SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 1–15

PARABOLIC BEHAVIOR OF A HYPERBOLIC DELAY EQUATION*

THOMAS LAURENT†, BRIAN RIDER‡, AND MICHAEL REED†

Abstract. It is shown that the fundamental solution of a hyperbolic partial differential equation with time delay has a natural probabilistic structure, i.e., is approximately Gaussian, as $t \to \infty$. The proof uses ideas from the DeMoivre proof of the central limit theorem. It follows that solutions of the hyperbolic equation look approximately like solutions of a diffusion equation with constant convection as $t \to \infty$.

Key words. hyperbolic equation, time delay, probabilistic behavior, parabolic equation

AMS subject classifications. 35R10, 35K15, 60F05

DOI. 10.1137/040611422

1. Introduction.

It has long been known that time delays have a smoothing effect on ordinary differential equations. For example, Kolmanovskii and Myshkis [8] state that, "This property of 'solution smoothing' ... together with some other properties make retarded differential equations resemble parabolic differential equations. However, the reasons for this resemblance are not entirely clear." In this paper, we provide an analytical foundation for these ideas by studying the initial value problem for the linear hyperbolic equation

$$\frac{\partial}{\partial t}u(t,x) \; + \; c\frac{\partial}{\partial x}u(t,x) = -Au(t,x) \; + Bu(t-\tau,x), \tag{1}$$

$$u(t,x) = f(t,x) \qquad \text{for} \quad -\tau \le t \le 0, \tag{2}$$

where A and B are positive constants and τ is the time delay.

In section 2, we introduce the "fundamental solution,"

$$K(t,x) = \sum_{k=0}^{n} \gamma_k(t)\delta_{ct-ck\tau}(x) \qquad \text{for } \; n\tau \le t < (n+1)\tau, \tag{3}$$

where $n = 0, 1, 2, \ldots$, and we define $K(t,x) \equiv 0$ for $t\epsilon[-\tau, 0)$. The coefficients $\gamma_k(t)$ are defined recursively: $\gamma_0(t)$ satisfies $\gamma_0'(t) = -A\gamma_0(t)$ with $\gamma_0(0) = 1$ and

$$\gamma_k'(t) = -A\gamma_k(t) \; + \; B\gamma_{k-1}(t-\tau) \qquad \text{with } \gamma_k(k\tau) = 0. \tag{4}$$

Using standard functional analytic methods, we prove that $K(t) \equiv K(t,\cdot)$ is a continuous $\mathcal{D}'(\mathbb{R})$-valued function on $[0,\infty)$ that is differentiable except at τ. We prove that if $f(t)$ is any continuous $\mathcal{D}'(\mathbb{R})$-valued initial data on $[-\tau, 0]$, then there is a unique $\mathcal{D}'(\mathbb{R})$-valued solution of (1) given by

$$u(t) = K(t) * f(0) \; + \; B\int_{-\tau}^{0} K(t-\theta-\tau) * f(\theta)d\theta \qquad \text{for } \; t > 0. \tag{5}$$

*Received by the editors July 12, 2004; accepted for publication (in revised form) May 31, 2005; published electronically March 31, 2006. The research in the paper and its preparation were supported by NSF grants DMS-9983320 and DMS-0109872.

http://www.siam.org/journals/sima/38-1/61142.html

†Department of Mathematics, Duke University, Durham, NC 27708 (thomas@math.duke.edu, reed@math.duke.edu).

‡Department of Mathematics, University of Colorado at Boulder, Boulder, CO 80309 (brider@euclid.colorado.edu)

If one integrates (1) with respect to x (assuming that $u(t,x) \to 0$ as $|x| \to \infty$), one obtains a solution of the retarded differential equation (RDE)

$$y'(t) = -Ay(t) + By(t-\tau) \quad \text{for } t > 0. \tag{6}$$

We shall denote the fundamental solution of (6) by $Y(t)$, i.e., $Y(t)$ is the solution with initial data equal to 1 at $t = 0$ and equal to 0 at earlier times. The general solution of (6) is given by a formula similar to (5), where $Y(t)$ replaces $K(t)$ and convolution is replaced by multiplication [7].

In section 3, the main section of the paper, we analyze the asymptotic behavior as $t \to \infty$ of the fundamental solution $K(t)$. The characteristic function of (6) is $h(\lambda) = \lambda + A - Be^{-\lambda\tau}$. The root of h with largest real part, λ_0, is real and we define $p = 1/h'(\lambda_0)$ and $\alpha = \tau c^2(1-p)p^2$. We prove that $K(t)$, after normalization, looks asymptotically like a Gaussian with standard deviation $\sigma = \sqrt{\alpha t}$ translating at speed cp. The proof depends on the combinatorics of the functions $\gamma_k(t)$ and follows many of the ideas of the DeMoivre proof of the central limit theorem.

In section 4, we assume that f is a continuous function on $[-\tau, 0]$ with values in $L^1(\mathbb{R})$. Using an appropriate seminorm, we show that asymptotically the solution $u(t)$ of (1) looks likes a weighted average of solutions of the transport heat equation,

$$\frac{\partial}{\partial t}v(t,x) \; + \; cp\frac{\partial}{\partial x}v(t,x) = \frac{\alpha}{2}\frac{\partial^2}{\partial x^2}v(t,x), \tag{7}$$

on the space scale $c\tau$.

We note that (1) has a natural interpretation in population dynamics. $u(t,x)$ is the population density in space at time t. Members of the population move to the right at speed c, die at rate A, and produce eggs at a rate B. The eggs are stationary, but after a time τ become moving members of the population. In this language, $Y(t)$ plays the role of the total population at time t. Hyperbolic equations with time delays occur frequently in ecology and cell biology [3], [9], [11], where $u(t,x)$ is the density of the population at different maturities x. Often these equations are nonlinear, boundary value problems with inhomogeneous velocities, so it is not clear whether the probabilistic methods of this study can be extended to those cases.

2. Existence, uniqueness, and representation. We denote by $\mathcal{D}(\mathbb{R})$ the usual space of test functions of compact support and by $\mathcal{D}'(\mathbb{R})$ the space of distributions. By convergence of a sequence $\mu_n \to \mu$ in $\mathcal{D}'(\mathbb{R})$, we always mean weak convergence, i.e., $\langle \mu_n, \psi \rangle \to \langle \mu, \psi \rangle$ for all $\psi \epsilon \mathcal{D}(\mathbb{R})$. For each fixed t, the sum defining $K(t)$ is finite, so it is clear that $K(t) \epsilon \mathcal{D}'(\mathbb{R})$ and the support of $K(t)$ is contained in the interval $[0, ct]$. Since we are interested in the properties of K as a function of t, we begin by summarizing briefly the basic definitions and properties of $\mathcal{D}'(\mathbb{R})$-valued functions that we use repeatedly.

A $\mathcal{D}'(\mathbb{R})$-valued function, f, is said to be *continuous* at t_o if $\langle f(t), \psi \rangle$ is continuous at t_o for all $\psi \epsilon \mathcal{D}(\mathbb{R})$.

PROPOSITION 2.1. *If f is continuous at t_o, then*

(a) *if $t_n \to t_o$, then $f(t_n) \to f(t_o)$ in $\mathcal{D}'(\mathbb{R})$.*

(b) *$\partial_x f$ is continuous at t_o.*

(c) *if $f(t)$ has support in $[0, ct]$ for each t and $v \epsilon \mathcal{D}'(\mathbb{R})$, then $f(t) * v$ is continuous at t_o.*

Proof. (a) simply reformulates the definition. Since $\langle \partial_x f(t), \psi \rangle = \langle f(t), -\partial_x \psi \rangle$, (b) is immediate. To prove (c), suppose $t_n \to t_o$. Since there is a ball $\mathbb{B}$ that

contains the supports of all the distributions $f(t_n)$ and since $f(t_n) \to f(t_o)$ in $\mathcal{D}'(\mathbb{R})$ by assumption, we conclude that $f(t_n) * v \to f(t_o) * v$ using the bicontinuity of convolution on $\mathcal{D}'(\mathbb{R}) \times \mathcal{D}'(\mathbb{R})$ [5, p. 105], [6, p. 71]. □

A $\mathcal{D}'(\mathbb{R})$-valued function, f, is said to be *differentiable* at t_o if $\langle \frac{f(t_n)-f(t_o)}{t_n-t_o}, \psi \rangle$ converges for all $\psi \epsilon \mathcal{D}(\mathbb{R})$ as $t_n \to t_o$.

PROPOSITION 2.2. *If f is differentiable at t_o, then*

(a) *there exists $f'(t_o)$ in $\mathcal{D}'(\mathbb{R})$ such that $\frac{f(t_n)-f(t_o)}{t_n-t_o} \to f'(t_o)$ in $\mathcal{D}'$.*

(b) *$\partial_x f$ is differentiable at t_o and $(\partial_x f)'(t_o) = \partial_x f'(t_o)$.*

(c) *if $f(t)$ has support in $[0, ct]$ for each t and $v \epsilon \mathcal{D}'(\mathbb{R})$, then $f(t) * v$ is differentiable at t_o and $(f(t) * v)' = f'(t) * v$.*

Proof. (a) A weakly convergent sequence in $\mathcal{D}'(\mathbb{R})$ has a limit in $\mathcal{D}'(\mathbb{R})$ [10, p. 15]. Since

$$\left\langle \partial_x \frac{f(t_n) - f(t_o)}{t_n - t_o}, \psi \right\rangle \;=\; -\left\langle \frac{f(t_n) - f(t_o)}{t_n - t_o}, \partial_x \psi \right\rangle,$$

(b) follows by taking limits. To prove (c), one writes the difference quotient and takes limits using the bicontinuity of convolution as above. □

A $\mathcal{D}'(\mathbb{R})$-valued function, f, is said to be *Riemann integrable* on $[a, b]$ if $\langle f(t), \psi \rangle$ is Riemann integrable on $[a, b]$ for all $\psi \epsilon \mathcal{D}(\mathbb{R})$.

PROPOSITION 2.3. *If f is Riemann integrable on $[a, b]$, then*

(a) *there exists an element of $\mathcal{D}'(\mathbb{R})$, denoted $\int_a^b f(t)\, dt$, such that*

$$\left\langle \int_a^b f(t)\, dt, \psi \right\rangle \;=\; \int_a^b \langle f(t), \psi \rangle \; dt \quad \textit{for all } \; \psi \, \epsilon \mathcal{D}(\mathbb{R}).$$

(b) $\partial_x \int_a^b f(t)\, dt \;=\; \int_a^b \partial_x f(t)\, dt$.

Proof. Let $\mathcal{P}_n$ be a sequence of partitions whose mesh size goes to zero as $n \to \infty$. Since $\langle \sum_{t_i \epsilon \mathcal{P}_n} f(t_i)(t_i - t_{i-1}), \psi \rangle$ converges for each ψ, $\sum_{t_i \epsilon \mathcal{P}_n} f(t_i)(t_i - t_{i-1})$ has a limit in $\mathcal{D}'(\mathbb{R})$. (b) follows, as above, by the adjoint relation for ∂_x. □

The following lemma summarizes the smoothness properties of $K(t)$.

LEMMA 2.4. *Let $K(t)$ be defined by* (3) *and* (4).

(a) *$K(t)$ is continuous on $[0, \infty)$.*

(b) *$K(t)$ is differentiable on $(0, \infty) \backslash \{\tau\}$ and*

$$K'(t) = -c\partial_x K(t) \;-\; AK(t) \;+\; BK(t - \tau) \quad \textit{for } \; t \epsilon (0, \infty) \backslash \{\tau\}. \tag{8}$$

Proof. Each $\delta_{ct-ck\tau}$ is a continuously differentiable $\mathcal{D}'(\mathbb{R})$-valued function of t and the coefficients $\gamma_k(t)$ are C^∞ functions of t since they are the solutions of ordinary differential equations with C^∞ source terms. Thus, $K(t)$ is continuously differentiable on each open interval of the form $(n\tau, (n+1)\tau)$ and we need only check the points $n\tau$ where the definition, (3), changes. Since $\gamma_n(n\tau) = 0$, for $n \geq 1$, $K(t)$ is continuous on $[0, \infty)$. Similarly, (4) shows that $\gamma_n'(n\tau) = 0$ for $n \geq 2$, which implies that $K(t)$ is differentiable at $n\tau$ for $n \geq 2$.

If $t \epsilon (0, \tau)$, then for all $\psi \epsilon \mathcal{D}(\mathbb{R})$, we know that $\langle K(t), \psi \rangle = \gamma_0(t)\psi(ct)$. Thus,

$$\begin{aligned} \frac{d}{dt}\langle K(t), \psi \rangle &= -A\gamma_0(t)\psi(ct) \;+\; c\gamma_0(t)\psi'(ct) \\ &= -A\langle K(t), \psi \rangle \;-\; c\langle \partial_x K(t), \psi \rangle. \end{aligned}$$

Since $K(t - \tau) = 0$, (8) holds.

Let $t\epsilon(n\tau,(n+1)\tau)$ for $n \geq 1$. Then,

$$\begin{aligned}\frac{d}{dt}\langle K(t),\psi\rangle &= \frac{d}{dt}\sum_{k=0}^{n}\gamma_k(t)\psi(ct-ck\tau)\\ &= -A\gamma_0(t)\psi(ct)+c\gamma_0(t)\psi'(ct)\\ &\quad+\sum_{k=1}^{n}(-A\gamma_k(t)+B\gamma_{k-1}(t-\tau))\psi(ct-ck\tau)\ +\ c\gamma_k(t)\psi'(ct-ck\tau)\\ &= -c\langle\partial_x K(t),\psi\rangle\ -\ A\langle K(t),\psi\rangle\ +\ B\langle K(t-\tau),\psi\rangle.\end{aligned}$$

The calculation for $t=(n+1)\tau$ is the same because $\gamma_{n+1}((n+1)\tau)=\gamma'_{n+1}((n+1)\tau)=0$. Thus, (8) holds on $(n\tau,(n+1)\tau]$ for all $n\geq 1$. □

LEMMA 2.5. *Let $f(t)$ be a continuous $\mathcal{D}'(\mathbb{R})$-valued function on $[-\tau,0]$. Then, the function $t\longmapsto\int_{-\tau}^{0}K(t-\theta-\tau)*f(\theta)\,d\theta$ is well defined and differentiable on $(0,\infty)$ and*

$$\text{for } t\geq\tau,\quad \frac{d}{dt}\int_{-\tau}^{0}K(t-\theta-\tau)*f(\theta)\,d\theta=\int_{-\tau}^{0}K'(t-\theta-\tau)*f(\theta)\,d\theta, \tag{9}$$

$$\begin{aligned}\text{for } t\epsilon(0,\tau),\quad \frac{d}{dt}\int_{-\tau}^{0}K(t-\theta-\tau)*f(\theta)\,d\theta &= f(t-\tau)\\ &\quad+\int_{-\tau}^{0}K'(t-\theta-\tau)*f(\theta)\,d\theta.\end{aligned} \tag{10}$$

Proof. For $n=-1,0,1,2,\ldots$, define the following regions of the $t-\theta$ plane:

$$R_n\ =\ \{(t,\theta)\mid n\tau\leq t-\theta-\tau<(n+1)\tau\quad\text{and}\quad -\tau\leq\theta\leq 0\}.$$

Since $K(t)$ is a C^∞ function of t in the open intervals $(n\tau,(n+1)\tau)$, $K(t-\theta-\tau)$ is C^∞ except on the boundaries of the regions R_n indicated by the dashed lines in Figure 1.

We suppose $\psi\epsilon\mathcal{D}$ and let T_a be the translation operator $T_a\psi(x)=\psi(x-a)$. For (t,θ) in $\bigcup_{-1}^{\infty}R_n$, we define

$$z(t,\theta)\ \equiv\ \langle K(t-\theta-\tau)*f(\theta),\psi\rangle$$

and note that $z\equiv 0$ on R_{-1} since $K(t)=0$ for $t<0$. On R_0, we have

$$\begin{aligned}z(t,\theta)&=\langle\gamma_0(t-\theta-\tau)\delta_{c(t-\theta-\tau)}*f(\theta),\psi\rangle\\ &=\gamma_0(t-\theta-\tau)\langle f(\theta),T_{-c(t-\theta-\tau)}\psi\rangle.\end{aligned}$$

Note that if $a_n\to a$, then $T_{a_n}\psi\to T_a\psi$ in $\mathcal{D}$, so by the continuity of f and the bicontinuity of $\langle\cdot,\cdot\rangle$ on $\mathcal{D}'\times\mathcal{D}$, we conclude that z is continuous on R_0. Since this is true for all $\psi\epsilon\mathcal{D}$, Proposition 2.3 assures us that for $0<t<\tau$ the integral $\int_{-\tau}^{0}K(t-\theta-\tau)*f(\theta)\,d\theta$ makes sense in $\mathcal{D}'$.

Let R_0^o denote the interior of R_0. If $(t,\theta)\epsilon R_0^o$, then $0<t-\theta-\tau<\tau$, so z is differentiable. Using the general fact [10, p. 105] that $\frac{d}{dy}\langle\mu(x),\psi(x+y)\rangle=\langle\mu(x),\psi'(x+y)\rangle$, we find

$$\begin{aligned}\frac{\partial}{\partial t}z(t,\theta)&=\gamma_0'(t-\theta-\tau)\langle f(\theta),T_{-c(t-\theta-\tau)}\psi\rangle\\ &\quad+c\gamma_0(t-\theta-\tau)\langle f(\theta),T_{-c(t-\theta-\tau)}\psi'\rangle.\end{aligned}$$

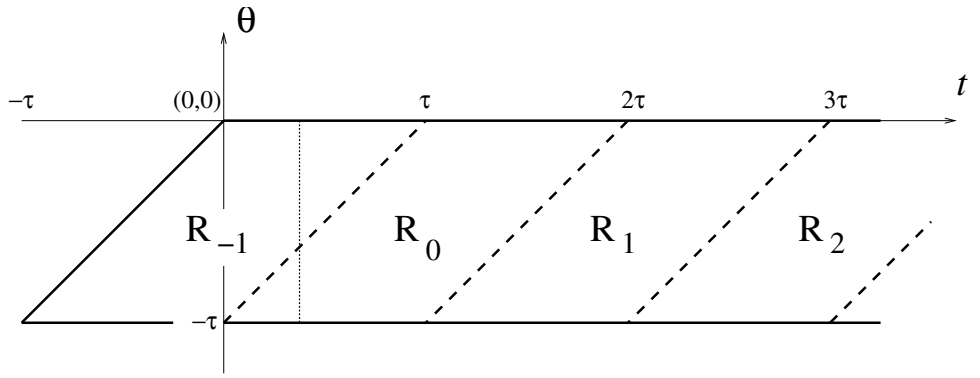


FIG. 1. *The regions R_n.*

This formula shows that $\frac{\partial}{\partial t}z(t,\theta)$ is continuous on R_0^o and can be extended continuously to the boundary of R_0^o. Thus, $\frac{\partial}{\partial t}z(t,\theta)$ is bounded in R_0^o and the usual proof (using the mean value theorem and the dominated convergence theorem) shows that

$$\frac{d}{dt}\int_{-\tau}^{0} z(t,\theta)\,d\theta = \frac{d}{dt}\int_{-\tau}^{t-\tau} z(t,\theta)\,d\theta = z(t,t-\tau)\ + \int_{-\tau}^{t-\tau}\frac{\partial}{\partial t}z(t,\theta)\,d\theta \qquad \text{for } 0<t<\tau.$$

Since this is true for all $\psi\epsilon\mathcal{D}$, Proposition 2.3 guarantees that (11) holds in $\mathcal{D}'$.

Formula (9) is proven similarly. The proof relies on the fact that z is continuous and $\frac{\partial}{\partial t}z(t,\theta)$ is bounded on $\bigcup_0^\infty R_n^o$. □

THEOREM 2.6. *Let $f:[-\tau,0]\to\mathcal{D}'$ be continuous and define*

$$u(t) = \begin{cases} K(t)*f(0)\ +\ B\int_{-\tau}^{0} K(t-\theta-\tau)*f(\theta)\,d\theta & \text{if } t>0,\\ f(t) & \text{if } -\tau\le t\le 0. \end{cases} \tag{11}$$

Then, $u(t)$ is the unique $\mathcal{D}'$-valued function that is continuous on $[-\tau,\infty)$, is differentiable on $(0,\infty)$, satisfies $u(t)=f(t)$ for $-\tau\le t\le 0$, and

$$u'(t)\ =\ -c\partial_x u(t)\ -\ Au(t)\ +\ Bu(t-\tau) \qquad \text{for } t>0. \tag{12}$$

Proof. Using the technical details from Propositions 2.1–2.3 and Lemma 2.5, the existence part of the proof is straightforward. For $t\ge\tau$,

$$\begin{aligned} u'(t) &= K'(t)*f(0)\ +\ B\int_{-\tau}^{0} K'(t-\theta-\tau)*f(\theta)\,d\theta\\ &= \{-c\partial_x K(t) - Ak(t) + BK(t-\tau)\}*f(0)\\ &\qquad + B\int_{-\tau}^{0}\{-c\partial_x K(t-\theta-\tau) - AK(t-\theta-\tau) + BK(t-\theta-2\tau)\}*f(\theta)\,d\theta\\ &= -c\partial_x u(t)\ -\ Au(t)\ +\ Bu(t-\tau), \end{aligned}$$

and for $0 < t < \tau$,

$$\begin{aligned} u'(t) &= K'(t) * f(0) \ + \ Bf(t-\tau) \ + \ B\int_{-\tau}^{t-\tau} K'(t-\theta-\tau) * f(\theta)\, d\theta \\ &= \{-c\partial_x K(t) - AK(t)\} * f(0) \ + \ Bu(t-\tau) \\ &\qquad + B\int_{-\tau}^{t-\tau} \{-c\partial_x K(t-\theta-\tau) - AK(t-\theta-\tau)\} * f(\theta)\, d\theta \\ &= -c\partial_x u(t) \ - \ Au(t) \ + \ Bu(t-\tau). \end{aligned}$$

To prove uniqueness, we need only show that $u(t) \equiv 0$ if $u(t)$ is a differentiable $\mathcal{D}'$-valued function on $(0,\infty)$ that satisfies (12) and $u(t) = 0$ for $-\tau \le t \le 0$. Let $\psi \epsilon \mathcal{D}$ and consider the function $\langle u(t), T_{ct}\psi\rangle$. By writing down the difference question and taking the limit (using the bicontinuity of $\langle\cdot,\cdot\rangle$ on $\mathcal{D}' \times \mathcal{D}$), one easily sees by using (12) that $\langle u(t), T_{ct}\psi\rangle$ is differentiable on $(0,\tau]$ and

$$\frac{d}{dt}\langle u(t), T_{ct}\psi\rangle = \langle u'(t) + c\partial_x u(t), T_{ct}\psi\rangle \tag{13}$$

$$= \langle -Au(t) + Bu(t-\tau), T_{ct}\psi\rangle \tag{14}$$

$$= -A\langle u(t), T_{ct}\psi\rangle. \tag{15}$$

Thus, for any $h > 0$, we have $\langle u(t), T_{ct}\psi\rangle = e^{-A(t-h)}\langle u(h), T_{ch}\psi\rangle$. Taking the limit as $h \to 0$ and using the continuity of u we conclude that $\langle u(t), T_{ct}\psi\rangle = 0$ for $0 \le t \le \tau$. Since this is true for all ψ, we have that $u(t) = 0$ in $\mathcal{D}'$ for $0 \le t \le \tau$. By iterating this argument, we find that $u(t) = 0$ for all $t > 0$. □

We remark that if the initial data, f, is a C^1 function on the strip $\mathbb{R} \times [-\tau, 0]$, the distribution solution is C^1 for $t > 0$ and satisfies the differential equation in the classical sense. Similarly, if f is a continuous function of t with values in $L^1(\mathbb{R})$ on $[-\tau, 0]$, then the solution $u(t)$ will be in $L^1(\mathbb{R})$ for all $t > 0$. These theorems can easily be proven by rewriting the differential equation as an integral equation and using the contraction mapping principle. Finally, consider the special case where $f(\theta) = 0$ for $\theta < 0$ and $f(0)$ is a nonnegative L^1 function. Then $u(t,\cdot)$ is the convolution of a nonnegative function with a finite linear combination of delta functions with positive coefficients, and so $u(t,\cdot)$ is nonnegative. Thus, integration in x shows that

$$||u(t,\cdot)||_1 \ = \ Y(t)||f(0)||_1,$$

where $Y(t)$ is the fundamental solution of (6).

3. The asymptotic behavior of K. The fundamental solution $K(t)$, which we denote now by K_t, is for each t a finite sum of point masses with weights $\gamma_k(t)$. K_t may be normalized to produce a proper probability measure, $\Pi_t = K_t/K_t(\mathbb{R})$, where $K_t(\mathbb{R}) = \sum_{k=0}^{n} \gamma_k(t)$ for $n\tau \le t < (n+1)\tau$. Note that $K_t(\mathbb{R}) = Y(t)$. The first form of our asymptotic result is stated in the language of convergence in distribution.

THEOREM 3.1. *Let $h(\lambda)$ be the characteristic polynomial of* (6) *and λ_0 be the root with the largest real part. Define $p = 1/h'(\lambda_0)$, $q = 1 - p$, and $\alpha = c^2\tau p^2 q$. Then,*

$$\lim_{t\to\infty} \Pi_t\Big[cpt + \sqrt{\alpha t}\ a,\ cpt + \sqrt{\alpha t}\ b\Big] = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

for any fixed $-\infty \le a < b \le \infty$.

From a dynamical perspective, Theorem 3.1 says that, normalized to have unit mass, the fundamental solution of the equation with delay τ and speed c resembles $G_t(x) = e^{-(x-cpt)^2/2\alpha t}/\sqrt{2\pi\alpha t}$ in the sense of measures. Theorem 3.1 follows easily from a stronger result of local limit type, Theorem 3.2 below.

Consider the special sequence of times $t_n = n\tau$, $n \to \infty$. For such t_n,

$$K_{t_n} \equiv \sum_{k\in\mathbb{Z}} b_n(k)\delta_{kc\tau},$$

where the $b_n(k)$ have the following form obtained by explicitly solving the equations (4):

$$b_n(k) = \begin{cases} \frac{k^{n-k}}{(n-k)!}\xi^{-k}(B\tau)^n & \text{for } 1 \le k \le n, \\ 0 & \text{otherwise,} \end{cases}$$

where $\xi = B\tau e^{A\tau}$. We let S_n be the total mass of K_{t_n}, i.e.,

$$S_n \;=\; \sum_{k=1}^{n} b_n(k) \;=\; Y(t_n). \tag{16}$$

In these terms, Π_{t_n} can be written

$$\Pi_{t_n} = \sum_{k\in Z} \frac{b_n(k)}{S_n}\delta_{kc\tau}. \tag{17}$$

The following local theorem is the main result of this section. By *local* we mean that we have uniform control of the individual masses $b_n(\cdot)$.

THEOREM 3.2. *Let $\pi_n(\mathbb{R})$ be the lattice $\{\frac{k}{\sqrt{n}}\}$ for k ranging over the integers, and let $[pn]$ denote the integer part of pn. Then*

$$\sup_{x\in\pi_n(\mathbb{R})} \left| \sqrt{n}\frac{b_n([pn]+x\sqrt{n})}{S_n} - \frac{1}{\sqrt{2\pi p^2 q}} e^{-\frac{x^2}{2p^2q}} \right| = O\left(\sqrt{\frac{\log n}{n}}\right) \tag{18}$$

as $n \to \infty$.

As already mentioned, the proof of Theorem 3.2 is similar in spirit to the classical Demoivre–Laplace calculation leading to the (local) central limit theorem for the binomial distribution (see, for example, [4, Chap. VII, sec. 3]). Note that the error on the right-hand side of (18) compares favorably with known estimates concerning Gaussian convergence of general lattice distributions (see [1, Chap. 5]).

We break the proof of Theorem 3.2 into several steps. First, we characterize p in a different way. Then, in Lemma 3.4 we describe the shape of the distribution $b_n(\cdot)$ and show that the mass peaks at a sequence $\{m_n\}$ that is within a fixed constant of $\{np\}$, the asymptotic mean. Lemma 3.5 contains the main asymptotic estimate; we renormalize by the mass at the maximum and define

$$f_n(x) \equiv \frac{b_n(m_n + x\sqrt{n})}{b_n(m_n)}. \tag{19}$$

The heart of the proof is a set of estimates on the sequence $\{f_n(x)\}$. Finally, using the asymptotic estimate, we prove Theorem 3.2 and sketch the proof of Theorem 3.1.

LEMMA 3.3. *For*

$$\psi(x) \equiv \frac{1-x}{x} e^{\frac{1-x}{x}} \qquad \text{and} \qquad \xi \equiv B\tau e^{A\tau},$$

p is the unique solution of $\xi = \psi(x)$ lying in $(0,1)$.

Proof. Previously, p was defined as $1/h'(\lambda_0)$, where $h(\lambda) = \lambda + A - Be^{-\lambda\tau}$ and λ_0 is the unique positive root of h. It follows that $(1-p)/p = h'(\lambda_0) - 1$ in which the right-hand side may be expressed as either $\tau(\lambda_0 + A)$ or $B\tau e^{-\lambda_0\tau}$. Thus,

$$\frac{1-p}{p} = B\tau e^{-\lambda_0\tau} = B\tau e^{A\tau} e^{-\tau(\lambda_0+A)} = \xi e^{-(1-p)/p},$$

as desired. For the uniqueness, observe that $\psi(x) < 0$ for $x < 0$ and $x > 1$ and is strictly decreasing on $[0,1]$ from $+\infty$ at $x = 0$ to 0 at $x = 1$. ∎

LEMMA 3.4. *For each n, the sequence $\{b_n(k)\}$ attains its maximum at a single index, m_n, and is increasing (decreasing) to the left (right) of that point, respectively. Furthermore, there exists a constant D such that $|m_n - np| \le D$.*

Proof. Define the following approximates to ψ:

$$\psi_n(x) \equiv \left(\frac{1-x}{x+1/n}\right)\left(1+\frac{1}{nx}\right)^{n(1-x)}.$$

It may be verified that each ψ_n lies under ψ and that $\lim_{n\to\infty} \psi_n = \psi$ holds pointwise. Moreover, the ψ_n have the same shape as ψ in that they decrease strictly from $+\infty$ to 0 as x ranges from 0 to 1. As such, there is a unique $p_n \in (0,1)$ satisfying $\xi = \psi_n(p_n)$.

For $k = 1, 2, \ldots, n-1$,

$$\frac{b_n(k+1)}{b_n(k)} = \frac{1}{\xi}\left(1+\frac{1}{k}\right)^{n-k}\left(\frac{n-k}{k+1}\right) = \frac{1}{\xi}\psi_n\left(\frac{k}{n}\right),$$

and so, $b_n(k)$ is increasing on $k \le np_n$ and then decreasing on $k \ge np_n$. If we let $m_n = [p_n n]$, the smallest integer $\ge p_n n$, it is left to prove that $p - p_n = O(1/n)$.

First note that the properties of ψ_n and ψ imply that $p_n < p$ and $\lim_{n\to\infty} p_n = p$. The first follows from $\psi(p) = \xi = \psi_n(p_n) < \psi(p_n)$ along with the strict decrease of ψ; the second follows from the first and the convergence $\psi_n \to \psi$. Now, set $q_n = 1 - p_n$. Since $\xi = \psi_n(p_n)$,

$$-\log\xi + nq_n \log\left(1+\frac{1}{np_n}\right) + \log\left(\frac{q_n}{p_n+1/n}\right) = 0,$$

and using the fact that $p_n \ge p/2$ for all large enough n, an expansion yields

$$-\log\xi + \frac{q_n}{p_n} + \log\left(\frac{q_n}{p_n}\right) = O\left(\frac{1}{n}\right). \tag{20}$$

Adding $\log\xi = \log\psi(p) = \log(q/p) + (q/p)$ to (20) gives

$$\frac{p-p_n}{pp_n} + \log\left(\frac{q_n}{q}\frac{p}{p_n}\right) = O\left(\frac{1}{n}\right).$$

The fact that $p > p_n$ (and so $q < q_n$) implies that each term on the left-hand side is strictly positive. Finally, $pp_n = O(1)$, showing that $p - p_n = O(\frac{1}{n})$, and the proof is complete. ∎

LEMMA 3.5. *There exists a constant C such that*

$$\sup_{x\in\pi_n(\mathbb{R})}\left|f_n(x)-e^{-\frac{x^2}{2p^2q}}\right|\le C\frac{1}{\sqrt{n}}. \tag{21}$$

Proof. From the definition of $b_n(k)$,

$$b_n(m_n+\sqrt{n}x)=\frac{(m_n+\sqrt{n}x)^{n-m_n-\sqrt{n}x}}{(n-m_n-\sqrt{n}x)!}\xi^{-(m_n+\sqrt{n}x)}(B\tau)^n$$

for $x\in\pi_n([\frac{1-m_n}{\sqrt{n}},\frac{n-m_n}{\sqrt{n}}])$. In this range, after dividing by $b_n(m_n)$ and taking the case $x>0$ we have

(22)

$$\begin{aligned}
f_n(x)&=\left(1+\frac{\sqrt{n}x}{m_n}\right)^{n-m_n}\frac{(n-m_n)!}{(n-m_n-\sqrt{n}x)!}\frac{1}{[\xi(m_n+\sqrt{n}x)]^{\sqrt{n}x}}\\
&=\left[\left(1+\frac{\sqrt{n}x}{m_n}\right)^{n-m_n}e^{-\sqrt{n}x\frac{q}{p}}\right]\left[\prod_{0\le k\le\sqrt{n}x-1}\frac{n-m_n-k}{nq}\right]\left[\frac{np}{(m_n+\sqrt{n}x)}\right]^{\sqrt{n}x}\\
&\equiv u_n(x)\times v_n(x)\times w_n(x).
\end{aligned}$$

The second line follows after multiplying through by $e^{-\sqrt{n}x\frac{q}{p}}\ (\xi\frac{p}{q})^{\sqrt{n}x}=1$ and some algebra. If $x<0$ is desired, $v_n(x)$ should be replaced by $\prod_{1\le k\le\sqrt{n}|x|}(\frac{nq}{n-m_n+k})$. However, the asymptotic considerations are similar for either $x>0$ or $x<0$; we assume $x>0$ for the rest of the proof. We note that $f_n(x)=0$ for x outside the interval $[\frac{1-m_n}{\sqrt{n}},\frac{n-m_n}{\sqrt{n}}]$, so we define u_n,v_n, and w_n to be zero there too.

We first estimate the quantities

$$|u_n(x)-e^{-\frac{qx^2}{2p^2}}|,\quad |v_n(x)-e^{-\frac{x^2}{2q}}|,\quad \text{and } |w_n(x)-e^{-\frac{x^2}{p}}|$$

on the interval $0\le x\le M\sqrt{\log n}$ for suitable M. This interval is convenient for the tail estimates below. Using the elementary inequalities $a-a^2\le\log(1+a)\le a+a^2$ for $|a|\le 1/2$, on the range $0\le x\le\frac{p}{4}n^{\frac{1}{2}}$ we have that

$$\begin{aligned}
\log w_n(x)&\le-\sqrt{n}x\log\left(1+\frac{x}{\sqrt{n}p}-\frac{D}{n}\right)\\
&\le-\sqrt{n}x\left(\frac{x}{\sqrt{n}p}-\frac{D}{n}\right)+2\sqrt{n}x\left(\left(\frac{x}{\sqrt{n}p}\right)^2+\left(\frac{D}{n}\right)^2\right)\\
&\le-\frac{x^2}{p}+2\left(\frac{Dx}{\sqrt{n}}+\frac{x^3}{\sqrt{n}p^2}\right)
\end{aligned}$$

and,

$$\log w_n(x)\ge-\frac{x^2}{p}-2\left(\frac{Dx}{\sqrt{n}}+\frac{x^3}{\sqrt{n}p^2}\right),$$

for all large n. It follows that there exist constants c_1,c_2,c_3, so that

$$|w_n(x)-e^{-x^2/p}|\le|e^{-x^2/p}(e^{\frac{c_1}{\sqrt{n}}(1+x^3)}-1)|\le\frac{c_2}{\sqrt{n}}(1+|x|^3)e^{-x^2/p}\le\frac{c_3}{\sqrt{n}} \tag{23}$$

for all x in $[0, n^{\frac{1}{6}}]$. Of course, for n large $[0, M\sqrt{\log n}]$ is contained in both $[0, \frac{p}{4}n^{\frac{1}{2}}]$ and $[0, n^{\frac{1}{6}}]$.

For $v_n(x)$,

$$\begin{aligned}(24)\quad \log v_n(x) &\leq \sum_{0\leq k\leq \sqrt{n}x-1} \log\left(1-\frac{k-D}{nq}\right) \\ &\leq \sum_{0\leq k\leq \sqrt{n}x-1} \left(\frac{-k}{nq}+\frac{2k^2}{n^2q^2}\right)+2\frac{Dx}{\sqrt{n}} \leq -\frac{x^2}{2q}+\frac{2}{\sqrt{n}}\left(1+Dx+\frac{x^3}{q^2}\right),\end{aligned}$$

with a lower bound of the form

$$(25)\qquad \log v_n(x) \geq -\frac{x^2}{2q}-\frac{2}{\sqrt{n}}\left(1+Dx+\frac{x^3}{q^2}\right).$$

For $u_n(x)$, one expands to third order to find

$$(26)\quad -\frac{qx^2}{2p^2}-\frac{q}{\sqrt{n}p}(2Dx+x^2+8x^3) \leq \log u_n(x) \leq -\frac{qx^2}{2p^2}+\frac{q}{\sqrt{n}p}(2Dx+x^2+8x^3).$$

Given the inequalities (24), (25), and (26), an argument similar to that for $w_n(x)$ shows that both $|v_n(x) - e^{-x^2/2q}|$ and $|u_n(x) - e^{-qx^2/2p^2}|$ are $O(1/\sqrt{n})$ for $x \in [0, M\sqrt{\log n}]$. It follows that the difference between $f_n(x)$ and $e^{-x^2/2p^2q} = e^{-x^2/p}e^{-x^2/2q}e^{-qx^2/2p^2}$ is also of order $1/\sqrt{n}$ for those values of x.

Finally, we prove a global estimate

$$(27)\qquad f_n(x) \leq c_4 e^{-c_5 x^2} \equiv f^*(x) \quad \text{for } x\epsilon\mathbb{R},$$

which will give us control over the tail region $x \geq M\sqrt{\log n}$. As above, we give the proof only for $x > 0$. First,

$$\log u_n(x) = (n-m_n)\log\left(1+\frac{\sqrt{n}x}{m_n}\right)-\frac{\sqrt{n}qx}{p} \leq \left(\frac{\sqrt{n}x}{m_n}\right)\left(\frac{np-m_n}{p}\right) \leq \frac{2D}{p^2},$$

and the product defining $v_n(x)$ has only a finite number of terms greater than one independent of n. Thus, $|u_n v_n|$ is bounded by a fixed constant for all n and x. To handle w_n, first notice that if n is large enough and $x \geq 2q\sqrt{n}$, then

$$\frac{n-m_n}{\sqrt{n}} \;\leq\; q\sqrt{n}+\frac{D}{\sqrt{n}} \;\leq\; x,$$

so $w_n(x) = 0$ by definition. Also, for n large enough and any $x > 1$, we know that $\frac{x}{\sqrt{n}p}-\frac{D}{n} > 0$, so there exists a constant, c_6, such that

$$\log\left(1+\frac{x}{\sqrt{n}p}-\frac{D}{n}\right) \;\geq\; c_6\left(\frac{x}{\sqrt{n}p}-\frac{D}{n}\right) \quad \text{for } 1 \leq x \leq 2q\sqrt{n}.$$

It follows that

$$(28)\qquad w_n(x) \leq e^{-c_6 x^2/p}e^{2c_6 D^2 q} \qquad \text{for } x \geq 1 \text{ and large } n.$$

By adjusting the multiplicative constant, this Gaussian bound extends to all of $x > 0$. The bound (27) follows from boundedness of $u_n v_n$ and (28).

To obtain the tail estimate, we choose M so that $M^2 \min\{c_5, \frac{1}{2p^2q}\} \geq \frac{1}{2}$, which implies that both $f_n(x)$ and $e^{-x^2/2p^2q}$ will be of order $1/\sqrt{n}$ on $x > M\sqrt{\log n}$. This concludes the proof. ∎

Proof of Theorem 3.2. We denote $f(x) = e^{-x^2/2p^2q}$ and define

$$\tilde{f}_n(x) \equiv \frac{b_n([np] + \sqrt{n}x)}{b_n(m_n)}.$$

Since $\tilde{f}_n(x) = f_n(x - \frac{1}{\sqrt{n}}(m_n - [np]))$, Lemma 3.5 and the triangle inequality imply

$$\left|\tilde{f}_n(x) - f(x)\right| \leq C\frac{1}{\sqrt{n}} + \sup_{|\delta| \leq D/\sqrt{n}} \left|f(x) - f(x+\delta)\right| \leq C_1\frac{1}{\sqrt{n}}. \tag{29}$$

Then,

$$\begin{aligned}\sqrt{2\pi p^2 q} - \frac{S_n}{\sqrt{n}b_n(m_n)} &= \int_{-\infty}^{\infty} f(x)dx - \sum_{k=-\infty}^{\infty} \frac{1}{\sqrt{n}} f_n\left(\frac{k}{\sqrt{n}}\right) \\ &= \left[\int_{-\infty}^{\infty} f(x)dx - \sum_{k=-\infty}^{\infty} \frac{1}{\sqrt{n}} f\left(\frac{k}{\sqrt{n}}\right)\right] + \frac{1}{\sqrt{n}}\left[\sum_{k=-\infty}^{\infty}\left(f\left(\frac{k}{\sqrt{n}}\right) - f_n\left(\frac{k}{\sqrt{n}}\right)\right)\right].\end{aligned}$$

Here the first term on the right-hand side is controlled by the error in the Riemann sum, which is $(1/\sqrt{n})\int_{-\infty}^{\infty}|f'|$. For the second term, recall that we know that $|f_n(\cdot) - f(\cdot)| \leq C/\sqrt{n}$ and the bound (27) both hold. Therefore,

$$\begin{aligned}\sum_{k=-\infty}^{\infty}\left(f\left(\frac{k}{\sqrt{n}}\right) - f_n\left(\frac{k}{\sqrt{n}}\right)\right) &\leq \sum_{|k| \leq \sqrt{\ell_n n}}\left[\frac{C}{\sqrt{n}}\right] + \sum_{|k| \geq \sqrt{\ell_n n}}(c_4 e^{-c_5 k^2/n} + e^{-k^2/2np^2q}) \\ &\leq 2C\sqrt{\ell_n} + c_6\sqrt{\frac{n}{\ell_n}}e^{-c_5\ell_n},\end{aligned}$$

and taking as ℓ_n as an appropriate multiple of $\log n$ shows that

$$\left|\frac{S_n}{\sqrt{n}b_n(m_n)} - \sqrt{2\pi p^2 q}\right| \leq C_2\sqrt{\frac{\log n}{n}}. \tag{30}$$

Putting together (29) and (30) completes the proof. ∎

Finally, we conclude this section with the following proof.

Proof of Theorem 3.1. First consider $t \to \infty$ along the special sequence $t_n = n\tau$. For any $a < b$, define

$$\begin{aligned}I_n(a,b) &= \Big[cp(n\tau) + \sqrt{\alpha(n\tau)}\ a, cp(n\tau) + \sqrt{\alpha(n\tau)}\ b\Big] \\ &= c\tau\Big[np + \sqrt{p^2qn}\ a, np + \sqrt{p^2qn}\ b\Big].\end{aligned}$$

Noting that the individual masses of Π_{t_n} are positioned at a distance of $c\tau$ from each

other, we see that

$$\begin{aligned}\lim_{n\to\infty} \Pi_{t_n}\Big[I_n(a,b)\Big] &= \lim_{n\to\infty} \sum_{\sqrt{p^2qna}\le k-np\le\sqrt{p^2qnb}} \left(\frac{b_n(k)}{S_n}\right) \\ &= \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{x\in\pi_n[\sqrt{p^2qa},\sqrt{p^2qb}]} \left(\sqrt{n}\frac{b_n([np]+\sqrt{n}x)}{S_n}\right) \\ &= \int_{-\sqrt{p^2qa}}^{\sqrt{p^2qb}} e^{-x^2/2p^2q}\frac{dx}{\sqrt{2\pi p^2q}}\end{aligned}$$

by the dominated convergence theorem. The pointwise convergence of the integrand is the statement of Theorem 3.2, and the domination is (27). This proves the result for the special sequence $t_n = n\tau$. The general statement as $t \to \infty$ follows by interpolation and the triangle inequality. □

4. Comparison to the transport heat equation. In Theorem 3.1 we showed that normalized $\Pi_t = K_t/Y(t)$ looks more and more like

$$G_t(dx) = \frac{e^{-(x-cpt)^2/2\alpha t}}{\sqrt{2\pi\alpha t}}\,dx$$

in the sense of probability measures as $t \to \infty$. This suggests that solutions of the hyperbolic equation with time delay (1) may look like solutions of the transport heat equation (7) for t large. We will prove two theorems that express this idea. For the first, Theorem 4.2, we shall consider the special case in which the initial data, f, is zero except at $t = 0$ and $f(0)\epsilon L^1(\mathbb{R})$. General initial conditions are considered in Theorem 4.3.

For $f\epsilon L^1(\mathbb{R})$, define $U(t)$ and $V(t)$ by

$$U(t)f \;\equiv\; K_t * f \quad \text{and} \quad V(t)f \;\equiv\; G_t * f.$$

$U(t)$ is a strongly continuous family of bounded operators and $V(t)$ is a strongly continuous semigroup on $L^1(\mathbb{R})$ for $t \ge 0$; $U(t)f$ satisfies (1) and $V(t)f$ satisfies (7). Furthermore, $||U(t)f||_1 \le Y(t)||f||_1$ and $||V(t)f||_1 \le ||f||_1$, with equality in both cases if f is nonnegative. We shall see that $U(t)/Y(t)$ and $V(t)$ are "comparable" for large t on the space scale $c\tau$.

Since $K_t/Y(t)$ is, for each t, a finite sum of point measures spaced at intervals of length $c\tau$, and G_t is smooth, we need a method of comparison that integrates over intervals of length $c\tau$. Let $\mathbb{M}(\mathbb{R})$ denote the finite Borel measures on $\mathbb{R}$, and for $\mu\epsilon\mathbb{M}(\mathbb{R})$ define

$$||\mu||_{1,c\tau} \;\equiv\; \sup_{-c\tau\le a\le 0} \sum_{k\epsilon\mathbb{Z}} |\mu([a+kc\tau, a+(k+1)c\tau))|.$$

Of course, any $f\epsilon L^1(\mathbb{R})$ corresponds to a finite Borel measure, and in that case

$$||f||_{1,c\tau} \;\equiv\; \sup_{-c\tau\le a\le 0} \sum_{k\epsilon\mathbb{Z}} \left| \int_{a+kc\tau}^{a+(k+1)c\tau} f(x)\,dx \right|.$$

We begin by collecting the properties of $||\cdot||_{1,c\tau}$.

PROPOSITION 4.1.
(a) $||\cdot||_{1,c\tau}$. *is a seminorm on* $\mathbb{M}(\mathbb{R})$ *that satisfies* $||\mu||_{1,c\tau} \le ||\mu||_{\mathbb{M}(\mathbb{R})}$.
(b) *If* $\mu\epsilon\mathbb{M}(\mathbb{R})$ *and* $f\epsilon L^1(\mathbb{R})$, *then* $||\mu * f||_{1,c\tau} \le ||\mu||_{1,c\tau}\,||f||_1$.
(c) *If* $g(t)$ *is a continuous function on* $[a,b]$ *with values in* $L^1(\mathbb{R})$, *then*

$$\left\| \int_a^b g(t)\,dt \right\|_{1,c\tau} \le \int_a^b ||g(t)||_{1,c\tau}\,dt.$$

Proof. The straightforward proof of (a) is omitted. Since g is a continuous function, (c) is proven by using the sublinearity of the seminorm in the standard proof for Riemann integrals. To prove (b),

$$\begin{aligned}
||\mu * f||_{1,c\tau} &= \sup_{-c\tau\le a\le 0} \sum_{k\epsilon\mathbb{Z}} |(\mu * f)([a+kc\tau, a+(k+1)c\tau))| \\
&= \sup_{-c\tau\le a\le 0} \sum_{k\epsilon\mathbb{Z}} \left| \int_{\mathbb{R}} \mu([a+kc\tau, a+(k+1)c\tau) - y) f(y)\,dy \right| \\
&\le \int_{\mathbb{R}} \left\{ \sup_{-c\tau\le a\le 0} \sum_{k\epsilon\mathbb{Z}} |\mu([a+kc\tau, a+(k+1)c\tau) - y)| \right\} |f(y)|\,dy \\
&= ||\mu||_{1,c\tau}\,||f||_1,
\end{aligned}$$

where we have used $(\mu * f)(A) = \int_{\mathbb{R}} \mu(A-y)f(y)dy$ to obtain the second line (see [2, p. 266]). □

THEOREM 4.2. *Suppose* $f\epsilon L^1(\mathbb{R})$. *Then, there exists a constant* C *such that for* t *large,*

$$||U(t)f/Y(t) \;-\; V(t)f||_{1,c\tau} \;\le\; C\frac{\log t}{\sqrt{t}}\,||f||_1. \tag{31}$$

Proof. According to Proposition 4.1(b),

$$\begin{aligned}
||U(t)f/Y(t) \;-\; V(t)f||_{1,c\tau} &= ||(K_t/Y(t)) * f \;-\; G_t * f||_{1,c\tau} \\
&\le ||(K_t/Y(t)) \;-\; G_t||_{1,c\tau}\,||f||_1,
\end{aligned}$$

so we need only prove

$$||K_t/Y(t) \;-\; G_t||_{1,c\tau} \le C\frac{\log t}{\sqrt{t}} \tag{32}$$

for large t. As in section 3, we give the details for the special sequence $t_n = n\tau$; the proof for general t follows from the triangle inequality. We set $g(x) = e^{-x^2/2p^2q}/\sqrt{2\pi p^2 q}$ and

$$g_n(x) = \sqrt{n}\frac{b_n([np] + \sqrt{n}x)}{S_n}.$$

First, we rewrite the left-hand side of (32) so that we can use the machinery and

results of section 3:

$$
\begin{aligned}
&||K_{t_n}/Y(t_n) - G_{t_n}||_{1,c\tau} \\
&\quad = \sup_{a\in(-c\tau,0]} \sum_{k\in\mathbb{Z}} \left| \frac{b_n(k)}{S_n} - \frac{1}{\sqrt{2\pi\alpha n\tau}} \int_{a+kc\tau}^{a+(k+1)c\tau} e^{-(x-cpn\tau)^2/2\alpha n\tau}\, dx \right| \\
&\quad = \sup_{a'\in(-1,0]} \sum_{k\in\mathbb{Z}} \left| \frac{b_n(k)}{S_n} - \frac{1}{\sqrt{n}} \int_{a'+k}^{a'+(k+1)} g\Big(\frac{z-np}{\sqrt{n}}\Big) dz \right| \\
&\quad = \sup_{a''\in(-\frac{1}{\sqrt{n}},0]} \sum_{k\in\mathbb{Z}} \left| \frac{1}{\sqrt{n}} g_n\left(\frac{k}{\sqrt{n}}\right) - \int_{a''+\frac{k}{\sqrt{n}}}^{a''+\frac{k+1}{\sqrt{n}}} g\left(y + \frac{[np]-np}{\sqrt{n}}\right) dy \right| \\
&\quad \le \frac{1}{\sqrt{n}} \sum_{k\in Z} \left| g_n\left(\frac{k}{\sqrt{n}}\right) - g\left(\frac{k}{\sqrt{n}}\right) \right| \\
&\qquad + \sum_{k\in\mathbb{Z}} \int_{\frac{k}{\sqrt{n}}}^{\frac{k+1}{\sqrt{n}}} \left(\sup_{a''\in(-\frac{1}{\sqrt{n}},0]} \left| g\left(\frac{k}{\sqrt{n}}\right) - g\left(y + a'' + \frac{[np]-np}{\sqrt{n}}\right) \right| \right) dy \\
&\quad \equiv \mathcal{A}_n + \mathcal{B}_n.
\end{aligned}
$$

To estimate $\mathcal{A}_n$, recall from Theorem 3.2 that $|g_n(\frac{k}{\sqrt{n}}) - g(\frac{k}{\sqrt{n}})| \le C_1(\log\frac{n}{n})^{\frac{1}{2}}$ independently of k. Also the bound (27) for f_n translates to $g_n(x) \le c_7 e^{-c_5 x^2}$ since $\sqrt{n}b_n(m_n)/S_n$ approaches a limit (see the proof of Theorem 3.2). It follows that

$$
\mathcal{A}_n \le \frac{1}{\sqrt{n}} \sum_{|k|\le M\sqrt{n\log n}} C_1\sqrt{\frac{\log n}{n}} + \frac{1}{\sqrt{n}} \sum_{|k|\ge M\sqrt{n\log n}} \left(c_6 e^{-c_5 k^2/n} + g\left(\frac{k}{\sqrt{n}}\right)\right) = O\left(\frac{\log n}{\sqrt{n}}\right)
$$

by the choice of M. The estimate for $\mathcal{B}_n$ follows the same strategy. The function g is globally Lipschitz, and so

$$
\sup_{|c|\le\frac{2}{\sqrt{n}}} \left| g\left(\frac{k}{\sqrt{n}}\right) - g\left(\frac{k}{\sqrt{n}} + c\right)\right| \le C_2 \frac{1}{\sqrt{n}},
$$

for all k. Using this for $|k| \le M\sqrt{n\log n}$ and the decay of g for larger k, we find

$$
\mathcal{B}_n \le \sum_{|k|\le M\sqrt{n\log n}} \frac{1}{\sqrt{n}} \cdot \frac{C_2}{\sqrt{n}} + \sum_{|k|\ge M\sqrt{n\log n}} \frac{1}{\sqrt{n}} \cdot 2g\left(\frac{k-2}{\sqrt{n}}\right) = O\left(\sqrt{\frac{\log n}{n}}\right).
$$

This proves (32) and thus completes the proof of (31). □

We now consider the case in which f, the initial data for (1), is a continuous $L^1(\mathbb{R})$-valued function of t for $-\tau \le t \le 0$. Define

$$
\begin{aligned}
u(t) &= U(t)f(0) \;+\; B\int_{-\tau}^{0} U(t-\theta-\tau)f(\theta)\,d\theta \\
v(t) &= Y(t)V(t)f(0) \;+\; B\int_{-\tau}^{0} Y(t-\theta-\tau)V(t-\theta-\tau)f(\theta)\,d\theta \\
y(t) &= Y(t)||f(0)||_1 \;+\; B\int_{-\tau}^{0} Y(t-\theta-\tau)||f(\theta)||_1\,d\theta;
\end{aligned}
$$

$u(t)$ is, of course, the solution of (1) with initial condition (2). Notice that $v(t)$ is not a solution of (7) but is a weighted average of solutions to (7); it will be clear from the proof why this weighted average is natural. $y(t)$ is the solution of (6) with initial data equal to $||f(t)||_1$ for $-\tau \le t \le 0$.

THEOREM 4.3. *Suppose that f is a continuous $L^1(\mathbb{R})$-valued function of t for $-\tau \le t \le 0$, and let $u(t)$, $v(t)$, and $y(t)$ be defined as above. Then, there is a constant C_1 so that for t large enough,*

$$||u(t) - v(t)||_{1,c\tau} \le C_1 \frac{\log t}{\sqrt{t}} y(t). \tag{33}$$

Proof. We subtract $v(t)$ from $u(t)$ and apply the seminorm $||\cdot||_{1,c\tau}$. Using Proposition 4.1(c) and estimate (32), we find

$$\begin{aligned} ||u(t) - v(t)||_{1,c\tau} &\le Y(t) C \frac{\log t}{\sqrt{t}} ||f(0)||_1 \\ &\quad + B \int_{-\tau}^{0} Y(t-\theta-\tau) C \frac{\log(t-\theta-\tau)}{\sqrt{t-\theta-\tau}} ||f(\theta)||_1 \, d\theta \\ &\le C_1 \frac{\log t}{\sqrt{t}} \left\{ Y(t)||f(0)||_1 + B \int_{-\tau}^{0} Y(t-\theta-\tau)||f(\theta)||_1 \, d\theta \right\} \end{aligned}$$

for large t, which proves (33). □

Note that if $f(t)$ is a nonnegative function for all $t\epsilon[-\tau, 0]$, then $y(t) = ||u(t)||_1 = ||v(t)||_1$, so (33) estimates the relative error.

REFERENCES

[1] R. N. BHATTACHARYA AND R. RANGA RAO, *Normal Approximation and Asymptotic Expansions*, John Wiley & Sons, New York, 1976.

[2] P. BILLINGSLEY, *Probability and Measure*, John Wiley & Sons, New York, 1995.

[3] J. DYSON, R. VILLELLA-BRESSAN, AND G. F. WEBB, *A semilinear transport equation with delays*, Int. J. Math. Sci., 32 (2003), pp. 2011–2026.

[4] W. FELLER, *An Introduction to Probability Theory and Its Applications, Volume* II, John Wiley & Sons, New York, 1971.

[5] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions, Vol.* I, Academic Press, New York, 1964.

[6] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions, Vol.* II, Academic Press, New York, 1968.

[7] J. K. HALE AND S. M. V. LUNEL, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[8] V. B. KOLMANOVSKII AND A. D. MYSHKIS, *Applied Theory of Functional Differential Equations*, Kluwer Academic Publishers, Boston, 1992.

[9] A. D. REY AND M. C. MACKEY, *Bifurcations and traveling waves in a delayed partial differential equation*, Chaos, 2 (1992), pp. 231–244.

[10] L. SCHWARTZ, *Théorie des Distributions*, Hermann, Paris, 1966.

[11] G. F. WEBB, *Theory of Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 16–36

VARIATIONAL PROPERTIES OF UNBOUNDED ORDER PARAMETERS*

BO LI†

Abstract. Order parameters in physical and biological systems can sometimes become unbounded as the size of an underlying system increases. It is proposed that such a quantity be modeled as a minimizer of the energy functional

$$I_\varepsilon(u) = \fint \left[\frac{\varepsilon^2}{2} |\nabla u|^2 - \frac{1}{2} \log(1 + |u|^2) \right] dx,$$

where u is constrained by a side condition, and $\varepsilon > 0$ is a parameter that is inversely proportional to the linear size of the system. It is shown that a minimizer of I_ε exists; the minimum value of I_ε scales as $\log \varepsilon$; and both the L^2 and H^1 norms of any minimizer of I_ε are of the order $O(1/\varepsilon)$, indicating the unboundedness of the order parameter. It is also shown that the renormalized energy functionals

$$J_\varepsilon(v) = I_\varepsilon \left(\frac{v}{\varepsilon} \right) - \log \varepsilon$$

Γ-converge to the functional

$$J(v) = \fint \left(\frac{1}{2} |\nabla v|^2 - \log |v| \right) dx.$$

Minimizers of this Γ-limit for scalar order parameters with the Dirichlet boundary condition are well characterized.

Key words. order parameters, variational models, energy asymptotics, renormalized energy, Γ-convergence

AMS subject classifications. 49J45, 49S05, 74G65

DOI. 10.1137/040621314

1. Introduction.

Order parameters in physical and biological systems, such as population, concentration, volume fractions, magnetization vectors, directors of liquid crystals, the slope of surface height profile of thin films, etc., are mathematically scalar or vector-valued functions, or gradients of functions. An order parameter can sometimes grow unbounded as the size of an underlying system increases. An example of such an unbounded order parameter is the slope of surface of an epitaxially growing thin film in some experimental situations [5, 12, 20].

We propose to model unbounded order parameters as possible minimizers or low energy configurations of the effective free energy functional

$$\hat{I}(\hat{u}) = \fint_{\widehat{\Omega}} \left[\frac{\alpha}{2} |\nabla \hat{u}|^2 - \frac{\beta}{2} \log \left(1 + |\hat{u}|^2 \right) \right] d\hat{x}, \tag{1.1}$$

where $\widehat{\Omega} \subset \mathbb{R}^n$ for some integer $n \geq 1$ is a bounded domain, $\alpha > 0$ and $\beta > 0$ are two material constants, and the functions $\hat{u} : \widehat{\Omega} \to \mathbb{R}^m$ with some integer $m \geq 1$ are constrained by a boundary condition or some other side conditions. Here and

*Received by the editors December 22, 2004; accepted for publication (in revised form) August 31, 2005; published electronically March 31, 2006. This work was partially supported by the NSF through grant DMS-0451466.

http://www.siam.org/journals/sima/38-1/62131.html

†Department of Mathematics, University of California at San Diego, 9500 Gilman Drive, Mail Code: 0112, La Jolla, CA 92093-0112 (bli@math.ucsd.edu).

below, we denote $\fint_E = \frac{1}{|E|}\int_E$ for a Lebesgue measurable set $E \subseteq \mathbb{R}^n$ with a finite and nonzero Lebesgue measure $|E|$. For any $a = (a_1, \dots, a_m) \in \mathbb{R}^m$, we denote $|a| = \sqrt{\sum_{i=1}^m a_i^2}$. For a differentiable function $u = (u_1, \dots, u_m) : D \to \mathbb{R}^m$ with $D \subset \mathbb{R}^n$ an open set, we define $\nabla u : D \to \mathbb{R}^{m\times n}$ to be the $m \times n$-matrix-valued function with $\partial_{x_j} u_i(x)$ the (i, j)-entry of $\nabla u(x)$ and denote $|\nabla u(x)| = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |\partial_{x_j} u_i(x)|^2}$ for $x = (x_1, \dots, x_n) \in D$.

In the special case with $\hat{u} = \nabla\hat{h}$ for some scalar function $\hat{h}$ defined on a two-dimensional domain, the functional (1.1) happens to be the Liapunov functional of the evolution equation

$$\frac{\partial\hat{h}}{\partial t} = -\alpha\Delta^2\hat{h} - \beta\nabla\cdot\left(\frac{\nabla\hat{h}}{1+|\nabla\hat{h}|^2}\right), \tag{1.2}$$

i.e., this equation is the gradient-flow induced by the functional (1.1). Equation (1.2) was first proposed phenomenologically in [12] to model the surface height profile $\hat{h}$, measured in a comoving frame, in epitaxial growth of thin films with a strong asymmetry of the adatom (adsorbed atom) attachment and detachment from lower and upper terraces to atomic step edges due to the existence of an energy barrier [4, 17, 18]. Numerical and analytical studies based on such a model have shown that the slope of the surface, $|\nabla\hat{h}|$, which is the order parameter in this case, grows unbounded, agreeing with experiments [10, 11, 12, 13, 15, 16, 21].

It is interesting to compare the energy functional (1.1) with a usual Ginzburg–Landau-type energy functional that has the term $(|\hat{u}|^2 - 1)^2$ or alike, instead of the negative logarithmic term in (1.1). Important examples of the latter include the Ginzburg–Landau energy for superconductors [3, 9] and the Cahn–Hilliard energy for phase separation [2]; both have been much studied. It is obvious that an order parameter modeled by a Ginzburg–Landau-type energy functional stays always bounded. If $|\hat{u}| \ll 1$, then by the Taylor expansion,

$$-\log(1+|\hat{u}|^2) = -|\hat{u}|^2 + \frac{1}{2}|\hat{u}|^4 + O(|\hat{u}|^6) = \frac{1}{2}(|\hat{u}|^2-1)^2 - \frac{1}{2} + O(|\hat{u}|^6).$$

Thus, both types of energy functionals have approximately the same energy landscape for admissible functions with very small magnitude. As a consequence, the zero function as a critical point is unstable in both types of models.

Intuitively, if the energy $\hat{I}(\hat{u})$ of an admissible function $\hat{u} : \widehat{\Omega} \to \mathbb{R}^m$ is very small, then the magnitude $|\hat{u}|$ of the function must be very large in some norm. But, the boundary condition (or other side conditions) and the presence of the gradient term in the energy $\hat{I}(\hat{u})$ prevent $|\hat{u}|$ from being too large. These competing mechanisms determine the magnitude of such a low energy function to be finite but to grow unbounded as the system size increases. Our primary goals of this work are to quantify such unboundedness and to characterize the asymptotic behavior of energy functionals for systems of large size.

We shall not, however, directly work with the functional $\hat{I}$ defined in (1.1). Rather, we shall first rescale the energy functional. The idea is clear for the special case that $\widehat{\Omega} = (0, \widehat{L})^n$, a cube in $\mathbb{R}^n$ of linear size $\widehat{L} > 0$: letting $u(x) = \hat{u}(\hat{x})$ with $x = \widehat{L}^{-1}\hat{x}$, one obtains that $I_\varepsilon(u) = \beta\hat{I}(\hat{u})$, where

$$I_\varepsilon(u) = \fint_\Omega \left[\frac{\varepsilon^2}{2}|\nabla u|^2 - \frac{1}{2}\log(1+|u|^2)\right] dx, \tag{1.3}$$

$\varepsilon = \sqrt{\alpha/\beta}\widehat{L}^{-1}$, and Ω is the unit cube of $\mathbb{R}^n$. Now, for a general bounded domain $\widehat{\Omega}$, one can fix some point $\hat{x}_0 \in \widehat{\Omega}$ and apply the change of variable $\hat{x} \to x = \widehat{L}^{-1}(\hat{x} - \hat{x}_0)$ with $\widehat{L}$ being the diameter of $\widehat{\Omega}$. One again obtains an equivalent variational problem with the energy functional given by (1.3), in which ε is inversely proportional to $\widehat{L}$ and $\Omega \subset \mathbb{R}^n$ is a fixed bounded domain whose diameter is independent of $\widehat{L}$.

Depending on how an underlying physical and biological system is modeled mathematically, the set of admissible functions, to be denoted by $\mathcal{H}(\Omega, \mathbb{R}^m)$, for the energy functional I_ε can be defined differently. In this work, we assume that $\Omega \subset \mathbb{R}^n$ in the definition of I_ε is a bounded domain with a Lipschitz-continuous boundary $\partial\Omega$, and define

$$\mathcal{H}(\Omega, \mathbb{R}^m) = H_0^1(\Omega, \mathbb{R}^m), \tag{1.4}$$

or

$$\mathcal{H}(\Omega, \mathbb{R}^m) = \left\{ u \in H^1(\Omega, \mathbb{R}^m) : \fint_\Omega u \, dx = 0 \right\}, \tag{1.5}$$

where $H^1(\Omega, \mathbb{R}^m)$ and $H_0^1(\Omega, \mathbb{R}^m)$ are the spaces of vector-valued functions whose components are in the usual Sobolev spaces of scalar functions $H^1(\Omega)$ and $H_0^1(\Omega)$, respectively [1, 8]. In both cases, $\mathcal{H}(\Omega, \mathbb{R}^m)$ is a closed subspace of the Hilbert space $H^1(\Omega, \mathbb{R}^m)$ that is equipped with the norm $\|u\| = \sqrt{\|u\|^2 + \|\nabla u\|^2}$ for all $u \in H^1(\Omega, \mathbb{R}^m)$, where $\|\cdot\|$ denotes the $L^2(\Omega)$-norm.

Our major results are as follows:

(1) For each $\varepsilon > 0$, there exists a minimizer of $I_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$. Moreover, the minimum energy scales as $\log \varepsilon$, and both the L^2 and H^1 norms of any minimizer of $I_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$ are of the order $O(1/\varepsilon)$, cf. Theorem 2.1;

(2) The renormalized energy functionals $J_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$, defined by

$$J_\varepsilon(v) = I_\varepsilon\left(\frac{v}{\varepsilon}\right) - \log \varepsilon \qquad \forall v \in \mathcal{H}(\Omega, \mathbb{R}^m), \tag{1.6}$$

Γ-converge to the energy functional $J : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R} \cup \{\infty\}$ defined by

$$J(v) = \fint_\Omega \left(\frac{1}{2}|\nabla v| - \log |v| \right) dx, \qquad v \in \mathcal{H}(\Omega, \mathbb{R}^m), \tag{1.7}$$

cf. Theorem 3.2. Moreover, if $v_\varepsilon \in \mathcal{H}(\Omega, \mathbb{R}^m)$ for $\varepsilon > 0$ is a minimizer of $J_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R} \cup \{\infty\}$, then there exists a subsequence of $\{v_\varepsilon\}_{\varepsilon>0}$ that converges *strongly* in $\mathcal{H}(\Omega, \mathbb{R}^m)$ to a minimizer of $J : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R} \cup \{\infty\}$, cf. Theorem 3.3;

(3) In the case of the scalar Dirichlet boundary-value problem, there exists a unique $v_+ \in H_0^1(\Omega)$ such that v_+ is smooth and positive in Ω, and that v_+ and $-v_+$ are the only minimizers of $J : H_0^1(\Omega) \to \mathbb{R} \cup \{\infty\}$, cf. Theorem 4.1.

The results in (1) and (2) hold also true if the set of admissible functions is $\{u \in H_0^1(\Omega, \mathbb{R}^m) : \fint_\Omega u \, dx = 0\}$, or if Ω is an open cube in $\mathbb{R}^n$ with its faces parallel to the coordinate planes and the corresponding set of admissible functions is $\{u \in H^1_{\rm per}(\Omega, \mathbb{R}^m) : \fint_\Omega u \, dx = 0\}$, where $H^1_{\rm per}(\Omega, \mathbb{R}^m)$ is the closure in $H^1(\Omega, \mathbb{R}^m)$ of the set of all C^∞, $\overline{\Omega}$-periodical functions from $\mathbb{R}^n$ to $\mathbb{R}^m$.

The heuristics behind the first part of our results is well illustrated in our previous work [13] through the calculation of trial functions with low energy using an ad hoc ansatz and the calculation of critical points of the energy functional using matched

asymptotics, with both calculations being done in a one-dimensional setting. Our results in (1) generalize those in [13] for more complicated domains and include the optimal lower bound as well as the precise asymptotics of the minimum energy.

Our results do not directly apply to continuum models, such as the Liapunov functional of the equation (1.2), of the epitaxial growth with a significant attachment-detachment asymmetry of adatoms. This is because the set of admissible functions $\hat{u}$ of the functional (1.1) is larger than the set of gradient vector fields. However, the approach developed in this work can be applied to the study of such continuum models and to obtain similar results. In particular, the large-system-size Γ-limit of the rescaled Liapunov functionals—the functional (1.3) with u replaced by ∇h for the surface height function h that models a finite energy barrier—is precisely Villain's model for an infinite energy barrier [21].

Potentially, the positive solution to the scalar Dirichlet boundary-value problem for the limiting functional can be a good alternative to the distance function in the reinitialization process of the widely used level-set numerical method [14, 19]. We will address these issues of application in separate works.

In section 2, we present and prove the results for the energy functionals $I_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$. In section 3, we prove the Γ-convergence of the renormalized energies $J_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$ to the functional $J : \mathcal{H}(\Omega, \mathbb{R}) \to \mathbb{R} \cup \{\infty\}$. Finally, in section 4, we characterize solutions to the scalar Dirichlet problem of infimizing the limiting energy defined in (1.7).

2. Energy asymptotics and bounds of energy minimizers. We consider the energy functionals $I_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$, defined in (1.3) for a general domain Ω, only for $\varepsilon \in (0, 1]$, though many of our results hold true also for any $\varepsilon > 0$. For convenience, we denote $\|\!|u|\!\| = \sqrt{⨍_\Omega |u(x)|^2 dx} = (1/\sqrt{|\Omega|})\|u\|$ for all $u \in L^2(\Omega)$.

The following is our main result in this section:

THEOREM 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with a Lipschitz-continuous boundary $\partial\Omega$. Let $\mathcal{H}(\Omega, \mathbb{R}^m)$ be defined as in* (1.4) *or* (1.5).

(1) *For each $\varepsilon \in (0, 1]$, there exists $u_\varepsilon \in \mathcal{H}(\Omega, \mathbb{R}^m)$ such that*

$$I_\varepsilon(u_\varepsilon) = \min_{u \in \mathcal{H}(\Omega, \mathbb{R}^m)} I_\varepsilon(u). \tag{2.1}$$

(2) *Let $\mu_\varepsilon = \min_{u \in \mathcal{H}(\Omega, \mathbb{R}^m)} I_\varepsilon(u)$. There exist constants C_1 and C_2 that depend only on Ω such that*

$$C_1 + \log \varepsilon \le \mu_\varepsilon \le C_2 + \log \varepsilon \qquad \forall \varepsilon \in (0, 1]. \tag{2.2}$$

Moreover, $\mu_\varepsilon - \log \varepsilon$ increases as $\varepsilon \in (0, 1]$ decreases, $\nu := \sup_{0 < \varepsilon \le 1}(\mu_\varepsilon - \log \varepsilon)$ is finite, and

$$\lim_{\varepsilon \to 0^+} (\mu_\varepsilon - \log \varepsilon) = \nu. \tag{2.3}$$

(3) *There exist constants $C_j > 0$ $(j = 3, 4, 5, 6)$ and $\varepsilon_0 \in (0, 1]$, all depending only on Ω, such that for any minimizer $u_\varepsilon \in \mathcal{H}(\Omega, \mathbb{R}^m)$ of $I_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$ and for all $\varepsilon \in (0, \varepsilon_0]$,*

$$\frac{C_3}{\varepsilon} \le \|\!|u_\varepsilon|\!\| \le \frac{C_4}{\varepsilon}, \tag{2.4}$$

$$\frac{C_5}{\varepsilon} \le \|\!|\nabla u_\varepsilon|\!\| \le \frac{C_6}{\varepsilon}. \tag{2.5}$$

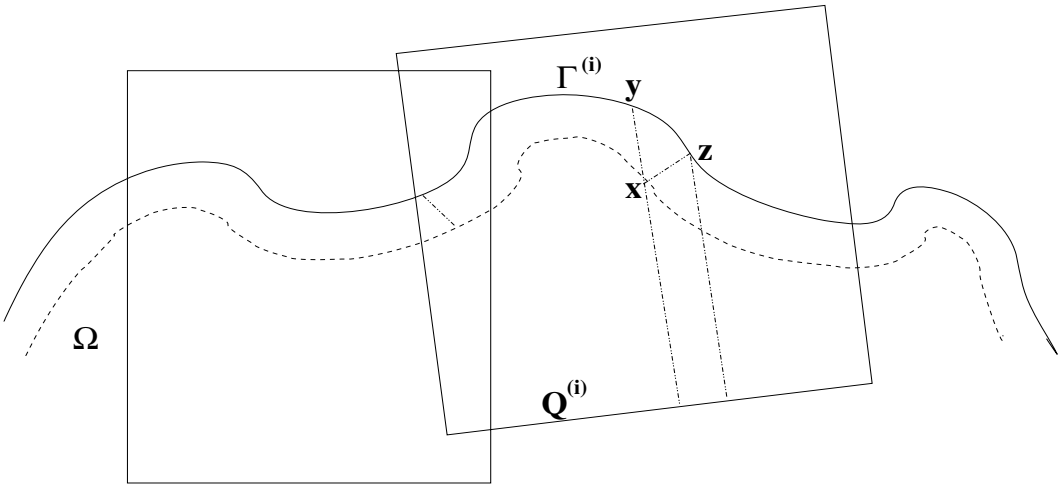


FIG. 2.2.1.

To prove this theorem, we need some preparations. We recall for any compact set $S \subset \mathbb{R}^n$ that the distance function $\text{dist}\,(\cdot, S) : \mathbb{R}^n \to \mathbb{R}$, defined by

$$\text{dist}\,(x, S) = \inf_{y \in S} |x - y| = \min_{y \in S} |x - y| \qquad \forall x \in \mathbb{R}^n,$$

is a Lipschitz-continuous function:

$$|\text{dist}\,(x, S) - \text{dist}\,(y, S)| \le |x - y| \qquad \forall x, y \in \mathbb{R}^n. \tag{2.6}$$

Moreover, it is differentiable almost everywhere in $\mathbb{R}^n$, and

$$|\nabla \text{dist}\,(x, S)| = 1 \qquad \text{a.e. } x \in \mathbb{R}^n, \tag{2.7}$$

cf. the proof of Lemma 3.2.34 in [7].

LEMMA 2.1. *If $\Omega \in \mathbb{R}^n$ is a bounded domain with a Lipschitz-continuous boundary $\partial\Omega$, then there exist constants $s_0 > 0$ and $C_0 > 0$, both depending only on Ω, such that*

$$|\{x \in \Omega : \text{dist}\,(x, \partial\Omega) \le s\}| \le C_0 s \qquad \forall s \in (0, s_0].$$

Proof. Since $\partial\Omega$ is Lipschitz-continuous, there exist finitely many Lipschitz-continuous functions $\phi^{(i)} : Q^{(i)} \to \mathbb{R}$ ($i = 1, \ldots, m$ for some integer $m \ge 1$) in local Cartesian coordinates with each $Q^{(i)} = \Pi_{j=1}^{n-1}[-\alpha_j^{(i)}, \alpha_j^{(i)}]$ for some $\alpha_j^{(i)} > 0$ ($1 \le i \le m$ and $1 \le j \le n-1$) a cube in $\mathbb{R}^{n-1}$, that satisfy the following properties (cf. Figure 2.2.1):

(1) For each i with $1 \le i \le m$, the local Cartesian coordinates $\xi^{(i)} = (\xi_1^{(i)}, \cdots, \xi_n^{(i)})$ are obtained by rotating and translating the original Cartesian coordinates $x = (x_1, \ldots, x_n)$;

(2) There exist $\alpha_n > 0$ and $\beta_n > 0$ such that for each cube $G^{(i)} := Q^{(i)} \times [-\alpha_n, \alpha_n] \subset \mathbb{R}^n$ ($1 \le i \le m$),

$$\Gamma^{(i)} := G^{(i)} \cap \partial\Omega = \left\{ \left(\hat{\xi}^{(i)}, \xi_n^{(i)} \right) \in \mathbb{R}^n : \hat{\xi}^{(i)} \in Q^{(i)},\, \xi_n^{(i)} = \phi^{(i)} \left(\hat{\xi}^{(i)} \right) \right\},$$

where $\hat{\xi}^{(i)} = \left(\xi_1^{(i)}, \ldots, \xi_{n-1}^{(i)}\right)$, and

$$U_+^{(i)} := \left\{ \left(\hat{\xi}^{(i)}, \xi_n^{(i)}\right) \in \mathbb{R}^n : \hat{\xi}^{(i)} \in Q^{(i)},\ \phi^{(i)}\left(\hat{\xi}^{(i)}\right) < \xi_n^{(i)} < \phi^{(i)}\left(\hat{\xi}^{(i)}\right) + \beta_n \right\} \subset G^{(i)} \cap \overline{\Omega}^c,$$

$$U_-^{(i)} := \left\{ \left(\hat{\xi}^{(i)}, \xi_n^{(i)}\right) \in \mathbb{R}^n : \hat{\xi}^{(i)} \in Q^{(i)},\ \phi^{(i)}\left(\hat{\xi}^{(i)}\right) - \beta_n < \xi_n^{(i)} < \phi^{(i)}\left(\hat{\xi}^{(i)}\right) \right\} \subset G^{(i)} \cap \Omega;$$

(3) The union $\cup_{i=1}^m G^{(i)}$ covers a neighborhood of the compact set $\partial\Omega$, and $\partial\Omega = \cup_{i=1}^m \Gamma_i$.

Now, since the distance function $\mathrm{dist}(\cdot, \partial\Omega) : \overline{\Omega} \to \mathbb{R}$ is Lipschitz-continuous and vanishes only on the boundary $\partial\Omega$ which is compact, by properties (2) and (3), there exists a constant $s_1 = s_1(\Omega) > 0$ such that

$$\{x \in \Omega : \mathrm{dist}\,(x, \partial\Omega) \le s_1\} \subseteq \bigcup_{i=1}^m U_-^{(i)}.$$

By this and property (3), there exist cubes $P^{(i)}$ in $\mathbb{R}^{n-1}$ with $P^{(i)} \subseteq Q^{(i)}$ $(1 \le i \le m)$ and a constant $s_0 = s_0(\Omega)$ with $0 < s_0 \le s_1$ that satisfy the following properties:

(4) For each i $(1 \le i \le m)$,

$$\{x \in \Omega : \mathrm{dist}\,(x, \partial\Omega) \le s_0\} \subseteq \bigcup_{i=1}^m V^{(i)},$$

where

$$V^{(i)} = \left\{ \left(\hat{\xi}^{(i)}, \xi_n^{(i)}\right) \in \mathbb{R}^n : \hat{\xi}^{(i)} \in \hat{P}^{(i)},\ \phi^{(i)}\left(\hat{\xi}^{(i)}\right) - \beta_n < \xi_n^{(i)} < \phi^{(i)}\left(\hat{\xi}^{(i)}\right) \right\} \subseteq U_-^{(i)};$$

(5) Let $x \in \Omega$ and $x' \in \partial\Omega$ be such that $\mathrm{dist}\,(x, \partial\Omega) = |x - x'| \le s_0$. If $x \in V^{(i)}$ for some i $(1 \le i \le m)$, then $x' \in \Gamma_i = G^{(i)} \cap \partial\Omega$.

Fix $s \in \mathbb{R}$ with $0 < s \le s_0$. Let $x \in \Omega$ be such that $\mathrm{dist}\,(x, \partial\Omega) \le s$. By property (4), we have $x \in V^{(i)}$ for some i with $1 \le i \le m$, cf. Figure 2.1. Let $\xi^{(i)} = (\hat{\xi}^{(i)}, \xi_n^{(i)})$ be the local coordinates of x in which $\hat{\xi}^{(i)} \in P^{(i)} \subseteq Q^{(i)}$. Let the point $y \in \mathbb{R}^n$ have the local coordinates $(\hat{\xi}^{(i)}, \phi^{(i)}(\hat{\xi}^{(i)}))$. Then, by property (2), $y \in \Gamma^{(i)} \subset \partial\Omega$, cf. Figure 2.1. Let $z \in \partial\Omega$ be such that $\mathrm{dist}\,(x, \partial\Omega) = |x - z| \le s$, cf. Figure 2.1. By property (5), $z \in \Gamma^{(i)} = G^{(i)} \cap \partial\Omega$. Thus, there exists $(\hat{\eta}^{(i)}, \eta_n^{(i)}) \in \Gamma^{(i)} = G^{(i)} \cap \partial\Omega$ such that $\hat{\eta}^{(i)} \in Q^{(i)}$, $(\hat{\eta}^{(i)}, \eta_n^{(i)}) = (\hat{\eta}^{(i)}, \phi^{(i)}(\hat{\eta}^{(i)}))$ are the local coordinates of z, and

$$\mathrm{dist}\,(x, \partial\Omega) = |x - z| = \left| \left(\hat{\xi}^{(i)}, \xi_n^{(i)}\right) - \left(\hat{\eta}^{(i)}, \phi^{(i)}\left(\hat{\eta}^{(i)}\right)\right) \right| \le s. \tag{2.8}$$

This implies that

$$\left|\hat{\xi}^{(i)} - \hat{\eta}^{(i)}\right| \le s. \tag{2.9}$$

Denoting by $L_i > 0$ the Lipschitz constant of the Lipschitz-continuous function $\phi^{(i)} : Q^{(i)} \to \mathbb{R}$, we have by (2.8) and (2.9) that

$$\begin{aligned} \left|\phi^{(i)}\left(\hat{\xi}^{(i)}\right) - \xi_n^{(i)}\right| &= |x - y| \\ &\le |x - z| + |y - z| \\ &\le s + \left|\hat{\xi}^{(i)} - \hat{\eta}^{(i)}\right| + \left|\phi^{(i)}\left(\hat{\xi}^{(i)}\right) - \phi^{(i)}\left(\hat{\eta}^{(i)}\right)\right| \\ &\le s + s + L_i \left|\hat{\xi}^{(i)} - \hat{\eta}^{(i)}\right| \le (2 + L_i)s. \end{aligned}$$

The arbitrariness of x now implies that

$$\{x \in \Omega : \operatorname{dist}(x, \partial\Omega) \le s\} \subseteq \bigcup_{i=1}^{m} \left\{ \left(\hat{\xi}^{(i)}, \xi_n^{(i)}\right) : \hat{\xi}^{(i)} \in \hat{Q}^{(i)},\ \phi^{(i)}(\hat{\xi}^{(i)}) - (2+L_i)s \le \xi_n^{(i)} \le \phi^{(i)}\left(\hat{\xi}^{(i)}\right) \right\}.$$

Consequently, we have

$$\begin{aligned} |\{x \in \Omega : \operatorname{dist}(x, \partial\Omega) \le s\}| &\le \sum_{i=1}^{m} \int_{\hat{Q}^{(i)}} \left[\phi^{(i)}\left(\hat{\xi}^{(i)}\right) - \left(\phi^{(i)}\left(\hat{\xi}^{(i)}\right) - (2+L_i)s \right) \right] d\hat{\xi}^{(i)} \\ &\le C_0 s, \end{aligned}$$

where $C_0 = \sum_{i=1}^{m}(2+L_i)|Q^{(i)}| > 0$, depending only on Ω, and $|Q^{(i)}|$ is the $(n-1)$-dimensional volume of $Q^{(i)}$ $(1 \le i \le m)$. □

LEMMA 2.2. *Given a bounded domain $\Omega \subset \mathbb{R}^n$ that has a Lipschitz-continuous boundary $\partial\Omega$, there exists a Lipschitz-continuous function $f : \overline{\Omega} \to \mathbb{R}$ such that*

$$f = 0 \text{ on } \partial\Omega, \quad \int_\Omega f\,dx = 0, \quad \text{and} \quad -\infty < \int_\Omega \log|f|\,dx < \infty.$$

Proof. Let $x_0 \in \Omega$ and $\rho > 0$ be such that the ball $B := B(x_0, \rho) = \{x \in \mathbb{R}^n : |x - x_0| < \rho\}$ is completely contained in Ω, i.e., $\overline{B} \subset \Omega$. For any $x \in \overline{\Omega}$, let $d(x)$ be the distance from x to the compact set $\partial\Omega \cup \partial B$ and define $f : \overline{\Omega} \to \mathbb{R}$ by

$$f(x) = \begin{cases} d(x) & \text{if } x \in \overline{\Omega} \setminus B, \\ -\gamma d(x) & \text{if } x \in B, \end{cases} \tag{2.10}$$

where $\gamma = \int_{\Omega\setminus B} d(x)\,dx / \int_B d(x)\,dx > 0$. Clearly, $f : \overline{\Omega} \to \mathbb{R}$ is continuous, $f = 0$ on $\partial\Omega$, and $\int_\Omega f(x)\,dx = 0$.

We show now that $f : \overline{\Omega} \to \mathbb{R}$ is Lipschitz-continuous. Fix $x, y \in \overline{\Omega}$. If both x and y are in B or both x and y are in $\overline{\Omega} \setminus B$, then we have by (2.6) and (2.10) that

$$|f(x) - f(y)| \le \max(1, \gamma)|x - y| \le (1+\gamma)|x - y|. \tag{2.11}$$

Assume now $x \in B$ but $y \in \overline{\Omega} \setminus B$. Choose $\delta \in \mathbb{R}$ so that $0 < \delta < 1$, the ball $B_1 := B(x_0, \rho + \delta) \subset \Omega$, and

$$0 < 2\delta < \operatorname{dist}(\partial\Omega, \overline{B}) := \inf_{x' \in \partial\Omega, y' \in \overline{B}} |x' - y'|. \tag{2.12}$$

If $y \in \overline{\Omega} \setminus B_1$, then $|x - y| \ge \delta$. Hence,

$$|f(x) - f(y)| \le \frac{2\max_{z\in\overline{\Omega}} |f(z)|}{\delta} |x - y|. \tag{2.13}$$

If $y \in B_1 \setminus B$, then

$$d(y) = \operatorname{dist}(y, \partial B) \le |x - y|.$$

Also,

$$d(x) = \operatorname{dist}(x, \partial B) \le |x - y|.$$

Thus,

$$|f(x) - f(y)| = \gamma d(x) + d(y) \le (1+\gamma)|x-y|. \tag{2.14}$$

Setting

$$L = \max\left(1+\gamma, \frac{2\max_{z\in\overline{\Omega}}|f(z)|}{\delta}\right) > 0,$$

we obtain from (2.11), (2.13), and (2.14) that

$$|f(x) - f(y)| \le L|x-y|.$$

Since $x, y \in \overline{\Omega}$ are arbitrary, the function $f : \overline{\Omega} \to \mathbb{R}$ is Lipschitz-continuous.

We show finally that $-\infty < \int_\Omega \log|f|\,dx < \infty$. Since $|f|$ is bounded from above on $\overline{\Omega}$,

$$\begin{aligned}\int_\Omega \log|f|\,dx &= \int_{\{x\in\Omega:|f(x)|<1\}} \log|f|\,dx + \int_{\{x\in\Omega:|f(x)|\ge 1\}} \log|f|\,dx \\ &\le \int_{\{x\in\Omega:|f(x)|\ge 1\}} \log|f|\,dx < \infty.\end{aligned}$$

So, we need only to show that

$$\int_\Omega \log|f|\,dx = \int_{B_1} \log|f|\,dx + \int_{\Omega\setminus B_1} \log|f|\,dx > -\infty, \tag{2.15}$$

where $B_1 = B(x_0, \rho+\delta) \subset \Omega$ is the same ball used before and $\delta > 0$ is given in (2.12).

Using (2.10) and (2.12), the polar coordinates, and a change of variables, we obtain

$$\begin{aligned}\int_{B_1} \log|f|\,dx &= \int_B \log(\gamma d(x))\,dx + \int_{B_1\setminus B} \log d(x)\,dx \\ &= |B|\log\gamma + \int_B \log|\rho - |x-x_0|\,|\,dx + \int_{B_1\setminus B} \log|\rho - |x-x_0|\,|\,dx \\ &= |B|\log\gamma + S_n\int_0^\rho r^{n-1}\log|\rho - r|\,dr + S_n\int_\rho^{\rho+\delta} r^{n-1}\log|\rho-r|\,dr \\ &> -\infty, \end{aligned} \tag{2.16}$$

where S_n is the surface area of the unit ball in $\mathbb{R}^n$.

Observe that for $x \in \Omega\setminus B_1$ with $d(x) < \delta$, we have by (2.12) that in fact $d(x) = \operatorname{dist}(x, \partial\Omega)$. Thus, by Lemma 2.1, there exists an integer $N \ge 1$ and a constant $C_0 > 0$ such that $|\omega_j| \le C_0\delta 2^{-j}$ $(j = N, \dots)$, where

$$\begin{aligned}\omega_j &:= \{x \in \Omega\setminus B_1 : 2^{-(j+1)}\delta < d(x) \le 2^{-j}\delta\} \\ &\subset \{x\in\Omega : \operatorname{dist}(x,\partial\Omega) \le 2^{-j}\delta\}, \qquad j = 0, \dots .\end{aligned}$$

Setting $E_\delta = \{x \in \Omega\setminus B_1 : d(x) > \delta\}$, we see that $\Omega\setminus B_1$ is the union of the pairwise

disjoint sets E_δ and ω_j $(j=0,\dots)$. Therefore, by the fact that $0<\delta<1$, we obtain

$$\begin{aligned}
\int_{\Omega\setminus B_1}\log|f|\,dx &= \int_{E_\delta}\log d(x)\,dx+\sum_{j=0}^{\infty}\int_{\omega_j}\log d(x)\,dx\\
&\geq |\Omega\setminus B_1|\log\delta+\sum_{j=0}^{N}|\omega_j|\log\left(2^{-(j+1)}\delta\right)+\sum_{j=N+1}^{\infty}|\omega_j|\log\left(2^{-(j+1)}\delta\right)\\
&\geq |\Omega\setminus B_1|\log\delta+|\Omega\setminus B_1|\log\left(2^{-(N+1)}\delta\right)\\
&\quad+C_0\delta\sum_{j=N+1}^{\infty}2^{-j}\log\left(2^{-(j+1)}\delta\right)\\
&= |\Omega\setminus B_1|\log\left(2^{-(N+1)}\delta^2\right)+C_0\delta\sum_{j=N+1}^{\infty}2^{-j}\left[\log\delta-(j+1)\log 2\right]\\
&> -\infty. \qquad (2.17)
\end{aligned}$$

Finally, (2.15) follows from (2.16) and (2.17). □

We are now ready to prove our main result in this section.

Proof of Theorem 2.1.

(1) Fix $\varepsilon\in(0,1]$. Recall the Poincaré inequality [1, 6, 8]

$$\|u\|\leq C_0\|\nabla u\| \qquad \forall u\in\mathcal{H}(\Omega,\mathbb{R}^m), \tag{2.18}$$

where $C_0>0$ is a constant depending only on Ω. Since $(1/s)\log(1+s)\to 0$ as $s\to\infty$, there exists $R_\varepsilon=R_\varepsilon(\Omega)>0$ such that

$$\log(1+s)\leq\frac{\varepsilon^2 s}{2C_0^2} \qquad \forall s\geq R_\varepsilon. \tag{2.19}$$

By (2.19) and (2.18), we have

$$\begin{aligned}
I_\varepsilon(u) &= \frac{\varepsilon^2}{2}\fint_\Omega|\nabla u|^2dx-\frac{1}{2|\Omega|}\int_{\{x\in\Omega:|u|^2\leq R_\varepsilon\}}\log(1+|u|^2)\,dx\\
&\quad-\frac{1}{2|\Omega|}\int_{\{x\in\Omega:|u|^2>R_\varepsilon\}}\log(1+|u|^2)\,dx\\
&\geq \frac{\varepsilon^2}{2}\fint_\Omega|\nabla u|^2dx-\frac{1}{2|\Omega|}\int_{\{x\in\Omega:|u|^2\leq R_\varepsilon\}}\log(1+R_\varepsilon)\,dx\\
&\quad-\frac{\varepsilon^2}{4C_0^2|\Omega|}\int_{\{x\in\Omega:|u|^2>R_\varepsilon\}}|u|^2dx\\
&\geq \frac{\varepsilon^2}{2}\fint_\Omega|\nabla u|^2dx-\frac{1}{2}\log(1+R_\varepsilon)-\frac{\varepsilon^2}{4C_0^2}\fint_\Omega|u|^2dx\\
&\geq \frac{\varepsilon^2}{4}\fint_\Omega|\nabla u|^2dx-\frac{1}{2}\log(1+R_\varepsilon) \qquad \forall u\in\mathcal{H}(\Omega,\mathbb{R}^m). \qquad (2.20)
\end{aligned}$$

Set $\mu_\varepsilon=\inf_{u\in\mathcal{H}(\Omega,\mathbb{R}^m)}I_\varepsilon(u)$. By (2.20), $\mu_\varepsilon>-\infty$. Let $\{u_j\}_{j=1}^\infty$ be an infimizing sequence of $I_\varepsilon:\mathcal{H}(\Omega,\mathbb{R}^m)\to\mathbb{R}$. It follows from (2.20) and (2.18) that $\{u_j\}_{j=1}^\infty$ is bounded in $\mathcal{H}(\Omega,\mathbb{R}^m)$. Thus, up to a subsequence, $u_j\rightharpoonup u_\varepsilon$ in $H^1(\Omega,\mathbb{R}^m)$ and

$u_j \to u_\varepsilon$ in $L^2(\Omega,\mathbb{R}^m)$ as $j \to \infty$ for some $u_\varepsilon \in H^1(\Omega,\mathbb{R}^m)$, where the symbol $\rightharpoonup$ and $\to$ denote the weak and strong convergence, respectively. We have in fact $u_\varepsilon \in \mathcal{H}(\Omega,\mathbb{R}^m)$, since $\mathcal{H}(\Omega,\mathbb{R}^m)$ is a closed subspace, hence a weakly closed subset, of $H^1(\Omega,\mathbb{R}^m)$.

For each $j \ge 1$, $|\nabla u_j|^2 + |\nabla u_\varepsilon|^2 \ge 2\nabla u_j \cdot \nabla u_\varepsilon$ in Ω, where the matrix dot-product is defined by $A \cdot B = \sum_{i=1}^m \sum_{j=1}^n A_{ij}B_{ij}$ for all $A = (A_{ij}), B = (B_{ij}) \in \mathbb{R}^{m\times n}$. Thus, by the weak convergence $u_j \rightharpoonup u_\varepsilon$ in $H^1(\Omega,\mathbb{R}^m)$, we have

(2.21)
$$\liminf_{j\to\infty} \fint_\Omega |\nabla u_j|^2 dx \ge \liminf_{j\to\infty} \left[2\fint_\Omega \nabla u_j \cdot \nabla u_\varepsilon \, dx - \fint_\Omega |\nabla u_\varepsilon|^2 dx \right] = \fint_\Omega |\nabla u_\varepsilon|^2 dx.$$

By the fact that $\log(1+s) \le s$ for all $s \ge 0$ and the Cauchy–Schwarz inequality, we imply from the strong convergence $u_j \to u_\varepsilon$ in $L^2(\Omega,\mathbb{R}^m)$ that

$$\begin{aligned} &\left| \fint_\Omega \left[\log(1+|u_j|^2) - \log(1+|u_\varepsilon|^2) \right] dx \right| = \left| \fint_\Omega \log\left(1 + \frac{|u_j|^2 - |u_\varepsilon|^2}{1+|u_\varepsilon|^2} \right) dx \right| \\ &\le \fint_\Omega \log\left(1 + \left| \frac{|u_j|^2 - |u_\varepsilon|^2}{1+|u_\varepsilon|^2} \right| \right) dx \le \fint_\Omega \left| \frac{|u_j|^2 - |u_\varepsilon|^2}{1+|u_\varepsilon|^2} \right| dx \\ &\le (\|\!| u_j |\!\| + \|\!| u_\varepsilon |\!\|) \|\!| u_j - u_\varepsilon |\!\| \to 0 \qquad \text{as } j \to \infty. \end{aligned} \tag{2.22}$$

This and (2.21) thus imply that

$$\mu_\varepsilon = \liminf_{j\to\infty} I_\varepsilon(u_j) \ge \fint_\Omega \left[\frac{\varepsilon^2}{2} |\nabla u_\varepsilon|^2 - \frac{1}{2}\log(1+|u_\varepsilon|^2) \right] dx = I_\varepsilon(u_\varepsilon) \ge \mu_\varepsilon,$$

leading to (2.1).

(2) Let $u_\varepsilon \in \mathcal{H}(\Omega,\mathbb{R}^m)$ be a minimizer of $I_\varepsilon : \mathcal{H}(\Omega,\mathbb{R}^m) \to \mathbb{R}$. The first variation of I_ε at u_ε then vanishes:

$$\delta I_\varepsilon(u_\varepsilon)(v) = \fint_\Omega \left(\varepsilon^2 \nabla u_\varepsilon \cdot \nabla v - \frac{u_\varepsilon v}{1+u_\varepsilon^2} \right) dx = 0 \qquad \forall v \in \mathcal{H}(\Omega,\mathbb{R}^m).$$

Choosing $v = u_\varepsilon$, we obtain that

$$\fint_\Omega |\nabla u_\varepsilon|^2 \le \frac{1}{\varepsilon^2}. \tag{2.23}$$

This and the Poincaré inequality (2.18) imply that

$$\fint_\Omega |u_\varepsilon|^2 \le \frac{C_0^2}{\varepsilon^2}. \tag{2.24}$$

Since the function $-\log(\cdot)$ is convex, Jensen's inequality and (2.24) then imply that

$$\begin{aligned} \mu_\varepsilon = I_\varepsilon(u_\varepsilon) &\ge -\frac{1}{2}\fint_\Omega \log(1+|u_\varepsilon|^2)\, dx \ge -\frac{1}{2}\log(1+\|\!| u_\varepsilon |\!\|^2) \\ &\ge -\frac{1}{2}\log\left(1 + \frac{C_0^2}{\varepsilon^2}\right) = -\frac{1}{2}\log\left(\varepsilon^2 + C_0^2\right) + \log\varepsilon \ge C_1 + \log\varepsilon, \end{aligned} \tag{2.25}$$

where $C_1 = -(1/2)\log(1+C_0^2)$.

Let $f : \overline{\Omega} \to \mathbb{R}$ be the Lipschitz-continuous function constructed in Lemma 2.2. Let $e_1 = (1, 0, \dots, 0) \in \mathbb{R}^m$ be the unit vector along the x_1-axis. Define $\hat{u}_\varepsilon = (f/\varepsilon)e_1$. Clearly, $\hat{u}_\varepsilon \in \mathcal{H}(\Omega, \mathbb{R}^m)$. Moreover,

$$\mu_\varepsilon \le I_\varepsilon(\hat{u}_\varepsilon) = \fint_\Omega \left[\frac{1}{2}|\nabla f|^2 - \frac{1}{2}\log\left(1 + \frac{|f|^2}{\varepsilon^2}\right)\right] dx \le C_2 + \log\varepsilon, \tag{2.26}$$

where

$$C_2 := \fint_\Omega \left(\frac{1}{2}|\nabla f|^2 - \log|f|\right) dx$$

is finite by Lemma 2.2. Now, (2.2) follows from (2.25) and (2.26).

Recall for each $\varepsilon \in (0, 1]$ that the renormalized energy functional, defined in (1.6), is

$$\begin{aligned} J_\varepsilon(v) &= I_\varepsilon\left(\frac{v}{\varepsilon}\right) - \log\varepsilon \\ &= \fint_\Omega \left[\frac{1}{2}|\nabla v|^2 - \frac{1}{2}\log\left(\varepsilon^2 + |v|^2\right)\right] dx \qquad \forall v \in \mathcal{H}(\Omega, \mathbb{R}^m), \end{aligned} \tag{2.27}$$

in which the variable v is scaled from the variable v/ε of the energy I_ε. It follows from (1) that for each $\varepsilon \in (0, 1]$ there exists a minimizer of $J_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$ and the minimum value of J_ε over $\mathcal{H}(\Omega, \mathbb{R}^m)$ is

$$\nu_\varepsilon := \min_{v \in \mathcal{H}(\Omega, \mathbb{R}^m)} J_\varepsilon(v) = \mu_\varepsilon - \log\varepsilon. \tag{2.28}$$

Consequently, by (2.2), $\{\nu_\varepsilon\}_{0<\varepsilon\le 1}$ is bounded. Moreover, for each fixed $v \in \mathcal{H}(\Omega, \mathbb{R}^m)$, we have by (2.27) that $J_\varepsilon(v)$ increases as $\varepsilon \in (0, 1]$ decreases. Therefore, ν_ε increases as $\varepsilon \in (0, 1]$ decreases. This and the boundedness of $\{\nu_\varepsilon\}_{0<\varepsilon\le 1}$ imply that $\nu \in \mathbb{R}$ as defined in (2) of Theorem 2.1 is finite and that (2.3) holds true.

(3) Let again $u_\varepsilon \in \mathcal{H}(\Omega, \mathbb{R}^m)$ be a minimizer of $I_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$. By (2.24) and (2.23), the upper bound in (2.4) and that in (2.5) hold true with $C_4 = C_0$ and $C_6 = 1$, respectively, for all $\varepsilon \in (0, 1]$. By (2.2) and Jensen's inequality, we obtain

$$C_2 + \log\varepsilon \ge \mu_\varepsilon = I_\varepsilon(u_\varepsilon) \ge -\frac{1}{2}\fint_\Omega \log\left(1 + |u_\varepsilon|^2\right) dx \ge -\frac{1}{2}\log\left(1 + \|u_\varepsilon\|^2\right),$$

leading to the lower bound in (2.4) for all $\varepsilon \in (0, e^{-C_2}/\sqrt{2}]$ with $C_3 = e^{-C_2}/\sqrt{2} > 0$. It, together with the Poincaré inequality (2.18), also implies the lower bound in (2.5) for ε in the same range with $C_5 = C_0 C_3 > 0$. Finally, letting $\varepsilon_0 = \min(1, e^{-C_2}/\sqrt{2}) \in (0, 1]$, we obtain all the desired inequalities in (2.4) and (2.5) for all $\varepsilon \in (0, \varepsilon_0]$. ∎

Remark 2.1. In the case that $\Omega = \Pi_{i=1}^n (a_i, b_i)$ with $-\infty < a_i < b_i < \infty$ $(i = 1, \dots, n)$ and the set of admissible functions is $\{u \in H^1_{\rm per}(\Omega, \mathbb{R}^m) : \fint_\Omega u\, dx = 0\}$, the upper bound (2.26) can be obtained by replacing f by $\sin(2\pi x_1/(b_1 - a_1))$.

3. Renormalized energies and their Γ-limit. We consider in this section the convergence of the renormalized energy functionals $J_\varepsilon : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$, defined in (2.27), to the energy functional $J : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R} \cup \{-\infty, \infty\}$, defined in (1.7).

THEOREM 3.1. *Let $\Omega \subset \mathbb{R}^n$ and $\mathcal{H}(\Omega, \mathbb{R}^m)$ be the same as in Theorem* 2.1. *We have*

$$-\infty < \inf_{v \in \mathcal{H}(\Omega, \mathbb{R}^m)} J(v) < \infty. \tag{3.1}$$

Moreover, there exists $v \in \mathcal{H}(\Omega, \mathbb{R}^m)$ *such that*

$$J(v) = \inf_{w \in \mathcal{H}(\Omega, \mathbb{R}^m)} J(w). \tag{3.2}$$

THEOREM 3.2. *Let* $\Omega \subset \mathbb{R}^n$ *and* $\mathcal{H}(\Omega, \mathbb{R}^m)$ *be the same as in Theorem* 2.1. *Let* $\{\varepsilon_j\}_{j=1}^{\infty}$ *be a decreasing sequence in* $(0, 1]$ *such that* $\lim_{j\to\infty} \varepsilon_j = 0$. *Then, the sequence of functionals* $J_{\varepsilon_j} : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$ $(j = 1, \dots)$ Γ*-converge with respect to the weak topology of* $\mathcal{H}(\Omega, \mathbb{R}^m)$ *to the functional* $J : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R} \cup \{\infty\}$, *i.e., the following hold true:*

(1) *If* $v_j \rightharpoonup v$ *in* $\mathcal{H}(\Omega, \mathbb{R}^m)$, *then*

$$\liminf_{j\to\infty} J_{\varepsilon_j}(v_j) \geq J(v); \tag{3.3}$$

(2) *For any* $w \in \mathcal{H}(\Omega, \mathbb{R}^m)$, *there exist* $w_j \in \mathcal{H}(\Omega, \mathbb{R}^m)$ $(j = 1, \dots)$ *such that* $w_j \rightharpoonup w$ *in* $\mathcal{H}(\Omega, \mathbb{R}^m)$ *and*

$$\lim_{j\to\infty} J_{\varepsilon_j}(w_j) = J(w).$$

THEOREM 3.3. *Let* $\Omega \subset \mathbb{R}^n$ *and* $\mathcal{H}(\Omega, \mathbb{R}^m)$ *be the same as in Theorem* 2.1. *Let* $\{\varepsilon_j\}_{j=1}^{\infty}$ *be a decreasing sequence in* $(0, 1]$ *such that* $\lim_{j\to\infty} \varepsilon_j = 0$. *For each integer* $j \geq 1$, *let* $v_j \in \mathcal{H}(\Omega, \mathbb{R}^m)$ *be a minimizer of* $J_{\varepsilon_j} : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$. *Then, there is a subsequence* $\{v_{j_i}\}_{i=1}^{\infty}$ *of* $\{v_j\}_{j=1}^{\infty}$ *and* $v \in \mathcal{H}(\Omega, \mathbb{R}^m)$ *that satisfy the following properties:*

(1) *As* $i \to \infty$, $v_{j_i} \to v$ (*strong convergence*) *in* $\mathcal{H}(\Omega, \mathbb{R}^m)$;
(2) $J(v) = \min_{w \in \mathcal{H}(\Omega, \mathbb{R}^m)} J(w)$;
(3) $\lim_{i\to\infty} \min_{w \in \mathcal{H}(\Omega, \mathbb{R}^m)} J_{\varepsilon_{j_i}}(w) = \min_{w \in \mathcal{H}(\Omega, \mathbb{R}^m)} J(w)$.

COROLLARY 3.1. *Let* $\Omega \subset \mathbb{R}^n$ *and* $\mathcal{H}(\Omega, \mathbb{R}^m)$ *be the same as in Theorem* 2.1.

(1) *Let* $\{\varepsilon_j\}_{j=1}^{\infty}$ *be a decreasing sequence in* $(0, 1]$ *such that* $\lim_{j\to\infty} \varepsilon_j = 0$. *For each integer* $j \geq 1$, *let* $u_j \in \mathcal{H}(\Omega, \mathbb{R}^m)$ *be a minimizer of* $I_{\varepsilon_j} : \mathcal{H}(\Omega, \mathbb{R}^m) \to \mathbb{R}$. *Then, there is a subsequence* $\{\varepsilon_{j_i} u_{j_i}\}_{i=1}^{\infty}$ *of* $\{\varepsilon_j u_j\}_{j=1}^{\infty}$ *and* $v \in \mathcal{H}(\Omega, \mathbb{R}^m)$ *that satisfy the following properties:* (i) *As* $i \to \infty$, $\varepsilon_{j_i} u_{j_i} \to v$ (*strong convergence*) *in* $\mathcal{H}(\Omega, \mathbb{R}^m)$; *and* (ii) $J(v) = \min_{w \in \mathcal{H}(\Omega, \mathbb{R}^m)} J(w)$.

(2) *We have*

$$\lim_{\varepsilon\to 0^+} \left(\min_{u \in \mathcal{H}(\Omega, \mathbb{R}^m)} I_\varepsilon(u) - \log \varepsilon \right) = \sup_{0 < \varepsilon \leq 1} \left(\min_{u \in \mathcal{H}(\Omega, \mathbb{R}^m)} I_\varepsilon(u) - \log \varepsilon \right) = \min_{w \in \mathcal{H}(\Omega, \mathbb{R}^m)} J(w).$$

We need several lemmas to prove our results.

LEMMA 3.1. *Let* $E \subset \mathbb{R}^n$ *be Lebesgue measurable with* $0 < |E| < \infty$. *Suppose* $g_j \to g$ *in* $L^1(E)$ *and* $\left\{\int_E \log |g_j|\, dx\right\}_{j=1}^{\infty}$ *is bounded. Then,* $\log |g| \in L^1(E)$ *and*

$$\liminf_{j\to\infty} \left(-\int_E \log |g_j|\, dx \right) \geq -\int_E \log |g|\, dx. \tag{3.4}$$

Proof. By the fact that $\log s \le (1/e)s$ for all $s > 0$, we have for each integer $j \ge 1$ that

$$\begin{aligned}\int_E |\log|g_j|\,|\,dx &= \int_{\{x\in E:|g_j(x)|\ge 1\}} \log|g_j|\,dx - \int_{\{x\in E:|g_j(x)|<1\}} \log|g_j|\,dx \\ &= 2\int_{\{x\in E:|g_j(x)|\ge 1\}} \log|g_j|\,dx - \int_E \log|g_j|\,dx \\ &\le \frac{2}{e}\int_E |g_j|\,dx - \int_E \log|g_j|\,dx. \end{aligned} \tag{3.5}$$

Since both $\{\int_E |g_j|\,dx\}_{j=1}^\infty$ and $\{\int_E \log|g_j|\,dx\}_{j=1}^\infty$ are bounded, we thus have

$$\sup_{j\ge 1}\int_E |\log|g_j|\,|\,dx < \infty. \tag{3.6}$$

Since $g_j \to g$ in $L^1(E)$, there exists a subsequence $\{g_{j_i}\}_{i=1}^\infty$ of $\{g_j\}_{j=1}^\infty$ such that $g_{j_i}(x) \to g(x)$ as $i \to \infty$ for a.e. $x \in E$. Consequently, by Fatou's Lemma and (3.6),

$$0 \le \int_E |\log|g|\,|\,dx = \int_E \liminf_{i\to\infty} |\log|g_{j_i}|\,|\,dx \le \liminf_{i\to\infty}\int_E |\log|g_{j_i}|\,|\,dx < \infty. \tag{3.7}$$

This implies that $\log|g| \in L^1(E)$, and, in particular, $|\{x \in E : g(x) = 0\}| = 0$.

For any $\sigma \in (0,1)$, we denote $S_\sigma = \{x \in E : 0 < |g(x)| \le \sigma\}$ and $m_\sigma = |S_\sigma|$. Since $|\log|g|\,| \ge |\log\sigma|$ on S_σ for any $\sigma \in (0,1)$, we have by (3.7) that

$$m_\sigma = \int_{S_\sigma} dx \le \int_{S_\sigma} \frac{|\log|g|\,|}{|\log\sigma|}\,dx \le \frac{1}{|\log\sigma|}\int_E |\log|g|\,|\,dx \to 0 \qquad \text{as } \sigma \to 0^+. \tag{3.8}$$

Thus, by (3.7), (3.8), and the absolute continuity of Lebesgue integrals, we obtain

$$\left|\int_{S_\sigma} \log|g|\,dx\right| \le \int_{S_\sigma} |\log|g|\,|\,dx \to 0 \qquad \text{as } \sigma \to 0^+. \tag{3.9}$$

Now, for each integer $j \ge 1$, we have by the fact that $-\log(\cdot)$ is convex and Jensen's inequality that

$$\begin{aligned} -\int_{S_\sigma} \log|g_j|\,dx &= -m_\sigma \fint_{S_\sigma} \log|g_j|\,dx \ge -m_\sigma \log\left(\fint_{S_\sigma} |g_j|\,dx\right) \\ &\ge m_\sigma \log m_\sigma - m_\sigma \log\left(\max_{i\ge 1}\|g_i\|_{L^1(E)}\right), \end{aligned} \tag{3.10}$$

in which $\max_{i\ge1}\|g_i\|_{L^1(E)} > 0$, since $g_i \to g$ in $L^1(E)$ as $i \to \infty$ and $\|g\|_{L^1(E)} \ne 0$. Thus, by (3.10) and (3.8),

$$\begin{aligned} \liminf_{j\to\infty}\left(-\int_{S_\sigma} \log|g_j|\,dx\right) &\ge m_\sigma \log m_\sigma - m_\sigma \log\left(\max_{i\ge1}\|g_i\|_{L^1(E)}\right) \\ &\to 0 \qquad \text{as } \sigma \to 0^+. \end{aligned} \tag{3.11}$$

Let $\delta > 0$. By (3.9) and (3.11), there exits $\sigma_0 \in (0,1)$ such that

$$\liminf_{j\to\infty}\left(-\int_{S_{\sigma_0}} \log|g_j|\,dx\right) \ge -\int_{S_{\sigma_0}} \log|g|\,dx - \delta. \tag{3.12}$$

Denoting $T_0 = \{x \in E : |g(x)| > \sigma_0\}$, we have by the fact that $\log(1+s) \le s$ for any $s \ge 0$ that

$$\begin{aligned}
&\left| \int_{T_0} \log |g_j| \, dx - \int_{T_0} \log |g| \, dx \right| \le \int_{T_0} \left| \log \frac{|g_j|}{|g|} \right| dx \\
&\le \int_{T_0} \log \left(1 + \left| \frac{|g_j| - |g|}{|g|} \right| \right) dx \le \int_{T_0} \frac{|g_j - g|}{\sigma_0} dx \\
&\le \frac{1}{\sigma_0} \|g_j - g\|_{L^1(E)} \to 0 \qquad \text{as } j \to \infty. \qquad (3.13)
\end{aligned}$$

It follows from (3.12) and (3.13) that

$$\liminf_{j\to\infty} \left(- \int_E \log |g_j| \, dx \right) \ge - \int_E \log |g| \, dx - \delta,$$

which implies (3.4) by the arbitrariness of $\delta > 0$. □

LEMMA 3.2. *Let $E \subset \mathbb{R}^n$ be Lebesgue measurable with $0 < |E| < \infty$ and $h \in L^1(E)$. Let $\{\varepsilon_j\}_{j=1}^\infty$ be a decreasing sequence in $(0,1]$ such that $\varepsilon_j \to 0$ as $j \to \infty$. Then,*

$$\lim_{j\to\infty} \int_E \log \sqrt{\varepsilon_j^2 + |h|^2} \, dx = \int_E \log |h| \, dx. \qquad (3.14)$$

Proof. Suppose first that $\int_E \log |h| \, dx = -\infty$. Set

$$\zeta_j = \int_E \log \sqrt{\varepsilon_j^2 + |h|^2} \, dx \qquad j = 1, \dots .$$

Then, $\{\zeta_j\}_{j=1}^\infty$ is a decreasing sequence. Thus, either $\lim_{j\to\infty} \zeta_j = -\infty$, leading to (3.14) in this case; or $\lim_{j\to\infty} \zeta_j$ exists and is finite. Suppose the latter were true. Then, $\{\zeta_j\}_{j=1}^\infty$ would be bounded from below. By the fact that $\log s \le (1/e)s$ for any $s > 0$, we have for any $j \ge 1$ that

$$\zeta_j \le \int_E \log \sqrt{1 + |h|^2} \, dx \le \frac{1}{e} \int_E \sqrt{1 + |h|^2} \, dx \le \frac{1}{e} \int_E (1 + |h|) \, dx < \infty. \qquad (3.15)$$

Thus, the sequence $\{\zeta_j\}_{j=1}^\infty$ is also bounded from above. In addition, $\sqrt{\varepsilon_j^2 + |h|^2} \to |h|$ in $L^1(E)$ as $j \to \infty$. Therefore, by Lemma 3.1, $\int_E \log |h| \, dx$ would be finite, leading to a contradiction in this case.

Suppose now that $\int_E \log |h| \, dx > -\infty$. Replacing g_j by h in (3.5), we obtain that $\log |h| \in L^1(E)$. By (3.15), $\log \sqrt{1 + |h|^2} \in L^1(E)$. Since for each $j \ge 1$,

$$\log |h| \le \log \sqrt{\varepsilon_j^2 + |h|^2} \le \log \sqrt{1 + |h|^2} \qquad \text{a.e. } E,$$

we thus obtain (3.14) in this case by Lebesgue's Dominated Convergence Theorem. □

Proof of Theorem 3.1. Let $\tau = \inf_{v \in \mathcal{H}(\Omega, \mathbb{R}^m)} J(v)$. Let $f : \overline{\Omega} \to \mathbb{R}$ be the Lipschitz-continuous function constructed in Lemma 2.2. Define $\hat{v} : \overline{\Omega} \to \mathbb{R}^m$ by $\hat{v} = f e_1$, where $e_1 = (1, 0, \dots, 0) \in \mathbb{R}^m$. Then, $\hat{v} \in \mathcal{H}(\Omega, \mathbb{R}^m)$ and $J(\hat{v}) < \infty$. Thus, $\tau < \infty$.

Since $(1/s)\log s \to 0$ as $s \to \infty$, there exists $R = R(\Omega) > 1$ such that

$$\log s \le \frac{s}{2C_0^2} \qquad \forall s \ge R, \tag{3.16}$$

where $C_0 > 0$ is the constant in the Poincaré inequality (2.18). Consequently, by (3.16) and the Poincaré inequality (2.18),

$$\begin{aligned}
J(v) &\ge \frac{1}{2}\fint_\Omega |\nabla v|^2 dx - \frac{1}{|\Omega|}\int_{\{x\in\Omega:|v|\ge 1\}} \log|v|\,dx \\
&= \frac{1}{2}\fint_\Omega |\nabla v|^2 dx - \frac{1}{2|\Omega|}\int_{\{x\in\Omega:1\le|v|^2\le R\}} \log\left(|v|^2\right)\,dx \\
&\quad - \frac{1}{2|\Omega|}\int_{\{x\in\Omega:|v|^2>R\}} \log\left(|v|^2\right)dx \\
&\ge \frac{1}{2}\fint_\Omega |\nabla v|^2 dx - \frac{1}{2|\Omega|}\int_{\{x\in\Omega:1\le|v|^2\le R\}} \log R\,dx - \frac{1}{4|\Omega|C_0^2}\int_{\{x\in\Omega:|v|^2>R\}} |v|^2 dx \\
&\ge \frac{1}{2}\fint_\Omega |\nabla v|^2 dx - \frac{1}{4C_0^2}\fint_\Omega |v|^2 dx - \frac{1}{2}\log R \\
&\ge \frac{1}{4}\fint_\Omega |\nabla v|^2 dx - \frac{1}{2}\log R \qquad \forall v \in \mathcal{H}(\Omega,\mathbb{R}^m).
\end{aligned} \tag{3.17}$$

This implies that $\tau > -\infty$. Hence, (3.1) is proved.

Let $\{v_j\}_{j=1}^\infty$ be an infimizing sequence of $J : \mathcal{H}(\Omega,\mathbb{R}^m) \to \mathbb{R}\cup\{\infty\}$. It follows from (3.17) and (2.18) that $\{v_j\}_{j=1}^\infty$ is bounded in $\mathcal{H}(\Omega,\mathbb{R}^m)$. Thus, up to a subsequence, $v_j \rightharpoonup v$ in $H^1(\Omega,\mathbb{R}^m)$ and $v_j \to v$ in $L^2(\Omega,\mathbb{R}^m)$ as $j \to \infty$ for some $v \in H^1(\Omega,\mathbb{R}^m)$. We have $v \in \mathcal{H}(\Omega,\mathbb{R}^m)$, since $\mathcal{H}(\Omega,\mathbb{R}^m)$ is weakly closed in $H^1(\Omega,\mathbb{R}^m)$.

As in the proof of Theorem 2.1, cf. (2.21), we have

$$\liminf_{j\to\infty} \fint_\Omega |\nabla v_j|^2 dx \ge \fint_\Omega |\nabla v|^2 dx. \tag{3.18}$$

Since $\{J(v_j)\}_{j=1}^\infty$ and $\{\|v_j\|\}_{j=1}^\infty$ are both bounded, the sequence

$$\left\{\int_\Omega \log|v_j|\,dx\right\}_{j=1}^\infty = \left\{\int_\Omega \frac{1}{2}|\nabla v_j|^2 dx - |\Omega| J(v_j)\right\}_{j=1}^\infty$$

is bounded. Thus, by Lemma 3.1, $\log|v| \in L^1(\Omega)$ and

$$\liminf_{j\to\infty}\left(-\fint_\Omega \log|v_j|\,dx\right) \ge -\fint_\Omega \log|v|\,dx. \tag{3.19}$$

Now, (3.2) follows from (3.18), (3.19), and the fact that $\{v_j\}_{j=1}^\infty$ is an infimizing sequence of $J : \mathcal{H}(\Omega,\mathbb{R}^m) \to \mathbb{R}\cup\{\infty\}$. □

Remark 3.1. In the case that $\Omega = \Pi_{i=1}^n (a_i, b_i)$ with $-\infty < a_i < b_i < \infty$ $(i = 1,\dots,n)$ and the set of admissible functions is $\{u \in H^1_{\rm per}(\Omega,\mathbb{R}^m) : \fint_\Omega u\,dx = 0\}$, we can still prove that $\tau < \infty$ by the same argument with f replaced by $\sin(2\pi x_1/(b_1 - a_1))$.

Proof of Theorem 3.2.

(1) Suppose $v_j \rightharpoonup v$ in $\mathcal{H}(\Omega,\mathbb{R}^m)$. We may assume that $\liminf_{j\to\infty} J_{\varepsilon_j}(v_j) < \infty$, otherwise (3.3) holds true trivially.

Notice that for each integer $j \geq 1$,

$$J_{\varepsilon_j}(v_j) \geq \fint_\Omega \left[\frac{1}{2}|v_j|^2 - \frac{1}{2}\log\left(1+|v_j|^2\right)\right] dx = I_1(v_j). \tag{3.20}$$

Thus, by (2.20) with $\varepsilon = 1$, the sequence $\{J_{\varepsilon_j}(v_j)\}_{j=1}^\infty$ is bounded from below. Let $\{v_{j_i}\}_{i=1}^\infty$ be a subsequence of $\{v_j\}_{j=1}^\infty$ such that $v_{j_i} \to v$ in $L^2(\Omega,\mathbb{R}^m)$ as $i\to\infty$ and

$$\liminf_{j\to\infty} J_{\varepsilon_j}(v_j) = \lim_{i\to\infty} J_{\varepsilon_{j_i}}(v_{j_i}) < \infty. \tag{3.21}$$

Then, the sequence $\left\{J_{\varepsilon_{j_i}}(v_{j_i})\right\}_{i=1}^\infty$ is bounded, from both above and below. Hence,

$$\left\{\int_\Omega \log\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2}\,dx\right\}_{i=1}^\infty = \left\{\int_\Omega \frac{1}{2}|\nabla v_{j_i}|^2 dx - |\Omega| J_{\varepsilon_{j_i}}(v_{j_i})\right\}_{i=1}^\infty$$

is bounded. Moreover, $\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2} \to |v|$ in $L^2(\Omega)$ as $i\to\infty$, since

$$\left|\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2}-|v|\right|^2 = \varepsilon_{j_i}^2+|v_{j_i}|^2+|v|^2-2|v|\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2} \leq \varepsilon_{j_i}^2+|v_{j_i}-v|^2 \qquad \forall i\geq 1.$$

Consequently, it follows from Lemma 3.1 that $\log|v| \in L^1(\Omega)$ and

$$\liminf_{i\to\infty}\left(-\fint_\Omega \log\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2}dx\right) \geq -\fint_\Omega \log|v|\,dx. \tag{3.22}$$

As before, we also have

$$\liminf_{j\to\infty}\fint_\Omega |\nabla v_j|^2 dx \geq \fint_\Omega |\nabla v|^2 dx, \tag{3.23}$$

cf. (2.21) and (3.18). Now, (3.3) follows from (3.21)–(3.23).

(2) Let $w\in\mathcal{H}(\Omega,\mathbb{R}^m)$ and $w_j = w$ for all integers $j\geq 1$. The assertion of this part follows from Lemma 3.2. □

Proof of Theorem 3.3. For each integer $j\geq 1$, $\varepsilon_j v_j$ is a minimizer of $I_{\varepsilon_j}: \mathcal{H}(\Omega,\mathbb{R}^m)\to\mathbb{R}$. Thus, by (3) of Theorem 2.1, $\{v_j\}_{j=1}^\infty$ is bounded in $\mathcal{H}(\Omega,\mathbb{R}^m)$. Hence, it has a subsequence $\{v_{j_i}\}_{i=1}^\infty$ such that $v_{j_i} \rightharpoonup v$ in $\mathcal{H}(\Omega,\mathbb{R}^m)$, $v_{j_i}\to v$ in $L^2(\Omega,\mathbb{R}^m)$, and $v_{j_i}(x)\to v(x)$ for a.e. $x\in\Omega$ as $i\to\infty$ for some $v\in\mathcal{H}(\Omega,\mathbb{R}^m)$.

Since v_{j_i} is a minimizer of $J_{\varepsilon_{j_i}}:\mathcal{H}(\Omega,\mathbb{R}^m)\to\mathbb{R}$, we have for any $w'\in\mathcal{H}(\Omega,\mathbb{R}^m)$ that

$$J_{\varepsilon_{j_i}}(w') \geq J_{\varepsilon_{j_i}}(v_{j_i}) = \min_{w\in\mathcal{H}(\Omega,\mathbb{R}^m)} J_{\varepsilon_{j_i}}(w) \qquad \forall i\geq 1.$$

Consequently, we have by Lemma 3.2 and (1) of Theorem 3.2 that

$$J(w') = \lim_{i\to\infty} J_{\varepsilon_{j_i}}(w') \geq \limsup_{i\to\infty} J_{\varepsilon_{j_i}}(v_{j_i}) \geq \liminf_{i\to\infty} J_{\varepsilon_{j_i}}(v_{j_i}) \geq J(v). \tag{3.24}$$

This proves (2). Setting $w' = v$ in (3.24), we obtain that

$$\lim_{i\to\infty}\min_{w\in\mathcal{H}(\Omega,\mathbb{R}^m)} J_{\varepsilon_{j_i}}(w) = \lim_{i\to\infty} J_{\varepsilon_{j_i}}(v_{j_i}) = J(v) = \min_{w\in\mathcal{H}(\Omega,\mathbb{R}^m)} J(v), \tag{3.25}$$

proving (3).

Notice that the sequence

$$\left\{\int_\Omega \log\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2}\,dx\right\}_{i=1}^{\infty}=\left\{\int_\Omega \frac{1}{2}|\nabla v_{j_i}|^2dx-|\Omega|J_{\varepsilon_{j_i}}(v_{j_i})\right\}_{i=1}^{\infty}$$

is bounded. Moreover, $\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2}\to|v|$ in $L^2(\Omega)$ as $i\to\infty$. Consequently, we have by Lemma 3.1 that $\log|v|\in L^1(\Omega)$ and

$$\liminf_{i\to\infty}\left(-\int_\Omega \log\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2}\,dx\right)\geq-\int_\Omega\log|v|\,dx. \tag{3.26}$$

Since $v_{j_i}\rightharpoonup v$ in $\mathcal{H}(\Omega,\mathbb{R}^m)$ as $i\to\infty$, we also have (cf. (2.21))

$$\liminf_{i\to\infty}\fint_\Omega|\nabla v_{j_i}|^2dx\geq\fint_\Omega|\nabla v|^2dx. \tag{3.27}$$

Now, it follows from (3.25)–(3.27) that

$$\begin{aligned}0&=\lim_{i\to\infty}J_{\varepsilon_{j_i}}(v_{j_i})-J(v)\\&\geq\left[\liminf_{i\to\infty}\fint_\Omega\frac{1}{2}|\nabla v_{j_i}|^2dx-\fint_\Omega\frac{1}{2}|\nabla v|^2dx\right]\\&\qquad+\left[\liminf_{i\to\infty}\left(-\fint_\Omega\log\sqrt{\varepsilon_{j_i}^2+|v_{j_i}|^2}\,dx\right)-\left(-\fint_\Omega\log|v|\,dx\right)\right]\\&\geq 0,\end{aligned}$$

which, together with (3.26) and (3.27), implies that

$$\liminf_{i\to\infty}\fint_\Omega|\nabla v_{j_i}|^2dx=\fint_\Omega|\nabla v|^2dx.$$

Thus,

$$\liminf_{i\to\infty}\int_\Omega|\nabla v_{j_i}-\nabla v|^2dx=\liminf_{i\to\infty}\int_\Omega\left(|\nabla v_{j_i}|^2+|\nabla v|^2-2\nabla v_{j_i}\cdot\nabla v\right)dx=0.$$

This and the Poincaré inequality (2.18) imply the strong convergence of a subsequence of $\{v_{j_i}\}_{i=1}^\infty$ to v in $\mathcal{H}(\Omega,\mathbb{R}^m)$. Thus (1) is proved. □

Proof of Corollary 3.1. Notice for any integer $j\geq1$ that u_j is a minimizer of $I_{\varepsilon_j}:\mathcal{H}(\Omega,\mathbb{R}^m)\to\mathbb{R}$ if and only if that $v_j:=\varepsilon_j u_j$ is a minimizer of $J_{\varepsilon_j}:\mathcal{H}(\Omega,\mathbb{R}^m)\to\mathbb{R}$. Thus, (1) follows from (1) and (2) of Theorem 3.3; (3) follows from (2) of Theorem 2.1 and (3) of Theorem 3.3. □

4. The scalar Dirichlet boundary-value problem. If the order parameter is a scalar function that satisfies the homogeneous Dirichlet boundary condition, then the solution of the corresponding limiting variational problem of infimizing $J:H_0^1(\Omega)\to\mathbb{R}\cup\{\infty\}$ can be well characterized. In what follows, we denote

$$\mathcal{H}_+(\Omega)=\left\{v\in H_0^1(\Omega):v(x)\geq0\text{ a.e. }x\in\Omega\right\}.$$

As usual, we also denote by $C_c^\infty(\Omega)$ the set of all $C^\infty(\Omega)$-functions that are compactly supported in Ω.

THEOREM 4.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with a Lipschitz-continuous boundary $\partial\Omega$. Then, there exists $v_+ \in \mathcal{H}_+(\Omega)$ that satisfies the following properties:*

(1) *The function v_+ is the unique minimizer of $J : \mathcal{H}_+(\Omega) \to \mathbb{R} \cup \{\infty\}$. Moreover, $v_+ \in C^\infty(\Omega)$, $v_+ > 0$ in Ω,*

$$\Delta v_+ + \frac{1}{v_+} = 0 \qquad \text{in } \Omega, \tag{4.1}$$

and

$$\fint_\Omega |\nabla v_+|^2 dx = 1; \tag{4.2}$$

(2) *The two functions v_+ and $v_- := -v_+$ are the unique minimizers of $J : H_0^1(\Omega) \to \mathbb{R} \cup \{\infty\}$.*

To prove this theorem, we need the following result.

LEMMA 4.1. *Let $\Omega \subset \mathbb{R}^n$ be the same as in Theorem 4.1. Let $\varepsilon \in (0, 1]$. There exists $v_{\varepsilon+} \in \mathcal{H}_+(\Omega)$ such that*

$$J_\varepsilon(v_{\varepsilon+}) = \min_{w \in \mathcal{H}_+(\Omega)} J_\varepsilon(w) = \min_{w \in H_0^1(\Omega)} J_\varepsilon(w). \tag{4.3}$$

Moreover, $\Delta v_{\varepsilon+} \le 0$ in Ω in the sense of distributions, i.e.,

$$\int_\Omega \nabla v_{\varepsilon+} \cdot \nabla \varphi \, dx \ge 0 \qquad \forall \varphi \in C_c^\infty(\Omega) \text{ with } \varphi \ge 0 \text{ in } \Omega, \tag{4.4}$$

and

$$\fint_\Omega |\nabla v_{\varepsilon+}|^2 dx = \fint_\Omega \frac{v_{\varepsilon+}^2}{\varepsilon^2 + v_{\varepsilon+}^2} \, dx. \tag{4.5}$$

Proof. Setting $\xi_\varepsilon = \inf_{w \in \mathcal{H}_+(\Omega)} J_\varepsilon(w)$, we have by Theorem 2.1 that

$$-\infty < \min_{w \in H_0^1(\Omega)} I_\varepsilon(w) - \log \varepsilon = \min_{w \in H_0^1(\Omega)} J_\varepsilon(w) \le \xi_\varepsilon \le J_\varepsilon(0) = -\log \varepsilon < \infty.$$

Let $\{w_j\}_{j=1}^\infty \subset \mathcal{H}_+(\Omega)$ be an infimizing sequence of $J_\varepsilon : \mathcal{H}_+(\Omega) \to \mathbb{R}$. Since

$$J_\varepsilon(w_j) \ge \fint_\Omega \left[\frac{1}{2} |\nabla w_j|^2 - \frac{1}{2} \log \left(1 + |w_j|^2\right) \right] dx = I_1(w_j) \qquad \forall j \ge 1,$$

we see from (2.20) with $\mathbb{R}^m = \mathbb{R}$ and $\varepsilon = 1$, and the Poincaré inequality, that the sequence $\{w_j\}_{j=1}^\infty$ is bounded in $H^1(\Omega)$. Thus, it has a subsequence $\{w_{j_i}\}_{i=1}^\infty$ such that $w_{j_i} \rightharpoonup v_{\varepsilon+}$ in $H^1(\Omega)$ and $w_{j_i} \to v_{\varepsilon+}$ in $L^2(\Omega)$ as $i \to \infty$ for some $v_{\varepsilon+} \in H^1(\Omega)$. We have in fact $v_+ \in \mathcal{H}_+(\Omega)$, since $\mathcal{H}_+(\Omega)$ is convex and strongly closed, and hence weakly closed, in $H^1(\Omega)$. Noting that $\varepsilon > 0$ is fixed, by the same argument in the proof of Theorem 2.1, cf. (2.21) and (2.22), we obtain that $J_\varepsilon(v_{\varepsilon+}) = \xi_\varepsilon$.

For any $w \in H_0^1(\Omega)$, we have $|w| \in \mathcal{H}_+(\Omega)$ and $J_\varepsilon(|w|) = J_\varepsilon(w)$, cf. Lemmas 7.6 and 7.7 in [8]. Thus,

$$\min_{w \in H_0^1(\Omega)} J_\varepsilon(w) \le \xi_\varepsilon = J_\varepsilon(v_{\varepsilon+}) \le J_\varepsilon(|w|) = J_\varepsilon(w) \qquad \forall w \in H_0^1(\Omega).$$

This leads to (4.3).

Since $v_{\varepsilon+}$ is a minimizer of $J_\varepsilon : H_0^1(\Omega) \to \mathbb{R}$, the first variation of J_ε at $v_{\varepsilon+}$ vanishes, i.e.,

$$\delta J(v_{\varepsilon+})(\varphi) = \int_\Omega \left(\nabla v_{\varepsilon+} \cdot \nabla \varphi - \frac{v_{\varepsilon+}\varphi}{\varepsilon^2 + v_{\varepsilon+}^2} \right) dx = 0 \qquad \forall \varphi \in H_0^1(\Omega). \tag{4.6}$$

This, together with the fact that $v_{\varepsilon+} \geq 0$ a.e. Ω and that $C_c^\infty(\Omega) \subset H_0^1(\Omega)$, implies (4.4). Finally, setting $\varphi = v_{\varepsilon+}$ in (4.6), we obtain (4.5). □

Proof of Theorem 4.1.

(1) By Lemma 4.1, there exists $v_j \in \mathcal{H}_+(\Omega)$ for each integer $j \geq 1$ such that

$$J_{1/j}(v_j) = \min_{w \in \mathcal{H}_+(\Omega)} J_{1/j}(w) = \min_{w \in H_0^1(\Omega)} J_{1/j}(w), \tag{4.7}$$

$$\Delta v_j \leq 0 \qquad \text{in } \Omega \tag{4.8}$$

in the sense of distributions, cf. (4.4), and

$$⨍_\Omega |\nabla v_j|^2 dx = ⨍_\Omega \frac{v_j^2}{j^{-2} + v_j^2}\, dx. \tag{4.9}$$

By (4.7) and Theorem 3.3, there exists a subsequence $\{v_{j_i}\}_{i=1}^\infty$ of $\{v_j\}_{j=1}^\infty$ such that $v_{j_i} \to v_+$ (strong convergence) in $H^1(\Omega)$ and $v_{j_i}(x) \to v_+(x)$ a.e. $x \in \Omega$ as $i \to \infty$ for some $v_+ \in H_0^1(\Omega)$ that is a minimizer of $J : H_0^1(\Omega) \to \mathbb{R} \cup \{\infty\}$. Since $\mathcal{H}_+(\Omega)$ is weakly closed in $H^1(\Omega)$, $v_+ \in \mathcal{H}_+(\Omega)$. Moreover, since $\mathcal{H}_+(\Omega) \subset H_0^1(\Omega)$, v_+ is also a minimizer of $J : \mathcal{H}_+(\Omega) \to \mathbb{R} \cup \{\infty\}$. The fact that v_+ is the unique minimizer of $J : \mathcal{H}_+(\Omega) \to \mathbb{R} \cup \{\infty\}$ follows from the strict convexity of $J : \mathcal{H}_+(\Omega) \to \mathbb{R} \cup \{\infty\}$.

Let $\eta_+ = \operatorname{ess\,inf}_\Omega v_+$. Since $v_+ \in \mathcal{H}_+(\Omega)$, $v_+ \geq 0$ a.e. Ω. Thus, $\eta_+ \geq 0$. If $\eta_+ > 0$, then there exists $\phi_j \in C^\infty(\overline{\Omega})$ for each integer $j \geq 1$ such that $\min_{\overline{\Omega}} \phi_j \geq \eta_+/2$ $(j = 1, \dots)$ and $\phi_j \to v_+$ in $H^1(\Omega)$ as $j \to \infty$. The trace of v_+, which is the limit of $\{\phi_j\}_{j=1}^\infty$ in $L^2(\partial\Omega)$, would then be positive a.e. $\partial\Omega$. This contradicts the fact that $v_+ \in H_0^1(\Omega)$. Thus, $\operatorname{ess\,inf}_\Omega v_+ = 0$. Since $v_+ \geq 0$ is a minimizer of $J : \mathcal{H}_+(\Omega) \to \mathbb{R} \cup \{\infty\}$, $\log v_+ \in L^1(\Omega)$. Thus, we also have that $v_+(x) > 0$ a.e. $x \in \Omega$. In particular, v_+ is not a constant in Ω. Since $v_{j_i} \to v_+$ in $H^1(\Omega)$ as $i \to \infty$, we obtain by (4.8) that $\Delta v_+ \geq 0$ in Ω in the sense of distributions. Applying the Strong Maximum Principle to $L = \Delta$ and $u = v_+$ in Theorem 8.19 in [8], we see that $\operatorname{ess\,inf}_B v_+ > 0$ for any ball $B \subset\subset \Omega$. (Here and below, the notation $\omega \subset\subset \Omega$ means $\overline{\omega} \subset \Omega$.) For any open set $\Omega' \subset \mathbb{R}^n$ with $\Omega' \subset\subset \Omega$, we can cover $\overline{\Omega'}$ by finitely many balls $B_i \subset\subset \Omega$, where $i = 1, \dots, N$ for some integer $N \geq 1$, so that $\operatorname{ess\,inf}_{B_i} v_+ > 0$ for all i $(1 \leq i \leq N)$. Thus, $\delta' := \operatorname{ess\,inf}_{\Omega'} v_+ > 0$, and hence $1/v_+ \in L^\infty(\Omega')$.

Let $\varphi \in C_c^\infty(\Omega)$ with $\operatorname{supp} \varphi \subset \Omega'$ and consider $q(\delta) := J(v_+ + \delta\varphi)$ for $\delta \in \mathbb{R}$. If $|\delta| \sup_{\Omega'} |\varphi| < \delta'$, then $v_+ + \delta\varphi > 0$ a.e. in Ω', and

$$\begin{aligned} q(\delta) &= ⨍_\Omega \left[\frac{1}{2} |\nabla v_+ + \delta\varphi|^2 - \log |v_+ + \delta\varphi| \right] dx \\ &= ⨍_\Omega \frac{1}{2} |\nabla v_+ + \delta\varphi|^2 dx - \frac{1}{|\Omega|} \int_{\Omega \setminus \Omega'} \log v_+ \, dx - \frac{1}{|\Omega|} \int_{\Omega'} \log(v_+ + \delta\varphi) dx. \end{aligned}$$

Since v_+ is a minimizer of $J : H_0^1(\Omega) \to \mathbb{R} \cup \{\infty\}$, $q(\delta)$ is minimized at $\delta = 0$. Thus, $q'(0) = 0$. This leads to

$$\Delta v_+ + \frac{1}{v_+} = 0 \qquad \text{in } \Omega' \tag{4.10}$$

in the sense of distributions.

Let $\Omega'' \subset \mathbb{R}^n$ be an open set such that $\Omega'' \subset\subset \Omega' \subset\subset \Omega$. Denoting $H^k(D) = W^{k,2}(D)$ the usual Sobolev space for an open set $D \subset \mathbb{R}^n$ and an integer $k \geq 1$ [1, 8], we claim:

($*$) For any integer $k \geq 1$, there exists an open set $\Omega_k \subset \mathbb{R}^n$ such that $\Omega'' \subset\subset \Omega_k \subset\subset \Omega'$ and $v_+ \in H^{k+1}(\Omega_k)$.

Let $f := 1/v_+ \in L^\infty(\Omega')$. By (4.10), $v_+ \in H^1(\Omega')$ is a weak solution of $\Delta v_+ = f$ in Ω', i.e.,

$$\int_{\Omega'} \nabla v_+ \cdot \nabla \varphi \, dx = \int_{\Omega'} f\varphi \, dx \qquad \forall \varphi \in C_c^\infty(\Omega'). \tag{4.11}$$

By the regularity theory of elliptic boundary-value problems, cf. Theorem 8.8 in [8], we have $v_+ \in H^2(\Omega_1)$ for any open set $\Omega_1 \subset \mathbb{R}^n$ such that $\Omega'' \subset\subset \Omega_1 \subset\subset \Omega'$. Thus, the statement ($*$) is true for $k = 1$.

Suppose the statement ($*$) is also true for a general $k \geq 1$. Then, $\partial^k f \in H^1(\Omega_k)$ for any partial derivative ∂^k of order k. Replacing φ in (4.11) by $\partial^k \psi$ for any $\psi \in C_c^\infty(\Omega_k)$, one easily verifies that

$$\int_{\Omega_k} \nabla \partial^k v_+ \cdot \nabla \psi \, dx = \int_{\Omega_k} \partial^k f \psi \, dx \qquad \forall \psi \in C_c^\infty(\Omega_k),$$

i.e., $\partial^k v_+ \in H^1(\Omega_k)$ satisfies $\Delta \partial^k v_+ = \partial^k f$ in Ω_k in the sense of distributions. Therefore, by the same regularity result, there exists an open set $\Omega_{k+1} \subset \mathbb{R}^n$ with $\Omega'' \subset\subset \Omega_{k+1} \subset\subset \Omega_k \subset\subset \Omega'$ such that $\partial^k v_+ \in H^2(\Omega_{k+1})$. By suitably enlarging Ω_{k+1} if necessary, we see that $v_+ \in H^{k+2}(\Omega_{k+1})$. Hence, the statement ($*$) is true for $k+1$. Thus, it is true for any integer $k \geq 1$.

By the statement ($*$), $v_+ \in C^\infty(\Omega'')$. It then follows from the arbitrariness of Ω'' and Ω' that $v_+ \in C^\infty(\Omega)$, and that $v_+ > 0$ in Ω, since $\operatorname{ess\,sup}_{\Omega'} v_+ > 0$ for any $\Omega' \subset\subset \Omega$. Moreover, (4.1) follows from (4.10) and the arbitrariness of $\Omega' \subset\subset \Omega$. Using the fact that $v_{j_i} \to v_+$ in $H^1(\Omega)$ and that $v_{j_i}(x) \to v_+(x) > 0$ a.e. Ω, and applying Lebesgue's Dominated Convergence Theorem, we obtain (4.2) from (4.9).

(2) Clearly, both v_+ and v_- are minimizers of $J : H_0^1(\Omega) \to \mathbb{R} \cup \{\infty\}$, cf. the proof of Lemma 4.1. Assume now $\tilde{v} \in H_0^1(\Omega)$ is a minimizer of $J : H_0^1(\Omega) \to \mathbb{R} \cup \{\infty\}$. Then, $|\tilde{v}| \in \mathcal{H}_+(\Omega)$ is a minimizer of $J : \mathcal{H}_+(\Omega) \to \mathbb{R} \cup \{\infty\}$. By (1), we must have that $|\tilde{v}| = v_+$ a.e. Ω. Thus, $|\tilde{v}| \in C^\infty(\Omega)$ and $\tilde{v} > 0$ in Ω. Consequently, we have for any ball $B \subset\subset \Omega$ that $\tilde{v}(x) > 0$ for all $x \in B$ or $\tilde{v}(x) < 0$ for all $x \in B$. Therefore, for any domain $\omega \subset\subset \Omega$, $\tilde{v}(x)$ has the same sign for each $x \in \bar{\omega}$. This implies that $\tilde{v}(x) = v_+(x)$ for all $x \in \Omega$ or $\tilde{v}(x) = v_-(x)$ for all $x \in \Omega$. □

Acknowledgments. The author thanks the referees for helpful comments.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

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

[3] M. Cyrot, *Ginzburg-Landau theory for superconductors*, Rep. Progr. Phys., 36 (1973), pp. 103–158.

[4] G. Ehrlich and F. G. Hudda, *Atomic view of surface diffusion: Tungsten on tungsten*, J. Chem. Phys., 44 (1966), p. 1036.

[5] H.-J. Ernst, F. Fabre, R. Folkerts, and J. Lapujoulade, *Observation of a growth instability during low temperature molecular beam epitaxy*, Phys. Rev. Lett., 72 (1994), pp. 112–115.

[6] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[7] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.
[8] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss., Springer-Verlag, Berlin, 1998.
[9] V. L. Ginzburg and L. D. Landau, Zh. Eksp. Teor. Fiz., 20 (1950), p. 1064.
[10] L. Golubović, *Interfacial coarsening in epitaxial growth models without slope selection*, Phys. Rev. Lett., 78 (1997), pp. 90–93.
[11] A. W. Hunt, C. Orme, D. R. M. Williams, B. G. Orr, and L. M. Sander, *Instabilities in MBE growth*, Europhys. Lett., 27 (1994), p. 611.
[12] M. D. Johnson, C. Orme, A. W. Hunt, D. Graff, J. Sudijono, L. M. Sander, and B. G. Orr, *Stable and unstable growth in molecular beam epitaxy*, Phys. Rev. Lett., 72 (1994), pp. 116–119.
[13] B. Li and J.-G. Liu, *Epitaxial growth without slope selection: Energetics, coarsening, and dynamic scaling*, J. Nonlinear Sci., 14 (2004), pp. 429–451.
[14] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, New York, 2002.
[15] P. Politi and A. Torcini, *Coarsening in surface growth models without slope selection*, J. Phys. A, 33 (2000), pp. L77–L82.
[16] M. Rost and J. Krug, *Coarsening of surface structures in unstable epitaxial growth*, Phys. Rev. E, 55 (1997), pp. 3952–3957.
[17] R. L. Schwoebel and E. J. Shipsey, *Step motion on crystal surfaces*, J. Appl. Phys., 37 (1966), p. 3682.
[18] R. L. Schwoebel, *Step motion on crystal surfaces II*, J. Appl. Phys., 40 (1969), p. 614.
[19] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, 1999.
[20] K. Thürmer, R. Koch, M. Weber, and K. H. Rieder, *Dynamic evolution of pyramid structures during growth of epitaxial Fe(001) films*, Phys. Rev. Lett., 75 (1995), pp. 1767–1770.
[21] J. Villain, *Continuum models of crystal growth from atomic beams with and without desorption*, J. de Phys. I, 1 (1991), pp. 19–42.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 37–63

EXISTENCE OF SOLUTIONS FOR A CLASS OF IMPACT PROBLEMS WITHOUT VISCOSITY*

JEONGHO AHN† AND DAVID E. STEWART‡

Abstract. In this paper we consider dynamic frictionless impact problems of elastic materials formulated in abstract settings. The contact conditions for the impact problem are Signorini-type complementarity conditions. Using time discretization and Galerkin approximation, we investigate the convergence of numerical fully discrete trajectories to a solution of the continuous-time problem. In this way we establish the existence of solutions for a class of impact problems, some of which have been previously studied, while others have not. Most of the impact problems to which this theory can be applied are "thick" obstacle problems, although it can also be applied to a number of boundary or "thin" obstacle problems. The crucial assumption for the theory is that the cone of possible contact forces satisfies a strong pointedness condition, which can usually be related to a Sobolev embedding condition.

Key words. impact problems, dynamic obstacle problems, complementarity conditions, strongly pointed cones

AMS subject classifications. Primary, 35L85; Secondary, 74M15

DOI. 10.1137/S0036141004444664

1. Introduction.

In this work, we show existence of solutions for a class of dynamic impact problems which are treated from an abstract point of view. The concrete impact problems that we discuss and that are in this class are mostly "thick" obstacle problems; the solutions are functions of time $t > 0$ and $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ for some d where Ω is an open domain, and the obstacle (or constraints) are applied over a subset $\Omega_1 \subset \overline{\Omega_1} \subset \Omega$. However, it is applicable to a small number of boundary contact problems. Unfortunately these do not include the wave equation with a boundary "thin" obstacle for which existence has been shown [14] or conservation of energy where Ω is a half-space [19]; the still-open case of a nonviscous elastic body with boundary contact requires further techniques beyond those described here.

In most papers on contact mechanics, the problem is cast as a variational inequality in terms of the main variable, the solution $u(t)$. This is the approach used in [9, 10, 16, 19], for example. As a result, the contact forces $N(t)$ are removed from the formulation, and their behavior remains somewhat difficult to elucidate. In this paper, we keep the contact forces as part of a complementarity formulation. In some ways this makes our approach more complicated as we now need to obtain suitable bounds on the normal contact forces. However, it means that more is known about the contact forces.

There is another aspect where this paper differs from most others on existence of solutions in contact mechanics: we are dealing with nonviscous elasticity. Thus, without contact, energy is conserved. The introduction of visco-elasticity is useful for improving the regularity of the solution; however, it makes the contact forces less

*Received by the editors June 22, 2004; accepted for publication (in revised form) August 3, 2005; published electronically March 31, 2006.
http://www.siam.org/journals/sima/38-1/44466.html.
†Department of Mathematics, Yousei University, Seoul 120-749, Korea (jh_ahn@yousei.ac.kr). The research of this author was partially supported by the Korean BK21 project.
‡Department of Mathematics, University of Iowa, Iowa City, IA 52242 (david-e-stewart@math.uiowa.edu). The research of this author was partially supported by NSF grant DMS-01-39708.

regular [3, 22, 23]. The question of the regularity of the contact forces is important as it is closely related to the question of proving energy balances. Only if the contact forces are sufficiently regular do we expect to prove that the contact forces do no work [30]. In rigid-body dynamics the contact forces are general measures and contain Dirac-δ functions; rigid-body dynamics also requires an additional parameter, the coefficient of restitution, to determine the consequences of a single impact, and to determine if energy is (possibly) conserved.

For example, consider the simple scalar problem $u_{tt} = f(t) + N(t)$ with f an integrable function, $0 \le u(t) \perp N(t) \ge 0$ for all t. Here $\mathbf{0} \le \mathbf{a} \perp \mathbf{b} \ge \mathbf{0}$ called complementarity condition implies that $\mathbf{a}, \mathbf{b} \ge \mathbf{0}$ component-wise and $\mathbf{a}^T\mathbf{b} = 0$, where we denote vectors and matrices by bold characters. Thus $\mathbf{0} \le \mathbf{a} \perp \mathbf{b} \ge \mathbf{0}$ means that for all indexes i, either $a_i = 0$ and $b_i \ge 0$, or $a_i \ge 0$ and $b_i = 0$. If $a_i > 0$, then $b_i = 0$, and conversely if $b_i > 0$, then $a_i = 0$. In our case $u(t)$ and $N(t)$ are scalars, so for all t if there is positive separation ($u(t) > 0$), there is no contact force ($N(t) = 0$); conversely, if there is a contact force ($N(t) > 0$), then there is no separation ($u(t) = 0$). Note that $u(t)$ cannot be negative (as $u(t) < 0$ would violate the no-penetration condition) while $N(t)$ cannot be negative (as $N(t) < 0$ would imply adhesion). This can be interpreted in a weak sense by requiring that u and N are nonnegative and that $\int_0^T u(t)\, N(t)\, dt = 0$.

Suppose that $u(t_*) = 0$. Then near $t = t_*$ we can set $N(t) = N_0(t) + N_* \, \delta(t - t_*)$ where N_0 is a measure without atom at t_* and δ is the Dirac-δ function. This means that $u_t(t_*^+) = u_t(t_*^-) + N_*$. Any value of N_* that exceeds $-u_t(t_*^-) \ge 0$ is permissible as then $u_t(t_*^+) > 0$ and $u(t) > 0$ for t in some interval $(t_*, t_* + \epsilon)$. For conservation of energy or an energy balance, we need $u_t(t_*^+)^2 = u_t(t_*^-)^2$; only one value of N_* will conserve energy, and yet all values of $N_* \ge -u_t(t_*^-)$ lead to solutions of the contact conditions $0 \le u(t) \perp N(t) \ge 0$. On the other hand, if $N(t) = N_0(t)$, an integrable function, then u_t is a continuous function. Thus, if $u(t_*) = 0$ in this case, $u_t(t_*) = 0$ as well, and so $\int_B u_t(t)\, N(t)\, dt = 0$ for any Borel set B, and so the contact forces do no work.

A related question is whether there should be a coefficient of restitution, as there is in rigid-body dynamics with impact. For example, the most commonly used coefficient of restitution is the Newton, or kinematic, coefficient of restitution, which is simply the negative of the ratio of the postimpact normal velocity to the preimpact normal velocity: $-u_t(t_*^+)/u_t(t_*^-)$.

Unlike the finite-dimensional case, the presence of unbounded operators means that the velocities do not necessarily have bounded variation, and care must be taken in dealing with the limits $u_t(t_*^+)$ and $u_t(t_*^-)$ in appropriate function spaces. However, it can be done in certain situations, such as a vibrating string (subject to the one-dimensional wave equation) which can contact a rigid obstacle along a portion of its length. The wave equation with an obstacle was first constructed by Amerio and Prouse [4] and was perfectly studied by Schatzman [26]. She was able to show the existence of solutions for a concave obstacle *which conserve energy*. An essential part of this was proving that $u_t(t^+, x) = -u_t(t^-, x)$ at almost every point of contact; that is, the kinematic coefficient of restitution is 1. However, the approach taken in [26] seems difficult to generalize to other contact problems.

For the sake of concreteness, we will discuss a one-dimensional vibrating string which satisfies the wave equation, fixed at endpoints $x = 0$ and $x = l$, and where contact can occur with a fixed obstacle over most of the length of the string. (Contact near the fixed endpoints will be assumed not to occur.)

We begin with the equations of motion for a body without contact:

$$\begin{aligned} u_{tt} &= -Au + f, \\ u(0) &= u^0, \\ u_t(0) &= v^0, \end{aligned}$$

where $(\cdot)_t$ denotes differentiation with respect to t, $A\colon V \to V'$ is an elliptic self-adjoint operator, V a Hilbert space, and f a constant vector in a suitable space. The results of this paper can be extended to nonconstant f; the main technical complication being obtaining suitable energy bounds for the time discretization. Incorporating contact conditions requires an obstacle represented by φ which is applied to some space W of restricted functions, a closed convex cone $K \subset W$, and a restriction or "trace" operator $\beta\colon V \to W$ which is used to identify when the solution u is "above" the obstacle: $\beta u - \varphi \in K$. In static (frictionless) contact problems, the normal contact force appears as a kind of Lagrange multiplier $N \in K^* \subset W'$, where K^* is the dual cone given by

$$K^* = \{z \mid \langle z, w\rangle \geq 0 \quad \text{for all } w \in K\}.$$

In our dynamic contact problems, we have "$N(t) \in K^*$ for all t", which must be suitably interpreted as N may be a W'-valued measure. This is transfered to V' via the adjoint of the restriction operator $\beta^*\colon W' \to V'$.

There should be a right inverse of the restriction operator (being an extension operator) $\gamma\colon W \to V$ so that $\beta\gamma\colon W \to W$ is the identity map. Note that this means that $\gamma^*\beta^*$ is the identity map on W'. This means that β is surjective and β^* is injective.

In the case of a vibrating string, $A = -\partial^2/\partial x^2$ which is defined on $H_0^1(0,l)$; the constraint on the string is that $u(t,x) - \varphi(x) \geq 0$ for all $x \in (\delta,\, l-\delta)$ for a suitable $\delta > 0$ to avoid contact near the fixed endpoints. Take $W = H^1(\delta,\, l-\delta)$. Our restriction operator $\beta\colon V \to W$ is simply given by $\beta u = u|_{(\delta,l-\delta)}$. The set K is then $\{w \in H_0^1(\delta,\, l-\delta) \mid w(x) \geq 0 \text{ for all } x \in (\delta,\, l-\delta)\}$, and K^* is the set of all nonnegative distributions in $H^{-1}(\delta,\, l-\delta)$ (and are therefore measures on $[0,\, l]$ [13]). Provided $\varphi(0),\, \varphi(l) < 0$, the energy boundedness of solutions ensures that there is a $\delta > 0$, where $u(t,x) - \varphi(x) > 0$ for all t and $x \notin (\delta, l-\delta)$. This problem was investigated by Schatzman [26] for φ concave.

It may seem unnecessary to have separate V and W spaces in this case; we could set $V = W$ and let K be a cone of functions on $(0,l)$. However, the extension of the results to show convergence of space and time discretizations assumes that K^* is the closure of K in the dual space. This does not hold if $V = W$ as then $K = \{w \in H^1(0,l) \mid w(x) \geq 0 \text{ for all } x \in [\delta, l-\delta]\}$ but $K^* = \{z \in H^{-1}(0,l) \mid z(x) \geq 0 \text{ for } x \in [\delta, l-\delta],\ z(x) = 0 \text{ for } x \in [0,\delta) \cup (l-\delta, l]\}$.

With contact the equations of motion become

$$\begin{aligned} u_{tt} &= -Au + \beta^* N(t) + f, \\ K \ni \beta u(t) - \varphi \perp N(t) &\in K^* \quad \text{for all } t \in [0,T], \end{aligned} \tag{1.1}$$

where (1.1) is called complementarity condition and it will be explained in Remark 3.1.

The impact problems have been studied in different physical situations using different approaches. In this paper, we consider the impact problems in a more general point of view which includes these problems studied in [26] and [1], and also the

Timoshenko beam (first described in [32, 33]) with Signorini contact conditions along the beam. A recent discussion of the Timoshenko beam with impact can be found in [20]. In section 4, we describe these and other impact problems in more detail. The abstract setting is described in section 3. The time discretization is given in section 5.1, and convergence of the discrete-time trajectories to a solution of our problem is proved in section 5.2 and also convergence of fully discrete trajectories is shown in section 5.3.

2. Preliminaries and functional spaces. The spaces we work are parts of a pair of *Gelfand triples* (see, for example, [35, section 17.1] for an extensive discussion of Gelfand triples):

$$V \subset H = H' \subset V'$$
$$W \subset Z = Z' \subset W',$$

where all spaces are separable Hilbert spaces, and all inclusions are dense and compact. The *pivot spaces* H and Z are typically L^2 spaces (although over different sets) while V and W are typically Sobolev spaces. Note that for a dual pair of spaces X and X' we use $\langle \cdot, \cdot \rangle_{X' \times X}$ or $\langle \cdot, \cdot \rangle$ to denote the duality pairing between X and X'; the latter form will be used where it is clear what the dual spaces are. Also inner product $(\cdot, \cdot)_H$ is used instead of $(\cdot, \cdot)_{H \times H}$. Note that the duality pairing in a Gelfand triple must be consistent with the inner product, i.e., for any $x \in H$ and $y \in V$, $\langle x, y \rangle_{V' \times V} = (x, y)_{H \times V} = (x, y)_H$. Thus the duality pairing $\langle \cdot, y \rangle_{V' \times V}$ is the continuous extension of the linear function $(\cdot, y)_{H \times V}$ or $(\cdot, y)_H$ over the largest possible subspace of V'. See [35, section 17] for the details.

For the example of a vibrating string, the pivot spaces are $H = L^2(0, l)$ and $Z = L^2(\delta, l - \delta)$; the duality pairing over H and V with V' is given by

$$\langle f, g \rangle = \int_0^l f(x)\, g(x)\, dx.$$

In order to state our existence results, we need scales of *interpolation spaces* V_θ and W_θ. While this can be done using either real or complex interpolation theories, we note that for Hilbert spaces these are equivalent to the following constructions. See, for example, Taylor [31, pp. 276–278] for complex interpolation, or Bramble and Zhang [7, Appendix B], Triebel [34, section 1.18.10], or Kuttler [15, section 22.6, exercises], for real interpolation theories. The real interpolation methods should use exponent 2 as appropriate for Hilbert spaces. Since $A \colon V \to V'$ is elliptic self-adjoint, the norm on V is equivalent to $\|u\|_A := \langle Au, u \rangle_{V' \times V}^{1/2}$.

For W we can use the natural map $J \colon W \to W'$ given by $J(w) = (\cdot, w)_{W' \times W}$ which is also elliptic and self-adjoint.

The spaces V_θ and W_θ are given by

$$V_\theta = \left\{ u \in H \mid \left\langle A^\theta u, u \right\rangle < +\infty \right\}, \quad \text{and}$$
$$W_\theta = \left\{ w \in Z \mid \left\langle J^\theta w, w \right\rangle < +\infty \right\}$$

which are inner product spaces with the inner products $(x, y)_{V_\theta} = \left\langle A^\theta x, y \right\rangle$ and $(w, z)_{W_\theta} = \left\langle J^\theta w, z \right\rangle$. Note that we will use $(\cdot, \cdot)_\theta$ for the inner product when it is clear which of the V_θ or W_θ spaces is being considered; similarly, we will use $\| \cdot \|_\theta$ for the norm in either V_θ or W_θ when the choice is clear.

Note that both V_θ and W_θ are well defined for any real number θ, and that $V = V_1$, $H = V_0$, and $V' = V_{-1}$; and the corresponding properties hold for the W_θ spaces. Note that $V_{\theta+\epsilon}$ embeds compactly in V_θ whenever $\epsilon > 0$. Furthermore, $(V_\theta)' = V_{-\theta}$ and $(W_\theta)' = W_{-\theta}$. Also continuous linear operators $F\colon X_0 \to Y_0$ and $F\colon X_1 \to Y_1$ can be extended to the continuous linear operators on the interpolation spaces, that is, $F\colon X_\theta \to Y_\theta$, $0 \le \theta \le 1$.

For the vibrating string problem, $V_\theta = \{u \in H^\theta(0,l) \mid u(0) = u(l) = 0\}$ for $\theta > 1/2$ and $H^\theta(0,l)$ otherwise. The reason for the difference is that the trace operators $u \mapsto u(0)$ and $u \mapsto u(l)$ fail to exist for $\theta \le 1/2$. Also, $W_\theta = H^\theta(\delta, l-\delta)$.

In order to obtain bounds on the contact forces, we make an assumption about the nature of the cones K and K^*.

DEFINITION 2.1. *A dual cone K^* is said to be* strongly pointed *if there exist $\kappa \in K$ and $\eta > 0$ such that for any $\zeta \in K^*$,*

$$\langle \zeta, \kappa \rangle_{W' \times W} \ge \eta \, \|\zeta\|_{W'} \, .$$

This terminology is first used in [29]; this property is equivalent to requiring that K has nonempty interior. Assuming that K^* is strongly pointed is not quite enough. In addition we need a "gap" in the scale of interpolation spaces: there should be a $\theta < 1$, $\kappa \in K$ and $\eta > 0$ such that

$$\langle \zeta, \kappa \rangle_{W_{-\theta} \times W_\theta} \ge \eta \|\zeta\|_{-\theta} \quad \text{for any } \zeta \in K^*. \tag{2.1}$$

Finally we need an additional property of the restriction operator β: there is an $\alpha < 1$, where $\beta\colon V_\alpha \to Z$ is a continuous map of Hilbert spaces.

In the case of the vibrating string problem, we can show that both of these conditions hold. Firstly, we can choose $\kappa(x) = x(l-x)$ which is positive and bounded away from zero on $[\delta, l-\delta]$. This means that

$$\begin{aligned} \langle \zeta, \kappa \rangle &= \int_{[\delta, l-\delta]} \zeta(x)\, \kappa(x)\, dx \\ &\ge \delta(l-\delta) \int_{[\delta, l-\delta]} \zeta(x)\, dx \\ &= \delta(l-\delta)\, \|\zeta\|_{\mathcal{M}[\delta, l-\delta]}. \end{aligned}$$

By the Sobolev embedding theorem (see [24, Thm. 6.91, p. 215] or [31, Prop. 3.3, p. 282]) $C[\delta, l-\delta] \supset H^\theta(\delta, l-\delta) = W_\theta$ as Banach spaces for $\theta > 1/2$. Thus by duality, $\mathcal{M}[\delta, l-\delta] \subset H^{-\theta}(\delta, l-\delta) = W_{-\theta}$ as Banach spaces, so there is a positive constant C_θ for $\theta > 1/2$, where $\|\zeta\|_{H^{-\theta}(\delta, l-\delta)} \le C_\theta \, \|\zeta\|_{\mathcal{M}[\delta, l-\delta]}$. Thus the strong pointedness assumption (2.1) holds for the vibrating string problem.

We assume that the initial displacement $u^0 \in V = V_1$ and the initial velocity $v^0 \in H = V_0$.

We introduce the following notation: if X and Y are Banach spaces, the set of all bounded linear operators from X to Y is denoted by $\mathcal{L}(X, Y)$, and, especially, $\mathcal{L}(X, X)$ is written as $\mathcal{L}(X)$.

3. Impact problem formulated in abstract setting. We seek a solution $u\colon [0, T] \to V$ satisfying the following conditions:

$$u_{tt}(t) = -Au(t) + \beta^* N(t) + f, \quad \text{in the sense of distributions,} \tag{3.1}$$

$$W \supset K \ni \beta\, u(t) - \varphi \perp N(t) \in K^* \subset W' \quad \text{for all } t \in [0, T], \tag{3.2}$$

$$u(0) = u^0, \tag{3.3}$$

$$u_t(0) = v^0. \tag{3.4}$$

Note that f does not depend on time t. The assumptions we will use throughout this paper are as follows:

(A1) We have the following Gelfand triples $V \subset H = H' \subset V'$ and $W \subset Z = Z' \subset W'$ with all spaces separable Hilbert spaces and all inclusions dense and compact.
(A2) The linear operator $A\colon V \to V'$ is elliptic and self-adjoint.
(A3) There is a linear operator $\beta\colon V \to W$ that is continuous and surjective, with a bounded right-inverse $\gamma\colon W \to V$, where $\beta\gamma = \mathrm{id}_W$.
(A4) The set $K \subset W$ is a closed convex cone, with dual cone $K^* \subset W'$.
(A5) $\varphi \in W$, $u^0 \in V$ and $v^0 \in H$.
(A6) There is $\eta > 0$, $\theta < 1$ and $\kappa \in K$, where $\langle \zeta, \kappa\rangle \geq \eta \|\zeta\|_{-\theta}$ for all $\zeta \in K^*$.
(A7) There is an $\alpha < 1$ where $\beta\colon V_\alpha \to Z$.

Note that (A6) is equivalent to requiring that the interior of K is nonempty in V_θ.

Our main result is the following theorem.

THEOREM 3.1. *Under assumptions* (A1)–(A7)*, for any $T > 0$ there is a solution (u, N) to* (3.1–3.4) *where u lies in $L^\infty(0,T;V) \cap W^{1,\infty}(0,T;H)$ and where N lies in $\mathcal{M}(0,T;W')$, the space of Borel measures on $[0,T]$ with bounded variation and values in W'.*

Remark 3.1.

(a) Note that (3.1) is interpreted in the sense of distributions in V', "$N(t) \in K^*$" is interpreted in the sense that N is a W'-valued measure with $N(B) \in K^*$ for any Borel set B, and "$\beta u - \varphi \perp N$" is interpreted in the sense that $\int_B \langle \beta u - \varphi, N\rangle = 0$ for any Borel set $B \subseteq [0,T]$.

(b) Note that since $\beta u(t) - \varphi \in K$ for all t and N is a K^*-valued measure, $\langle \beta u - \varphi, N\rangle \geq 0$ in the sense of measures. Thus if $\int_{[0,T]} \langle \beta u - \varphi, N\rangle = 0$, then $\int_B \langle \beta u - \varphi, N\rangle = 0$ for any Borel set $B \subseteq [0,T]$.

(c) The condition "$N(t) \in K^*$" for a measure N is equivalent to that used in [21] and [29].

4. Examples of dynamic frictionless impact problems. In this section, we list examples of concrete impact problems to which this abstract formulation can be applied. In section 1 we have seen the vibrating string problem with contact along the length of the string.

The Euler–Bernoulli beam equation with contact along the length of the beam can also be handled in this way. Another equation to which this approach can be applied is the Timoshenko beam with contact occurring along the length of the beam. Higher-dimensional systems require higher-order differential equations. For example, the biharmonic equation can be used for two- or three-dimensional systems with contact either on a subdomain of positive measure, or on the boundary.

The most difficult assumption to satisfy is (A6): there is a $\kappa \in K$, where $\langle \zeta, \kappa\rangle \geq \eta \|\zeta\|_{W_{-\theta}}$ for all $\zeta \in K^*$. In thick obstacle problems, $W = H^m(\Omega_1)$, where $\Omega_1 \subset \overline{\Omega_1} \subset \Omega \subset \mathbb{R}^d$. If $m > d/2$, then we have the Sobolev embedding $H^m(\Omega_1) \subset C(\overline{\Omega_1})$ and we can choose $\kappa(x) = 1$ for all $x \in \Omega_1$. Then by duality, $H^{-m}(\Omega_1) \supset \mathcal{M}(\overline{\Omega_1})$, the space of Borel measures on $\overline{\Omega_1}$. Provided K is a cone of nonnegative functions, K^* is the corresponding set of nonnegative measures. The norm on the space of measures is essentially the same as the L^1 norm, which is equal to the integral over $\overline{\Omega_1}$, which is $\langle \kappa, \zeta\rangle$. In boundary thin obstacle problems, $W = H^m(\Gamma_C)$, where $\Gamma_C \subset \partial\Omega \subset \mathbb{R}^{d-1}$, so our condition becomes $m > (d-1)/2$.

For the case of the wave equation or equations of elasticity on a domain $\Omega \subset \mathbb{R}^d$ for $d \geq 2$ the condition (A6) does not hold. The reasons are as described in the above

paragraph: we take $m = 1$ as we must take V to be a subspace of $H^1(\Omega)$ satisfying the essential homogeneous boundary conditions, and so the Sobolev embedding and the strong pointedness condition fail.

Now we introduce those problems in detail.

4.1. The wave equation in one dimension. The vibrating string problem has the formulation:

$$\begin{aligned} u_{tt} &= u_{xx} + \beta^* N(t,x) \quad \text{in } (0,T] \times (0,l), \\ 0 \le \beta u(t,x) - \varphi(x) \perp N(t,x) &\ge 0 \quad \text{in } (0,T] \times (0,l), \\ u(t,0) = u(t,l) &= 0 \quad \text{in } (0,T], \end{aligned}$$

where l is the length of a string and $\varphi(x)$ is an obstacle. Schatzman showed the existence of solutions for φ concave down; furthermore, she showed that the solutions she constructed conserved energy. However, this was done using the method of characteristics. An essential part of this work was showing that for the solutions constructed, $u_t(t^+, x) = -u_t(t^-, x)$ for almost everywhere of (t, x), where $u(t, x) = 0$. Thus it behaves as if it has a kinematic coefficient of restitution set equal to 1 for perfectly elastic impacts. However, there is no coefficient of restitution in the equations of motion. This problem has already been discussed in the introduction, and the precise characterization of the spaces V, W, etc., and the operators involved, can be found there.

The contact problem with the wave equation in one dimension and contact only at an endpoint is an old problem, and was discussed by Routh nearly a century and a half ago [25]. A more recent analysis can be found in [28]; a variation of this idea with a general point obstacle can be found in [27]. This can also be (essentially) treated in this framework, although the operator A is not coercive: it has one zero eigenvalue and all others are positive. Firstly, we take $V = H^1(0, l)$ and $W = \mathbb{R}$. Thus $Z = W_\theta = W_{-\theta} = \mathbb{R}$ for all θ. We take $K = \mathbb{R}_+$ to $K^* = \mathbb{R}_+$ and assumption (A6) holds. Assumption (A7) holds as the restriction operator $\beta u = u(0)$ is a bounded linear operator $\beta\colon V_\alpha = H^\alpha(0, l) \to Z = \mathbb{R}$ for $\alpha > 1/2$ by the usual Sobolev embedding theorems. Note that in this problem the operator $\beta^*\colon \mathbb{R} \to H^{-1}(0, l)$ is the operator $\beta^*(N) = N\,\delta$ where δ is the "Dirac-δ function" of the theory of distributions [12]. Such a term in the right-hand side of the wave equation

$$u_{tt} = u_{xx} + N(t)\,\delta(x) \quad \text{for } x \in (0, l)$$

requires less delicate treatment. If we integrate the above equation against a function $\omega \in H^1(0, l)$, we see that

$$\begin{aligned} \frac{d^2}{dt^2} \langle u, \omega \rangle &= \int_0^l u_{xx}\, \omega(x)\, dx + N(t)\, \omega(0) \\ &= u_x \omega(x)|_{x=0}^{x=l} - \int_0^l u_x\, \omega_x\, dx + N(t)\, \omega(0). \end{aligned}$$

Taking the end $x = l$ to be free, that is, to have the natural boundary condition $u_x(t, l) = 0$, we see that the appropriate boundary condition at $x = 0$ is $u_x(t, 0) = N(t)$. This is consistent with traction boundary conditions for a force $N(t)$ applied at the endpoint $x = 0$. The regularity of the solutions of these "point contact" problems is considerably better than for the problems with distributed contact.

4.2. Euler–Bernoulli beam equations. The problem of a Euler–Bernoulli beam that can contact a rigid frictionless obstacle along its length has been studied in [2, 1], both theoretically and numerically. If we assume that the beam is clamped at $x = 0$ but free at $x = l$, we get the system

$$\begin{aligned} u_{tt} &= -u_{xxxx} + \beta^* N(t,x) \quad \text{in } (0,T] \times (0,l), \\ 0 \le N(t,x) \perp \beta u(t,x) - \varphi(x) &\ge 0 \quad \text{in } (0,T] \times (0,l), \\ u(t,0) = u_x(t,0) &= 0 \quad \text{on } (0,T], \\ u_{xx}(t,l) = u_{xxx}(t,l) &\quad \text{in } (0,T]. \end{aligned}$$

As you can see, we have essential and natural boundary conditions. Notice that Euler–Bernoulli beam equation has a fourth-order differential operator. We assume that $-\varphi(0) > 0$ as for the vibrating string problem. This ensures that contact is kept away from the essential boundary $x = 0$.

More specifically, $A = \partial^4/\partial x^4$ while $V = H^2_{cf}(0,l)$ (the first subscript "c" indicates clamped boundary conditions as $x = 0$ and the second subscript "f" means free) and $W = H^2(\delta, l)$. Also $H = L^2(0,l)$ and $Z = L^2(\delta, l)$. As for the vibrating string example, we assume that $\beta : H \to Z$. One of the interesting results from the numerical study was that energy was close to being conserved even though the numerical formulation was designed to mimic inelastic impacts. It is possible to set up initial conditions for this problem which would result in dissipation of energy, but the numerical simulations seem to indicate that energy may be conserved for "most" initial conditions.

Boundary contact problems with Euler–Bernoulli beams have been investigated in other works, such as [5, 17], although the beam is typically assumed to be visco-elastic; that is, the equations of motion have the form

$$u_{tt} = -a u_{xxxx} - b\, u_{xxxxt} + f(t,x).$$

The inclusion of viscosity makes the solution u more regular (at least in space) but tends to make the normal contact force less regular. (See, for example, the results of Petrov and Schatzman [22, 23] and Ahn and Stewart [3] for a visco-elastic wave equation

$$u_{tt} = u_{xx} + b\, u_{xxt}$$

with contact at the endpoint $x = 0$.) While visco-elastic problems are not treated in this paper, the corresponding nonviscous problem with boundary contact can be dealt with in this framework.

4.3. Timoshenko beam formulation. The Timoshenko beam equations are more realistic than the Euler–Bernoulli beam equation, as is well known in engineering [32, 33]. Imposing frictionless Signorini's contact conditions along the length of the bar, we establish the following system:

$$\begin{aligned} \rho\mathcal{A}\psi_{tt} &= \mathcal{E}M\psi_{xx} + D\mathcal{A}G(w_x - \psi) \quad \text{in } (0,l) \times [0,T], \\ \rho\mathcal{A}w_{tt} &= D\mathcal{A}G(w_{xx} - \psi_x) + N(t,x) \quad \text{in } (0,l) \times [0,T], \\ 0 \le N(t,x) \perp \beta u(t,x) - \varphi(x) &\ge 0 \quad \text{in } (0,l) \times [0,T], \\ w(t,0) = \psi(t,0) &= 0 \quad \text{on } (0,T], \\ w_x(t,l) - \psi(t,l) &= 0 \quad \text{on } (0,T]. \end{aligned}$$

In this case $u(t,x) = [\psi(t,x),\, w(t,x)]^T \in \mathbb{R}^2$ and $V = H^1(0,l;\mathbb{R}^2)$. Our restriction operator $\beta : H^1(0,l;\mathbb{R}^2) \to H^1(\delta,l)$ is given by $\beta u(t,x) = w(t,x)$; note that $W = H^1(\delta,l)$. The pivot spaces are $H = L^2(0,l;\mathbb{R}^2)$ and $Z = L^2(\delta,l)$. The cone $K = \{w \in H^1(\delta,l) \mid w(x) \geq 0 \text{ for all } x \in (\delta,l)\}$. The quantity ψ is the angular rotation of the beam, while w is the vertical displacement. The physical constants involved in this equation are: ρ is the density of beam, $\mathcal{A}$ is the area of cross section of the beam, $\mathcal{E}$ is Young's modulus, M is the second moment of area, G is the modulus of elasticity of shear, and D is introduced to account for the geometric dependent distribution of shear stress.

This system behaves very similarly to the wave equation in impact, although the Euler–Bernoulli beam system can be obtained as a limit of this system as certain combinations of ρ, $\mathcal{A}$, $\mathcal{E}$, M, G and D go to infinity. Note that the approach of Schatzman for the vibrating string problem would be considerably complicated in this context by having two sets of characteristics to follow.

4.4. Thin plates in impact. The equations of motion for thin elastic plates are fourth-order equations of which the simplest uses the biharmonic operator. Incorporating frictionless contact forces over a subdomain $\Omega_1 \subset \overline{\Omega_1} \subset \Omega \subset \mathbb{R}^d$, again so as to avoid contact close to the essential boundary, we have the thick obstacle problem for a clamped boundary:

$$\begin{aligned} u_{tt} &= -\Delta^2 u + \beta^* N(t,\mathbf{x}) \quad \text{in } (0,T]\times\Omega,\\ 0 \leq \beta u(t,\mathbf{x}) - \varphi(\mathbf{x}) \perp N(t,\mathbf{x}) &\geq 0 \quad \text{on } (0,T]\times\Omega,\\ u(t,\mathbf{x}),\, \frac{\partial u}{\partial n}(t,\mathbf{x}) &= 0 \quad \text{on } (0,T]\times\partial\Omega. \end{aligned}$$

Specifically, $V = H_0^2(\Omega)$ and $H = L^2(\Omega)$, while $W = H^2(\Omega_1)$ and $Z = L^2(\Omega_1)$, and the measure of Ω_1 is strictly positive. This can be treated in our system provided $d \leq 3$. The main reason for this restriction is assumption (A6). As before, we note that this is connected to the Sobolev embedding $W_\theta = H^{2\theta}(\Omega_1) \subset C(\overline{\Omega_1})$ for $2\theta > d/2$. We can pick $1 > \theta > d/4$ provided $d \leq 3$. Then we can proceed as in the previous examples and show that $W_{-\theta} \supset \mathcal{M}(\overline{\Omega_1})$, and we can take $\kappa(\mathbf{x}) = 1$ for all $\mathbf{x} \in \Omega_1$ for the purposes of (A6).

A boundary thin obstacle problem based on the biharmonic operator can be treated by the approach here for $d \leq 3$. Let $\partial\Omega = \Gamma_D \cup \Gamma_C$ with clamped (Dirichlet) boundary conditions applied on Γ_D and the contact conditions applied on Γ_C. To avoid the problems of contact close to the essential boundary Γ_D we assume that $-\varphi$ is bounded above a positive constant on Γ_D. The boundary conditions that are applied are then

$$\begin{aligned} u(t,\mathbf{x}),\, \frac{\partial u}{\partial n}(t,\mathbf{x}) &= 0 \quad \text{for } t>0 \text{ and } \mathbf{x}\in\Gamma_D,\\ \Delta u(t,\mathbf{x}) &= 0 \quad \text{for } t>0 \text{ and } \mathbf{x}\in\Gamma_C,\\ \frac{\partial}{\partial n}\Delta u(t,\mathbf{x}) &= N(t,\mathbf{x}) \quad \text{for } t>0 \text{ and } \mathbf{x}\in\Gamma_C,\\ 0 \leq \beta u(t,\mathbf{x}) - \varphi(\mathbf{x}) \perp N(t,\mathbf{x}) &\geq 0 \quad \text{for } t>0 \text{ and } \mathbf{x}\in\Gamma_C. \end{aligned}$$

We take $V = \{u \in H^2(\Omega) \mid u|\Gamma_D = (\partial u/\partial n)|\Gamma_D = 0\}$ and $H = L^2(\Omega)$. The restriction operator $\beta\colon V \to W$ is the trace operator $H^2(\Omega) \to H^{3/2}(\Gamma_C)$. Since the dimension of Γ_C is $d-1$, we have a Sobolev embedding $H^{3/2}(\Gamma_C) \subset C(\Gamma_C)$ provided $(d-1)/2 < 3/2$; that is, provided $d \leq 3$. This means that for assumption (A6) we

can take $\kappa(\mathbf{x}) = 1$ for all $\mathbf{x} \in \Gamma_C$ provided $-\varphi(\mathbf{x}) > 0$ for all $\mathbf{x} \in \partial\Omega\backslash\Gamma_C$. This also guarantees that there is no contact within a certain small distance of $\Gamma_D = \partial\Omega\backslash\Gamma_C$.

A problem of this kind was treated numerically in [8]. The authors are not aware of any theoretical treatment of impact problems for plates of either the "thick" or "boundary thin" kinds.

5. Convergence of numerical formulation.

5.1. Time-discrete approximation. Using time discretization, we set up a numerical formulation. Our numerical scheme is to employ the midpoint rule for elasticity and the implicit Euler method for contact condition. In order to do so, we first partition time interval $[0, T]$:

$$0 = t_0 < t_1 < \cdots < t_{k-1} < t_k < t_{k+1} < \cdots < t_P = T,$$

where $h_t = t_{k+1} - t_k$ for $k \geq 0$ is the time step size. We will assume for analytical convenience that T is a multiple of h_t; that is, $h_t = T/P$ for some positive integer P. Then we denote by u^k a numerical solution of displacement $u(t_k)$, by v^k a numerical solution of velocity $v(t_k)$ and by N^k a numerical solution of contact force $N(t_k)$, respectively, at each discretized time $t_k = k \cdot h_t$. Now we establish the numerical formulation:

$$\frac{v^{k+1} - v^k}{h_t} = -\left(\frac{A\,u^{k+1} + A\,u^k}{2}\right) + \beta^* N^k + f \tag{5.1}$$

$$\frac{u^{k+1} - u^k}{h_t} = \frac{v^{k+1} + v^k}{2} \tag{5.2}$$

$$K \ni \beta u^{k+1} - \varphi \perp N^k \in K^*, \tag{5.3}$$

where the complementarity condition (5.3) implies $\left\langle N^k,\, \beta u^{k+1} - \varphi \right\rangle_{W_{-\theta}\times W_\theta} = 0$. Also we can write the numerical solution quantities as

$$u^k = \sum_{j=1}^{\infty} \left\langle u^k, \phi_j \right\rangle \phi_j,$$

$$v^k = \sum_{j=1}^{\infty} \left(v^k, \phi_j\right) \phi_j,$$

$$\beta^* N^k = \sum_{j=1}^{\infty} \left\langle \beta^* N^k, \phi_j \right\rangle \phi_j,$$

where ϕ_j forms an H-orthonormal basis and is an eigenfunction of an elliptic self-adjoint operator A. Now we define energy functional of the impact problems as

$$E[u,\, v] = \frac{1}{2}\left(\langle Au,\, u\rangle_{V'\times V} + (v,\, v)_H\right) - \langle f,\, u\rangle_{V'\times V}.$$

$\frac{1}{2}\langle Au,\, u\rangle$ is the elastic energy, $\frac{1}{2}(v,\, v)_H$ is the kinetic energy, and $-\langle f,\, u\rangle$ is the potential energy due to the external forces. In the case where $A = -\Delta$ with Dirichlet or Neumann boundary conditions, $\langle Au,\, u\rangle = \int_\Omega |\nabla u|^2 dx$.

Using (5.1) and (5.2), u^{k+1} and v^{k+1} can be expressed in terms of u^k and v^k in the following way:

$$\begin{bmatrix} u^{k+1} \\ v^{k+1} \end{bmatrix} = \left(I + h_t^2 A/4\right)^{-1} \left(\begin{bmatrix} I - h_t^2 A/4 & h_t \\ -h_t A & I - h_t^2 A/4 \end{bmatrix} \begin{bmatrix} u^k \\ v^k \end{bmatrix} + \begin{bmatrix} h_t^2/2 \\ h_t \end{bmatrix} \left(\beta^* N^k + f\right) \right). \tag{5.4}$$

5.2. Convergence theory in the semidiscrete case. We first show that we can solve the above complementarity problem.

LEMMA 5.1. *Provided $u^k \in V$ and $v^k \in H$, then there is a unique solution (u^{k+1}, v^{k+1}, N^k) to (5.1–5.3) with $u^{k+1} \in V$, $v^{k+1} \in H$ and $N^k \in W'$.*

Proof. Employing (5.4), the solution to (5.1–5.3) can be found by reducing it to the complementarity problem

$$K \ni (\beta u^k - \varphi) + h_t\,\beta(I + h_t^2 A/4)^{-1}\left[-\frac{h_t}{2}Au^k + v^k + \frac{h_t}{2}(\beta^* N^k + f)\right] \quad \perp \quad N^k \in K^*.$$

This complementarity problem is equivalent to the variational inequality: find $N^k \in K^*$ such that for all $N \in K^*$,

$$\left\langle (\beta u^k - \varphi) + h_t\,\beta(I + h_t^2 A/4)^{-1}\left[-\frac{h_t}{2}Au^k + v^k + \frac{h_t}{2}(\beta^* N^k + f)\right],\, N^k - N\right\rangle \le 0.$$

Since the operator $\beta(I + h_t^2 A/4)^{-1}\beta^*\colon W' \to W$ is elliptic and self-adjoint, we can apply the Lions–Stampacchia theorem [6, Thm. 3.1, pp. 24–29] to show that there is a unique $N^k \in W'$ satisfying these conditions. Noting that $\beta^* N^k \in V'$ and $(I + h_t^2 A/4)^{-1}\colon V' \to V$ by (5.4), we see that $u^{k+1} \in V$ and $v^{k+1} \in H$. ∎

In the discrete-time case, we will see in the next lemma that our numerical scheme ensures that energy does not increase.

LEMMA 5.2. *Suppose that numerical solutions satisfy* (5.1), (5.2) and (5.3). *Then energy does not increase.*

Proof. We claim that $E[u^{k+1}, v^{k+1}] \le E[u^k, v^k]$. Using (5.1) and (5.2), by the extension of $(\cdot,\cdot)_H$ on $V' \times V$ for $h_t > 0$ we have

$$\left(\frac{v^{k+1} - v^k}{h_t},\, \frac{v^{k+1} + v^k}{2}\right)_H = \left\langle -\left(\frac{A\,u^{k+1} + A\,u^k}{2}\right) + \beta^* N^k + f,\, \frac{u^{k+1} - u^k}{h_t}\right\rangle_{V' \times V}.$$

Thus we obtain

$$\begin{aligned}
&\frac{1}{2h_t}\left(\left(v^{k+1}, v^{k+1}\right)_H - \left(v^k, v^k\right)_H\right)\\
&= -\frac{1}{2h_t}\left(\left\langle A\,u^{k+1},\, u^{k+1}\right\rangle_{V' \times V} - \left\langle A\,u^k,\, u^k\right\rangle_{V' \times V}\right)\\
&\quad + \frac{1}{h_t}\left(\left\langle \beta^* N^k + f,\, u^{k+1}\right\rangle_{V' \times V} - \left\langle \beta^* N^k + f,\, u^k\right\rangle_{V' \times V}\right)\\
&= -\frac{1}{2h_t}\left(\left\langle A\,u^{k+1},\, u^{k+1}\right\rangle_{V' \times V} - \left\langle A\,u^k,\, u^k\right\rangle_{V' \times V}\right)\\
&\quad + \frac{1}{h_t}\left(\left\langle N^k,\, \beta u^{k+1}\right\rangle_{W' \times W} - \left\langle N^l,\, \beta u^k\right\rangle_{W' \times W}\right)\\
&\quad + \frac{1}{h_t}\left(\left\langle f,\, u^{k+1}\right\rangle_{V' \times V} - \left\langle f,\, u^k\right\rangle_{V' \times V}\right)\\
&= -\frac{1}{2h_t}\left(\left\langle A\,u^{k+1} - 2f,\, u^{k+1}\right\rangle_{V' \times V} - \left\langle A\,u^k - 2f,\, u^k\right\rangle_{V' \times V}\right)\\
&\quad + \frac{1}{h_t}\left(\left\langle N^k, \beta u^{k+1} - \varphi\right\rangle_{W' \times W} - \left\langle N^k, \beta u^k - \varphi\right\rangle_{W' \times W}\right).
\end{aligned}$$

Since $\left\langle N^k,\, \beta u^{k+1} - \varphi\right\rangle_{W' \times W} = 0$ and $\left\langle N^k,\, \beta u^k - \varphi\right\rangle_{W' \times W} \ge 0$ from the contact

condition (5.3),

$$E[u^{k+1}, v^{k+1}] = \frac{1}{2}\left(\langle Au^{k+1} - 2f,\, u^{k+1}\rangle_{V'\times V} + \left(v^{k+1}, v^{k+1}\right)_H\right)$$
$$\leq \frac{1}{2}\left(\langle Au^{k} - 2f,\, u^{k}\rangle_{V'\times V} + \left(v^{k}, v^{k}\right)_H\right) = E[u^k,\, v^k],$$

as required. ∎

Since we assumed that $u^0 \in V$ and $v^0 \in H$, the initial energy is finite, i.e., $E^0 := E(u^0,\, v^0) < \infty$. In the next Lemma, we shall see that approximate solutions at each time t_k are uniformly bounded. Notice that C represents a quantity that is independent of h and k, but the value of C will be different in each occurrence.

LEMMA 5.3. *For all $k \geq 0$,*

$$\|u^k\|_V \leq 2\,\|f\|_{V'} + \sqrt{2E^0}, \tag{5.5}$$

$$\|v^k\|_H \leq 2\,\|f\|_{V'} +^4 \sqrt{2E^0\,\|f\|_{V'}^2} + \sqrt{E^0}, \tag{5.6}$$

where $E^0 = E[u^0,\, v^0]$ is the initial energy.

Proof. From the Lemma 5.2, $E[u^k, v^k] \leq E[u^0, v^0] < \infty$ for each $k > 0$. Then we have

$$E^0 = E[u^0,\, v^0] \geq E[u^k, v^k]$$
$$= \frac{1}{2}\left(\langle u^k, A\,u^k - 2f\rangle + \left(v^k,\, v^k\right)\right)$$
$$= \frac{1}{2}\left\langle \sum_{j=1}^{\infty} u_j^k\phi_j, \sum_{j=1}^{\infty}\lambda_j u_j^k\phi_j \right\rangle + \frac{1}{2}\left(\sum_{j=1}^{\infty} v_j^k\phi_j, \sum_{j=1}^{\infty} v_j^k\phi_j\right) - \langle u^k,\, f\rangle$$
$$\geq \frac{1}{2}\left(\sum_{j=1}^{\infty}\lambda_j\left(u_j^k\right)^2 + \sum_{j=1}^{\infty}\left(v_j^k\right)^2\right) - \|u^k\|_V\,\|f\|_{V'}$$
$$= \frac{1}{2}\left(\left\|u^k\right\|_V^2 + \left\|v^k\right\|_H^2\right) - \|u^k\|_V\,\|f\|_{V'},$$

where $u_j^k = \langle u^k, \phi_j\rangle$ and $v_j^k = (v^k, \phi_j)$. Thus after some computations with quadratic equations, we see (5.5) and (5.6) which are independent of k or $h_t > 0$. ∎

Let the numerical trajectories $u_{h_t}(t)$ be linear continuous interpolant of $u_{h_t}(t_k) = u^k$ and $u_{h_t}(t_{k+1}) = u^{k+1}$ for $t \in [t_k, t_{k+1}]$. For the velocity, let $v_{h_t}(t)$ be a constant interpolant and $v_{h_t}(t) = v^{k+1}$ for $t \in (t_k, t_{k+1})$. Then $u_{h_t}(t) = u^k + \frac{1}{2}\int_{t_k}^{t}[v_{h_t}(\tau - h_t) + v_{h_t}(\tau)]\,d\tau$ for all $t \in (t_k, t_{k+1})$. In order to approach contact force $N \in \mathcal{M}(0, T; W')$, we first set up a step function $N(t) = N^k$ for $t \in [t_k, t_{k+1})$ on W'. This implies that contact force N is constant on W' for only time $t \in [t_k, t_{k+1})$. Thus it is easy to see that the approximate contact force N_{h_t} can be defined as

$$N_{h_t} = h_t \sum_{k=0}^{T/h_t - 1} \delta(t - (k+1)h_t)N^k. \tag{5.7}$$

So from Lemma 5.3, $u_{h_t} \in L^\infty(0, T; V_1)$ and $v_{h_t} \in L^\infty(0, T; H)$. Note that V_1 is compactly embedded in V_θ for any $\theta < 1$.

We now wish to bound the norm of the contact forces N_{h_t} in $\mathcal{M}(0, T; W_{-\theta})$ using the strong pointedness assumption (A6). However, we will find it convenient to work

with $\langle N^k, \beta(I + h_t^2 A/4)^{-1}\gamma\,\kappa\rangle$, instead of $\langle N^k, \kappa\rangle$, so we need an approximation result for κ.

LEMMA 5.4. *Suppose that $w \in W$. Then,*

$$\beta\left(I + h_t^2 A/4\right)^{-1}\gamma w \to w \quad \textit{in } W \quad \textit{as } h_t \downarrow 0.$$

Proof. Since $w \in W$, $\left(I + h_t^2 A/4\right)^{-1}\gamma w \in V$. Let $y = \gamma w$. Putting $y = \sum_j y_j \phi_j$, we have

$$\begin{aligned}\left\|\left(I + h_t^2 A/4\right)^{-1} y - y\right\|_V^2 &= \left\|A^{1/2}\left(\left(I + h_t^2 A/4\right)^{-1} y - y\right)\right\|_H^2 \\ &= \sum_{j=1}^{\infty}\left(\frac{h_t^2\lambda_j/4}{1 + h_t^2\lambda_j/4}\right)^2 \lambda_j y_j^2.\end{aligned}$$

Since $y \in V$, for any $\delta > 0$, there is an $n > 0$, independent of $h_t > 0$ such that $\sum_{j>n}\lambda_j y_j^2 < \delta$. Now for such n, we have

$$\sum_{j=1}^{n}\left(\frac{h_t^2\lambda_j/4}{1 + h_t^2\lambda_j/4}\right)^2 \lambda_j y_j^2 \to 0 \quad \text{as } h_t \downarrow 0. \tag{5.8}$$

On the other hand,

$$\sum_{j>n}^{\infty}\left(\frac{h_t^2\lambda_j/4}{1 + h_t^2\lambda_j/4}\right)^2 \lambda_j y_j^2 \le \sum_{j>n}^{\infty}\lambda_j y_j^2 < \delta. \tag{5.9}$$

Therefore from (5.8) and (5.9),

$$\left\|\left(I + h_t^2 A/4\right)^{-1} y - y\right\|_V \to 0 \quad \text{as } h_t \downarrow 0.$$

Thus as $h_t \downarrow 0$,

$$\begin{aligned}\left\|\beta\left(I + h_t^2 A/4\right)^{-1}\gamma w - w\right\|_W &= \left\|\beta\left(I + h_t^2 A/4\right)^{-1}\gamma w - \beta\gamma w\right\|_W \\ &= \left\|\beta\left(\left(I + h_t^2 A/4\right)^{-1} y - y\right)\right\|_W \\ &\le \|\beta\|_{\mathcal{L}(V,W)}\left\|\left(I + h_t^2 A/4\right)^{-1} y - y\right\|_V \to 0,\end{aligned}$$

as required. ∎

The above approximation result can now be used to bound the norm of N_h as a $W_{-\theta}$-valued measure.

LEMMA 5.5. *Suppose that* (A6) *holds with $\kappa \in \operatorname{int} K \subset W$. Then there is a constant C (independent of h) where*

$$\int_{[0,T]}\|N_{h_t}(t)\|_{W_{-\theta}}\,dt \le C \quad \textit{for all sufficiently small } h_t > 0.$$

Proof. Recall the matrix form (5.4). By extension of $\left(v^{k+1} - v^k, \gamma\kappa\right)_{H\times V}$ in $V' \times V$,

$$\begin{aligned}\left(v^{k+1} - v^k, \gamma\kappa\right)_{H\times V} &= \Big\langle \left(I + h_t^2 A/4\right)^{-1}\left(-h_t A u^k + \left(I - h_t^2 A/4\right) v^k\right. \\ &\quad \left. - \left(I + h_t^2 A/4\right) v^k + h_t(\beta^* N^k + f)\right), \gamma\kappa\Big\rangle_{V'\times V} \\ &= \left\langle \left(I + h_t^2 A/4\right)^{-1}\left(-h_t A u^k - \frac{h_t^2}{2} A v^k + h_t(\beta^* N^k + f)\right), \gamma\kappa\right\rangle_{V'\times V}.\end{aligned}$$

Thus we have

$$\begin{aligned} & h_t \left\langle \left(I + h_t^2 A/4\right)^{-1} \beta^* N^k,\, \gamma\kappa \right\rangle_{V' \times V} \\ & \quad = \left(v^{k+1} - v^k, \gamma\kappa\right)_H + h_t \left\langle \left(I + h_t^2 A/4\right)^{-1} A u^k,\, \gamma\kappa \right\rangle_{V' \times V} \\ & \quad + \frac{h_t^2}{2} \left\langle \left(I + h_t^2 A/4\right)^{-1} A v^k,\, \gamma\kappa \right\rangle_{V' \times V} - h_t \left\langle \left(I + h_t^2 A/4\right)^{-1} f,\, \gamma\kappa \right\rangle_{V' \times V}. \end{aligned} \tag{5.10}$$

Also it follows that

$$\begin{aligned} h_t \left\langle \left(I + h_t^2 A/4\right)^{-1} \beta^* N^k,\, \gamma\kappa \right\rangle_{V' \times V} &= h_t \left\langle \beta^* N^k,\, \left(I + h_t^2 A/4\right)^{-1} \gamma\kappa \right\rangle_{V' \times V} \\ &= h_t \left\langle N^k,\, \beta \left(I + h_t^2 A/4\right)^{-1} \gamma\kappa \right\rangle_{W' \times W}. \end{aligned} \tag{5.11}$$

Consider the second term of the right-hand side on (5.10). Since for $h_t > 0$ $\left\|(I + h_t^2 A/4)^{-1}\right\|_{\mathcal{L}(H)} = \sup_{\lambda_j \in \sigma(A)} (1 + h_t^2 \lambda_j/4)^{-1} \leq 1$, we have

$$\begin{aligned} \left| \left\langle \left(I + h_t^2 A/4\right)^{-1} A u^k,\, \gamma\kappa \right\rangle_{V' \times V} \right| &= \left| \left(\left(I + h_t^2 A/4\right)^{-1} A^{1/2} u^k,\, A^{1/2} \gamma\kappa \right)_H \right| \\ &\leq \left\| \left(I + h_t^2 A/4\right)^{-1} A^{1/2} u^k \right\|_H \left\| A^{1/2} \gamma\kappa \right\|_H \\ &\leq \left\| \left(I + h_t^2 A/4\right)^{-1} \right\|_{\mathcal{L}(H)} \left\| A^{1/2} u^k \right\|_H \left\| A^{1/2} \gamma\kappa \right\|_H \\ &\leq \left\| u^k \right\|_V \left\| \gamma\kappa \right\|_V. \end{aligned} \tag{5.12}$$

Consider the third term of the right-hand side of (5.10). Since

$$\left\|(I + h_t^2 A/4)^{-1} h_t A^{1/2}/2\right\|_{\mathcal{L}(H)} = \sup_{\lambda_j \in \sigma(A)} (h_t \lambda_j^{1/2}/2)(1 + h_t^2 \lambda_j/4)^{-1} \leq 1 \text{ for any } h_t > 0,$$

we obtain

$$\begin{aligned} & h_t \left| \left\langle \left(I + h_t^2 A/4\right)^{-1} A v^k,\, \gamma\kappa \right\rangle_{V' \times V} \right| \\ & = h_t \left| \left(\left(I + h_t^2 A/4\right)^{-1} A^{1/2} v^k,\, A^{1/2} \gamma\kappa \right)_H \right| \\ & \leq \left\| \left(I + h_t^2 A/4\right)^{-1} h_t A^{1/2} v^k \right\|_H \left\| A^{1/2} \gamma\kappa \right\|_H \\ & \leq 2 \left\| \left(I + h_t^2 A/4\right)^{-1} h_t A^{1/2}/2 \right\|_{\mathcal{L}(H)} \left\| v^k \right\|_H \left\| A^{1/2} \gamma\kappa \right\|_H \\ & \leq 2 \left\| v^k \right\|_H \left\| \gamma\kappa \right\|_V. \end{aligned} \tag{5.13}$$

Consider the fourth term of the right-hand side of (5.10). Then we have

$$\begin{aligned} h_t \left| \left\langle \left(I + h_t^2 A/4\right)^{-1} f,\, \gamma\kappa \right\rangle_{V' \times V} \right| &= h_t \left| \left(\left(I + h_t^2 A/4\right)^{-1} A^{-1/2} f,\, A^{1/2} \gamma\kappa \right)_H \right| \\ &\leq h_t \left\| \left(I + h_t^2 A/4\right)^{-1} \right\|_{\mathcal{L}(H)} \left\| A^{-1/2} f \right\|_H \left\| A^{1/2} \gamma\kappa \right\|_H \\ &\leq h_t \left\| f \right\|_{V'} \left\| \gamma\kappa \right\|_V. \end{aligned}$$

By Lemma 5.4, $\beta(I+h_t^2A/4)^{-1}\gamma\,\kappa \in \operatorname{int} K$ for sufficiently small $h_t>0$. Thus if $\|\kappa-\beta(I+h_t^2A/4)^{-1}\gamma\kappa\|_W \le \eta/2$ for sufficiently small $h_t>0$, then for all $\zeta\in K^*$,

$$\begin{aligned}&\langle\zeta,\,\kappa\rangle_{W'\times W}-\left\langle\zeta,\,\beta(I+h_t^2A/4)^{-1}\gamma\kappa\right\rangle_{W'\times W}\\&=\left\langle\zeta,\,\kappa-\beta(I+h_t^2A/4)^{-1}\gamma\kappa\right\rangle_{W'\times W}\\&\le\|\zeta\|_{W'}\left\|\kappa-\beta(I+h_t^2A/4)^{-1}\gamma\kappa\right\|_W\le\frac{\eta}{2}\|\zeta\|_{W'}.\end{aligned}$$

Since $\langle\zeta,\kappa\rangle\ge\eta\|\zeta\|_{W_{-\theta}}$, we have

$$\begin{aligned}&\eta\|\zeta\|_{W_{-\theta}}-\left\langle\zeta,\,\beta(I+h_t^2A/4)^{-1}\gamma\kappa\right\rangle_{W'\times W}\\&\le\langle\zeta,\,\kappa\rangle_{W'\times W}-\left\langle\zeta,\,\beta(I+h_t^2A/4)^{-1}\gamma\kappa\right\rangle_{W'\times W}\le\frac{\eta}{2}\|\zeta\|_{W_{-\theta}}.\end{aligned}$$

Therefore we obtain

$$\left\langle\zeta,\,\beta(I+h_t^2A/4)^{-1}\gamma\kappa\right\rangle_{W'\times W}\ge\frac{\eta}{2}\|\zeta\|_{W_{-\theta}}\quad\text{for all }\zeta\in K^*.\tag{5.14}$$

Now by (5.10) and (5.11), taking ζ as N^k, we obtain

$$\begin{aligned}&h_t\sum_{k=0}^{T/h_t-1}\left\langle N^k,\,\beta\left(I+h_t^2A/4\right)^{-1}\gamma\kappa\right\rangle_{W'\times W}\\&=\sum_{k=0}^{T/h_t-1}\left(v^{k+1}-v^k,\,\gamma\kappa\right)_{H\times V}+h_t\sum_{k=0}^{T/h_t-1}\left\langle\left(I+h_t^2A/4\right)^{-1}Au^k,\,\gamma\kappa\right\rangle_{V'\times V}\\&\quad+\frac{h_t^2}{2}\sum_{k=0}^{T/h_t-1}\left\langle\left(I+h_t^2A/4\right)^{-1}Av^k,\,\gamma\kappa\right\rangle_{V'\times V}+h_t\left\langle\left(I+h_t^2A/4\right)^{-1}f,\,\gamma\kappa\right\rangle_{V'\times V}.\end{aligned}$$

Thus using Lemma 5.3, by (5.12), (5.13), and (5.14)

$$\begin{aligned}&\frac{\eta h_t}{2}\sum_{k=0}^{T/h_t-1}\left\|N^k\right\|_{W_{-\theta}}\\&\le h_t\sum_{k=0}^{T/h_t-1}\left\langle N^k,\,\beta\left(I+h_t^2A/4\right)^{-1}\gamma\kappa\right\rangle_{W'\times W}\\&\le\|\gamma\kappa\|_H\left(\left\|v^{T/h_t}\right\|_H+\left\|v^0\right\|_H\right)+h_t\|\gamma\kappa\|_V\left(\sum_{k=0}^{T/h_t-1}\left(\left\|u^k\right\|_V+\left\|v^k\right\|_H\right)\right)\\&\quad+T\,\|f\|_{V'}\|\gamma\kappa\|_V\\&\le\|\gamma\kappa\|_H\left(\left\|v^{T/h_t}\right\|_H+\left\|v^0\right\|_H\right)+T\,\|\gamma\kappa\|_V\max_{0\le k\le T/h_t-1}\left(\left\|u^k\right\|_V+\left\|v^k\right\|_H\right)\\&\quad+T\,\|f\|_{V'}\|\gamma\kappa\|_V\\&\le C\left(\|f\|_{V'}+\sqrt{E^0}+^4\sqrt{E^0\|f\|_{V'}^2}\right),\end{aligned}\tag{5.15}$$

where C does not depend on h_t. Note that E^0 is the initial energy. Since

$$\sum_{k=0}^{T/h_t-1}\left\|N^k\right\|_{W_{-\theta}}=h_t^{-1}\int_{[0,T]}\|N_{h_t}(t)\|_{W_{-\theta}}\,dt,$$

so

$$\int_{[0,T]} \|N_{h_t}(t)\|_{W_{-\theta}} dt \le \frac{2C}{\eta}\left(\|f\|_{V'} + \sqrt{E^0} + {}^4\sqrt{E^0\,\|f\|_{V'}^2}\right) \text{ for sufficiently small } h_t > 0,$$

as required. □

For any $\sigma \in \mathbb{R}$ we define the space of Hölder continuous functions with exponent $0 < p \le 1$ in $C^p(0,T;V_\sigma)$ with norm

$$\|u\|_{C^p(0,T;V_\sigma)} = \|u\|_{C(0,T;V_\sigma)} + \sup_{t\neq s} \frac{\|u(t)-u(s)\|_{V_\sigma}}{|t-s|^p}.$$

In the next Lemma, we will see that interpolant u_{h_t} is uniformly Hölder continuous from $[0,T]$ to V_σ, $0 \le \sigma < 1$. In order to do so, we need the following inequality: for $u \in V$ we have

$$\|u\|_{V_\sigma} \le C_\sigma \|u\|_H^{1-\sigma} \|u\|_V^{\sigma}, \tag{5.16}$$

where $0 < \sigma < 1$. This inequality is presented in Kuttler [15, section 22.6, equation (62)], Triebel [34, Thm. 1.3.3(g)], and Bramble & Xu [7, Appendix A, Thms. A.1 and A.2]. Also, if $\sigma = 0$ or $\sigma = 1$, clearly (5.16) is satisfied.

LEMMA 5.6. *The discrete-time solutions u_{h_t} are uniformly Hölder continuous from $[0,T] \to V_\sigma$ with exponent $p = 1-\sigma$, where $0 \le \sigma < 1$.*

Proof. We claim that for s, $t \in [0,T]$, $\|u_{h_t}(s) - u_{h_t}(t)\|_{V_\sigma} \le C\,|s-t|^p$. Using the interpolation space bounds (5.16),

$$\begin{aligned}
\|u_{h_t}(s) - u_{h_t}(t)\|_{V_\sigma} &\le \|u_{h_t}(s) - u_{h_t}(t)\|_H^{1-\sigma}\, \|u_{h_t}(s) - u_{h_t}(t)\|_V^{\sigma} \\
&\le C_\sigma \left\| \int_s^t v_{h_t}(\tau)\, d\tau \right\|_H^{1-\sigma} \|u_{h_t}(s) - u_{h_t}(t)\|_V^{\sigma} \\
&\le C_\sigma \left(\int_s^t \|v_{h_t}(\tau)\|_H \, d\tau \right)^{1-\sigma} \left(\|u_{h_t}(s)\|_V + \|u_{h_t}(t)\|_V\right)^{\sigma}.
\end{aligned}$$

Then by the energy bounds, we obtain

$$\begin{aligned}
&\|u_{h_t}(s) - u_{h_t}(t)\|_{V_\sigma} \\
&\le C_\sigma \left(\|f\|_{V'} + \sqrt{2E^0}\right)^{\sigma} \left(2\,\|f\|_{V'} + {}^4\sqrt{2E^0\,\|f\|_{V'}^2} + \sqrt{E^0}\right)^{(1-\sigma)} |s-t|^{1-\sigma} \\
&\le C_\sigma\, |s-t|^p,
\end{aligned}$$

where C_σ is a constant independent of h_t, s and t. □

Thus we can see that the interpolant u_{h_t} is uniformly bounded in $C^p(0,T;V_\sigma)$. Now we need compactness to show that u_{h_t} converges strongly in $C([0,T];V_\sigma)$, as $h_t \downarrow 0$. The continuous linear interpolant $u_{h_t} : [0,T] \to V$ is bounded in V_σ with $\sigma < 1$, so $u_{h_t}(t)$ is in a compact subset of V_σ for all $t \in [0,T]$ and $h_t > 0$, and $u_{h_t} : [0,T] \to V_\sigma$ is an equicontinuous family. By the *Arzela–Ascoli Theorem* [18, p. 57], $C^p(0,T;V_\sigma)$ is compactly embedded in $C(0,T;V_\sigma)$ for any $\sigma < 1$.

Thus u_{h_t} has a subsequence which converges strongly in $C(0,T;V_\sigma)$. Denoting this subsequence by u_{h_t}, we restrict our attention to this subsequence.

In the next Lemma, we will see that solutions obtained by our numerical trajectories satisfy differential equations in the sense of measure.

LEMMA 5.7. *The limits (u, v, N) obtained from the numerical trajectories $(u_{h_t}, v_{h_t}, N_{h_t})$ satisfy $du/dt = v$ in $L^\infty(0, T; H)$, and the equation of motion (3.1) holds in $\mathcal{M}(0, T-\epsilon; V')$, the space of V'-valued measures on $[0, T-\epsilon]$, for any $\epsilon > 0$, where $v(0) = v^0$.*

Proof. First, we claim that $du_{h_t}/dt(t) = (v_{h_t}(t-h) + v_{h_t}(t))/2$ for all $t \neq t_k$ with u_{h_t} absolutely continuous. From (5.2) and the fact that u_{h_t} is the piecewise linear interpolant of $u_h(t_k) = u^k$ for all k, we have

$$\frac{du_{h_t}}{dt}(t) = \frac{v^{k+1} + v^k}{2} \quad \text{for all } t \in (t_k, t_{k+1}).$$

Because v_{h_t} is piecewise constant, $du_{h_t}/dt(t) = (v_{h_t}(t - h_t) + v_{h_t}(t))/2$ for $t \in (t_k,\, t_{k+1})$. Thus for any $0 \le s < t \le T$,

$$u_{h_t}(t) - u_{h_t}(s) = \int_s^t \frac{1}{2}(v_{h_t}(\tau - h_t) + v_{h_t}(\tau))\, d\tau.$$

As $u_{h_t} \to u$ in $C(0, T; V_\sigma) \subset C(0, T; H)$, and $v_{h_t} \rightharpoonup^* v$ in $L^\infty(0, T; H)$, taking limits in any convergent subsequence gives

$$u(t) - u(s) = \int_s^t v(\tau)\, d\tau,$$

and so $du/dt(t) = v(t)$ for almost all t.

We claim that for all $\xi \in \mathrm{Lip}(0, T; V)$ (the space of Lipschitz functions $[0, T] \to V$) with $\xi(T) = 0$, (u, v, N) obtained from the numerical trajectory $(u_{h_t}, v_{h_t}, N_{h_t})$ satisfies

$$\int_{[0,T]} \left\langle \frac{dv}{dt}(t),\, \xi(t) \right\rangle_{V' \times V} dt = \int_{[0,T]} \langle -Au(t) + \beta^* N(t) + f,\, \xi(t) \rangle_{V' \times V}\, dt \tag{5.17}$$

interpreted in the sense of distributions. Note that a Lipschitz function $\xi\colon [0, T] \to V$ is differentiable almost everywhere with a derivative in $L^\infty(0, T; V)$ as V has the Radon–Nikodym property [11]. Note that on $[0, T]$,

$$\frac{dv_{h_t}}{dt}(t) = \sum_{k=0}^{T/h_t - 1} \left(v^{k+1} - v^k\right) \delta\left(t - t_k\right)$$

in the sense of distributions. Then using (5.1), we have, noting that T is a multiple of h_t:

$$\begin{aligned}
&\int_{[0,T]} \left\langle \frac{dv_{h_t}}{dt}(t),\, \xi(t) \right\rangle_{V' \times V} dt \\
&= \sum_{k=0}^{T/h_t - 1} \left\langle v^{k+1} - v^k,\, \xi(t_k) \right\rangle \\
&= h_t \sum_{k=0}^{T/h_t - 1} \left\langle -A\left(\frac{u^{k+1} + u^k}{2}\right) + \beta^* N^k + f,\, \xi(t_k) \right\rangle \\
&= h_t \sum_{k=0}^{T/h_t - 1} \left\langle -A\left(\frac{u^{k+1} + u^k}{2}\right),\, \xi(t_k) \right\rangle + h_t \sum_{k=0}^{T/h_t} \left\langle \beta^* N^k + f,\, \xi(t_k) \right\rangle. \qquad (5.18)
\end{aligned}$$

Consider the first sum in (5.18). Since u_{h_t} is a piecewise linear interpolant, we obtain

$$\begin{aligned}
&h_t \sum_{k=0}^{T/h_t-1} \left\langle -A\left(\frac{u^{k+1}+u^k}{2}\right), \xi(t_k)\right\rangle \\
&= \sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle -Au_{h_t}, \xi(t_k)\rangle \, dt \\
&= \sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle -Au_{h_t}(t), \xi(t)-\xi(t)+\xi(t_k)\rangle \, dt \\
&= -\sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle Au_{h_t}(t), \xi(t)\rangle \, dt + \sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle Au_{h_t}(t), \xi(t)-\xi(t_k)\rangle \, dt \\
&= -\int_{[0,T]} \langle Au_{h_t}(t), \xi(t)\rangle \, dt + \int_{[0,T]} \langle Au_{h_t}(t), \xi(t)-\xi(h_t\lfloor t/h_t\rfloor)\rangle \, dt.
\end{aligned} \tag{5.19}$$

Consider the second sum in (5.18). Then by the similar argument to above,

$$\begin{aligned}
&h_t \sum_{k=0}^{T/h_t-1} \left\langle \beta^* N^k + f, \xi(t_k)\right\rangle \\
&= \sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle \beta^* N_{h_t}(t) + f, \xi(t_k)\rangle \, dt \\
&= \sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle \beta^* N_{h_t}(t) + f, \xi(t)\rangle \, dt + \sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle \beta^* N_h(t) + f, \xi(t_k)-\xi(t)\rangle \, dt \\
&= \int_{[0,T]} \langle \beta^* N_{h_t}(t) + f, \xi(t)\rangle \, dt + \int_{[0,T]} \langle \beta^* N_{h_t}(t) + f, \xi(h_t\lfloor t/h_t\rfloor)-\xi(t)\rangle \, dt.
\end{aligned} \tag{5.20}$$

Let C_ξ be the Lipschitz constant of ξ. Then

$$\begin{aligned}
&\left| \int_{[0,T]} \langle \beta^* N_{h_t}(t) + f, \xi(h_t\lfloor t/h_t\rfloor)-\xi(t)\rangle \, dt \right| \\
&\le \int_{[0,T]} \|\beta^* N_{h_t}(t) + f\|_{V'} \, \|\xi(h_t\lfloor t/h_t\rfloor)-\xi(t)\|_V \, dt \\
&\le C_\xi \, h_t \int_{[0,T]} \left(\|\beta^* N_{h_t}(t)\|_{V'} + \|f\|_{V'}\right) \, dt \le C \, h_t
\end{aligned}$$

for some C not depending on h_t. It follows that

$$\int_{[0,T]} \langle \beta^* N_{h_t}(t) + f, \xi(h_t\lfloor t/h_t\rfloor)-\xi(t)\rangle \, dt \to 0 \quad \text{as } h_t \downarrow 0.$$

Similarly, using the fact that Au_{h_t} is uniformly bounded in $L^\infty(0,T;V')$,

$$\int_{[0,T]} \langle Au_{h_t}(t), \xi(h_t\lfloor t/h_t\rfloor - \xi(t)\rangle \, dt \to 0 \quad \text{as } h_t \downarrow 0.$$

Extend ξ to $\xi\colon \mathbb{R} \to V$ by setting $\xi(t) = \xi(0)$ for $t < 0$ and $\xi(t) = \xi(T) = 0$ for $t > T$. Now, using integration by parts,

$$\int_{[0,T]} \left\langle \frac{dv_{h_t}}{dt}(t),\, \xi(t) \right\rangle dt = \left\langle v_{h_t}(T^+),\, \xi(T^+) \right\rangle - \left\langle v_{h_t}(0^-),\, \xi(0^-) \right\rangle - \int_0^T \left\langle v_{h_t}(t),\, \frac{d\xi}{dt}(t) \right\rangle dt.$$

But $v_{h_t}(0) = v^0$, the initial velocity, and $\xi(T^+) = \xi(T) = 0$, so

$$\int_{[0,T]} \left\langle \frac{dv_{h_t}}{dt}(t),\, \xi(t) \right\rangle dt = -\left\langle v^0,\, \xi(0) \right\rangle - \int_{[0,T]} \left\langle v_{h_t}(t),\, \frac{d\xi}{dt}(t) \right\rangle dt. \tag{5.21}$$

Since $v_{h_t} \rightharpoonup^* v$ in $L^\infty(0,T;H)$, the right-hand side of (5.21) converges to

$$-\left\langle v^0,\, \xi(0) \right\rangle - \int_{[0,T]} \left\langle v(t),\, \frac{d\xi}{dt}(t) \right\rangle dt.$$

Using integration by parts again, we see that this limit is just

$$\int_{[0,T]} \left\langle \frac{dv}{dt}(t),\, \xi(t) \right\rangle dt.$$

Now taking the weak* limits $\beta^* N_{h_t} \rightharpoonup^* \beta^* N$ in $\mathcal{M}(0,T;V')$ and $Au_{h_t} \rightharpoonup^* Au$ in $L^\infty(0,T;V')$, we get

$$\int_{[0,T]} \left\langle \frac{dv}{dt}(t),\, \xi(t) \right\rangle dt = \int_{[0,T]} \langle -Au(t) + \beta^* N(t) + f,\, \xi(t) \rangle \, dt,$$

for all $\xi \in \mathrm{Lip}(0,T;V)$, $\xi(T) = 0$. Since any function $\widetilde{\xi} \in \mathrm{Lip}(0, T-\epsilon; V)$ can be extended to a function $\xi \in \mathrm{Lip}(0,T;V)$ with $\xi(T) = 0$,

$$\begin{aligned} \frac{dv}{dt} &= -Au + \beta^* N + f \quad \text{on } [0, T-\epsilon], \\ v(0) &= v^0, \end{aligned}$$

in the sense of distributions for any $\epsilon > 0$, as desired. ∎

In the next Lemma, we will see that the limit of the numerical trajectories, as $h_t \downarrow 0$, satisfies the complementarity condition in the weak sense.

LEMMA 5.8. *Setting $\sigma = (1-\theta)\alpha + \theta$ with $\theta < 1$ and $\alpha < 1$ satisfying* (A6) *and* (A7)*, we have*

$$0 \le \int_{[0,T]} \langle N_{h_t}(t), \beta u_{h_t}(t) - \varphi \rangle_{W_{-\theta} \times W_\theta} \, dt \le C_{\theta,\sigma} h_t^{1-\sigma} \int_{[0,T]} \|N_{h_t}(t)\|_{W_{-\theta}} \, dt.$$

Proof. Let $q_{h_t}(t) = u_{h_t}(t) - u^{k+1}$ for $t_k \le t \le t_{k+1}$. Then by complementarity condition, for $t_k \le t \le t_{k+1}$ we have

$$\begin{aligned} \langle N_{h_t}(t), \beta u_{h_t}(t) - \varphi \rangle_{W_{-\theta} \times W_\theta} &= \left\langle N_{h_t}(t),\, \beta q_{h_t}(t) + \beta u^{k+1} - \varphi \right\rangle_{W_{-\theta} \times W_\theta} \\ &= \langle N_{h_t}(t),\, \beta q_{h_t}(t) \rangle_{W_{-\theta} \times W_\theta}. \end{aligned}$$

Since $\left\|N^k\right\|_{W_{-\theta}} = \|N_{h_t}(t)\|_{W_{-\theta}}$ for $t_k \le t < t_{k+1}$, we obtain

$$\begin{aligned}\int_{t_k}^{t_{k+1}} \langle N_{h_t}(t),\, \beta q_{h_t}(t)\rangle_{W_{-\theta}\times W_\theta}\, dt &= \int_{t_k}^{t_{k+1}} \left\langle N^k, \beta q_{h_t}(t)\right\rangle_{W_{-\theta}\times W_\theta} dt \\ &\le C_\theta \int_{t_k}^{t_{k+1}} \|q_{h_t}(t)\|_{V_\theta} \left\|N^k\right\|_{W_{-\theta}} dt \\ &= C_\theta \left\|N^k\right\|_{W_{-\theta}} \int_{t_k}^{t_{k+1}} \|q_{h_t}(t)\|_{V_\theta}\, dt.\end{aligned}$$

Using trapezoidal rule for the integral of linear interpolant u_{h_t} and Lemma 5.6, we have

$$\begin{aligned}\int_{t_k}^{t_{k+1}} \langle N_{h_t}(t),\, \beta q_{h_t}(t)\rangle_{W_{-\theta}\times W_\theta}\, dt &\le \frac{C_\theta}{2} \left\|N^k\right\|_{W_{-\theta}} \left\|u^{k+1} - u^k\right\|_{V_\sigma} \cdot h_t \\ &\le C_{\theta,\sigma} \left\|N^k\right\|_{W_{-\theta}} h_t^{1-\sigma} h_t.\end{aligned}$$

Thus

$$\begin{aligned}\int_{[0,T]} \langle N_{h_t}(t), \beta u_{h_t}(t) - \varphi\rangle_{W_{-\theta}\times W_\theta}\, dt &= \sum_{k=0}^{T/h_t-1} \int_{t_k}^{t_{k+1}} \langle N_{h_t}(t),\, \beta q_{h_t}(t)\rangle_{W_{-\theta}\times W_\theta}\, dt \\ &\le C_{\theta,\sigma} \sum_{k=0}^{T/h_t-1} \left\|N^k\right\|_{W_{-\theta}} h_t^{1-\sigma} h_t \\ &= C_{\theta,\sigma} h_t^{1-\sigma} \int_{[0,T]} \|N_{h_t}(t)\|_{W_{-\theta}}\, dt,\end{aligned}$$

as required. ∎

LEMMA 5.9. *Suppose that dual cone K^* is strongly pointed and $\theta < 1$ and $\sigma < 1$. Then u_{h_t} strongly converges to solution u in $C(0,T;V_\sigma)$ and N_{h_t} weakly* converges to solution N in the measure space $\mathcal{M}(0,T;W_{-\theta})$, as $h_t \downarrow 0$, and the limits satisfy complementarity condition in the weak sense.*

Proof. Let $\sigma = (1-\theta)\alpha + \theta < 1$ where $\alpha < 1$ and $\theta < 1$ from (A6) and (A7). By the Arzela–Ascoli Theorem there is a strongly convergent subsequence u_{h_t} to u in $C(0,T;V_\sigma)$. An elementary calculation and the interpolation theory of operators show that $\beta\colon V_\sigma \to W_\theta$ is a bounded linear operator. On the other hand, there is a subsequence in which N_{h_t} converges weakly* to N in $\mathcal{M}(0,T;W_{-\theta})$ by the *Alouglu's Theorem* [24, Thm. 6.62, p. 203]. Then by Lemma 5.8, as $h_t \downarrow 0$, we have

$$0 \le \int_{[0,T]} \langle N_{h_t}(t), \beta u_{h_t}(t) - \varphi\rangle_{W_{-\theta}\times W_\theta}\, dt \le C_{\theta,\sigma} h_t^{1-\sigma} \int_{[0,T]} \|N_{h_t}(t)\|_{W_{-\theta}}\, dt \to 0.$$

Since K is a closed convex cone in W, for any $\zeta \in K^*$ and $0 \le s < t \le T$, $0 \le \int_s^t \langle \zeta,\, \beta u_{h_t} - \varphi\rangle\, d\tau \to \int_s^t \langle \zeta,\, \beta u - \varphi\rangle\, d\tau$, and so $\beta u(t) - \varphi \in K$ for almost all t. Similarly, $N_{h_t} \rightharpoonup^* N \in K^*$. Thus,

$$\int_{[0,T]} \langle N_{h_t}(t), \beta u_{h_t}(t) - \varphi\rangle_{W_{-\theta}\times W_\theta}\, dt \to \int_{[0,T]} \langle N(t), \beta u(t) - \varphi\rangle_{W_{-\theta}\times W_\theta}\, dt.$$

As the integral on the left is $O(h_t^{1-\sigma})$ as $h_t \downarrow 0$, $\int_{[0,T]} \langle N(t), \beta u(t) - \varphi\rangle\, dt = 0$. This implies that the limits of the discrete-time trajectories satisfy the complementarity condition in the weak sense. The proof is complete. ∎

5.3. Convergence theory in the fully discrete case. Let $\{\psi_i \mid i = 1, \ldots, \mathcal{N}_{h_x}\}$ be a basis for a subspace $V^{h_x} \subset V$; this is the basis that is used for space discretization: $u(t_k, x) \approx \sum_{j=1}^{\mathcal{N}_{h_x}} u_j^k \psi_j(x)$ on each time t_k.

Then let $\{\Phi_i \mid \Phi_i = \beta\psi_i \text{ for } i = 1, \ldots, \mathcal{N}_{h_x}\}$ be a basis for a finite dimensional space $W_\theta^{h_x} \subset W_\theta$, and $K^{h_x} \subset K$ be the cone generated by $\{\beta\psi_i \mid 1 \le i \le \mathcal{N}_{h_x}\}$ such that $K^{h_x} \subset W_\theta^{h_x}$.

In order to show the full convergence of our numerical trajectories, we need the following conditions:

(B1) The dual cone K^* is a closure of K in $W_{-\theta}$.
(B2) For every $w \in K$, $\min_{z \in K^{h_x}} \|w - z\|_W \to 0$ as $n \to 0$.
(B3) There is a constant $C > 0$, where $\|\psi_i\|_{W_{-\theta}} \le C$ for all i, independent of $h_x > 0$.
(B4) There is a constant $C_1, C_2 > 0$, where $\left\|\mathbf{M}^{-1}\right\| \le C_1$ and $\|\mathbf{M}\| \le C_2$, independent of $h_x > 0$.
(B5) There is a constant η_1, where $\min_i \kappa_i / \max_i \kappa_i \ge \eta_1 > 0$ with $\kappa_{h_x} := \sum_{j=1}^{\mathcal{N}_{h_x}} \kappa_j \times \beta\psi_j \in W_\theta \cap K^{h_x}$, independent of $h_x > 0$.
(B6) There is a constant η_2 where $\|\boldsymbol{\kappa}\|_\infty \ge \eta_2 > 0$, independent of $h_x > 0$ ($\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_{\mathcal{N}_{h_x}})$).

Note that (B1) is a self-duality condition. Also note that we can suppose (B4), by the condition number of $\mathbf{M}$, $\mathrm{cond}(\mathbf{M}) = \left\|\mathbf{M}^{-1}\right\| \|\mathbf{M}\| \le C$ and scalar-invariance. The condition (B2) implies that we can approximate w by $z \in K^{h_x}$. We write the approximate solution $N_{h_x}^k \in K^{h_x}$ each time step t_k as

$$N_{h_x}^k = \sum_{j=1}^{\mathcal{N}_{h_x}} N_j^k \beta\psi_j,$$

where $\mathbf{N}^k = (N_1^k, N_2^k, \cdots, N_{\mathcal{N}_{h_x}}^k)^T$. In the same fashion, we write approximate solution $(u_{h_x}^k, v_{h_x}^k) \in V_\sigma^{h_x} (\subset V_\sigma) \times H^{h_x} (\subset H)$ each time step t_k as

$$u_{h_x}^k = \sum_{j=1}^{\mathcal{N}_{h_x}} u_j^k \psi_j \quad \text{and} \quad v_{h_x}^k = \sum_{j=1}^{\mathcal{N}_{h_x}} v_j^k \psi_j,$$

where $\mathbf{u}^k = (u_1^k, u_2^k, \ldots, u_{\mathcal{N}_{h_x}}^k)^T$, $\mathbf{v}^k = (v_1^k, v_2^k, \ldots, v_{\mathcal{N}_{h_x}}^k)^T$. Then we notice that approximate solutions $u_{h_t,h_x} \in C(0,T;V_\sigma)$ and $v_{h_t,h_x} \in L^\infty(0,T;H)$ in the fully discrete case.

From (5.3) the approximate solutions satisfy the complementarity condition:

$$\mathbf{0} \le \mathbf{g}^{k+1} \quad \perp \quad \mathbf{N}^k \ge \mathbf{0}, \tag{5.22}$$

where $\beta u^{k+1} - \varphi = \sum_{j=1}^{\mathcal{N}_{h_x}} \left(u_j^{k+1} - \varphi_j\right) \beta\psi_j$ and $\mathbf{g}^k = \left(g_1^k, g_2^k, \cdots, g_n^k\right)^T$ with $g_j^k = u_j^k - \varphi_j$. Then employing the fully discrete version, we have the norm of N_{h_t,h_x} as a $W_{-\theta}$-valued measure given by

$$\begin{aligned} \int_{[0,T]} \|N_{h_t,h_x}\|_{W_{-\theta}} \, dt &= h_t \sum_{k=0}^{T/h_t - 1} \left\| \sum_{j=1}^{\mathcal{N}_{h_x}} N_j^k \beta\psi_j \right\|_{W_{-\theta}} \\ &\le h_t \sum_{k=0}^{T/h_t - 1} \sum_{j=1}^{\mathcal{N}_{h_x}} \left|N_j^k\right| \|\beta\psi_j\|_{W_{-\theta}} . \end{aligned} \tag{5.23}$$

The mass matrix $\mathbf{M}$ is given by $m_{ji} = (\beta\psi_j, \beta\psi_i)_Z$. Also we assume that $\mathbf{M}$ is a positive definite. Attributed to the basis functions and mass matrix, now we modify the complementarity condition (5.22) into the alternative complementarity condition that is necessary to consider full convergence:

$$\mathbf{0} \le \mathbf{g}^{k+1} \quad \perp \quad \mathbf{M}\mathbf{N}^k \ge \mathbf{0}. \tag{5.24}$$

LEMMA 5.10. *Suppose that K^* is strongly pointed;* (B1–B6) *hold. Then we have*

$$\int_{[0,T]} \|N_{h_t,h_x}\|_{W_{-\theta}}\, dt \le C \quad \textit{for sufficiently small } h_t,\, h_x > 0,$$

where $\theta \le 1$ and C does not depend on h_t and h_x.

Proof. By Lemma 5.4, $\beta(I + h_t^2 A/4)^{-1}\gamma\omega \to \omega$ on W for any $\omega \in W$. Then we approximate κ by $\kappa_{h_x} := \sum_{j=1}^{\mathcal{N}_{h_x}} \kappa_j \beta\psi_j \in W_\theta \cap K^{h_x}$, where $\kappa_j > 0$ for any $j \ge 1$, since K is a convex cone. Recalling that (A6) is equivalent to $\kappa \in \operatorname{int} K$ in W_θ, it follows that $\kappa_{h_x} \to \kappa$ as $h_x \downarrow 0$ and $\beta(I + h_t^2 A/4)^{-1}\gamma\kappa \to \kappa$ as $h_t \downarrow 0$. Thus we choose sufficiently small $h_x > 0$ and $h_t > 0$, such that $\|\kappa_{h_x} - \kappa\|_{W_\theta} \le \eta/4$ and $\|\beta(I+h_t^2A/4)^{-1}\gamma\kappa - \kappa\|_{W_\theta} \le \eta/4$. Now employing the estimate (5.15), for sufficiently small $h_t > 0$ and $h_x > 0$ we obtain

$$\begin{aligned}
C &\ge h_t \sum_{k=0}^{T/h_t-1} \left\langle N_{h_x}^k,\, \beta\left(I + h_t^2 A/4\right)^{-1}\gamma\kappa \right\rangle_{W'\times W} \\
&\ge h_t \sum_{k=0}^{T/h_t-1} \left\langle N_{h_x}^k,\, \beta\left(I + h_t^2 A/4\right)^{-1}\gamma\kappa - \kappa + \kappa - \kappa_{h_x} + \kappa_{h_x} \right\rangle_{W'\times W} \\
&= h_t \sum_{k=0}^{T/h_t-1} \left\langle N_{h_x}^k,\, \kappa_{h_x} \right\rangle_{W'\times W} - h_t \sum_{k=0}^{T/h_t-1} \left\langle N_{h_x}^k,\, \kappa_{h_x} - \kappa \right\rangle_{W'\times W} \\
&\qquad - h_t \sum_{k=0}^{T/h_t-1} \left\langle N_{h_x}^k,\, \kappa - \beta\left(I + h_t^2 A/4\right)^{-1}\gamma\kappa \right\rangle_{W'\times W} \\
&\ge h_t \sum_{k=0}^{T/h_t-1} \left(\left\langle N_{h_x}^k,\, \kappa_{h_x} \right\rangle_{W'\times W} - \frac{\eta}{2} \left\| N_{h_x}^k \right\|_{W_{-\theta}} \right).
\end{aligned} \tag{5.25}$$

Then using (5.15) again, from (5.25)

$$h_t \sum_{k=0}^{T/h_t-1} \left\langle N_{h_x}^k,\, \kappa_{h_x} \right\rangle_{W'\times W} \le \frac{\eta h_t}{2} \sum_{k=0}^{T/h_t-1} \left\| N_{h_x}^k \right\|_{W_{-\theta}} + C \le C.$$

Thus we have

$$\begin{aligned} C &\geq h_t \sum_{k=0}^{T/h_t-1} \left\langle N_{h_x}^k,\, \kappa_{h_x} \right\rangle_{W'\times W} \\ &= h_t \sum_{k=0}^{T/h_t-1} \left(N_{h_x}^k,\, \kappa_{h_x} \right)_Z \\ &= h_t \sum_{k=0}^{T/h_t-1} \left(\sum_{j=1}^{\mathcal{N}_{h_x}} N_j^k \beta\psi_j, \sum_{i=1}^{\mathcal{N}_{h_x}} \kappa_i \beta\psi_i \right)_Z \\ &= h_t \sum_{k=0}^{T/h_t-1} \sum_{i=0}^{\mathcal{N}_{h_x}} \kappa_i \cdot (\mathbf{M}\mathbf{N}^k)_i \\ &\geq h_t \sum_{k=0}^{T/h_t-1} \sum_{i=0}^{\mathcal{N}_{h_x}} \min_{1\leq i\leq \mathcal{N}_{h_x}} \kappa_i \cdot (\mathbf{M}\mathbf{N}^k)_i. \end{aligned}$$

So it follows from condition (B5) that

$$h_t \sum_{k=0}^{T/h_t-1} \left\|\mathbf{M}\mathbf{N}^k\right\|_1 \leq \frac{C}{\eta_1 \left\|\boldsymbol{\kappa}\right\|_\infty}. \tag{5.26}$$

Therefore using (5.23) and (5.26),

$$\begin{aligned} \int_{[0,T]} \|N_{h_t,h_x}\|_{W_{-\theta}}\, dt &= h_t \sum_{k=0}^{T/h_t-1} \left\| \sum_{j=1}^{\mathcal{N}_{h_x}} N_j^k \beta\psi_j \right\|_{W_{-\theta}} \\ &\leq h_t \sum_{k=0}^{T/h_t-1} \sum_{j=1}^{\mathcal{N}_{h_x}} \left|N_j^k\right| \|\psi_j\|_{W_{-\theta}} \\ &= Ch_t \sum_{k=0}^{T/h_t-1} \left\|\mathbf{N}^k\right\|_1 \\ &= Ch_t \sum_{k=0}^{T/h_t-1} \left\|\mathbf{M}^{-1}\mathbf{M}\mathbf{N}^k\right\|_1 \\ &\leq Ch_t \sum_{k=0}^{T/h_t-1} \left\|\mathbf{M}^{-1}\right\|_1 \left\|\mathbf{M}\mathbf{N}^k\right\|_1 \\ &\leq Ch_t \sum_{k=0}^{T/h_t-1} \left\|\mathbf{M}\mathbf{N}^k\right\|_1 \\ &\leq \frac{C}{\eta_1 \cdot \eta_2} = C, \end{aligned}$$

where C does not depend on h_t and h_x. The proof is complete. ∎

As we mentioned in the semidiscrete case, in the next Lemmas we will restrict our attention to subsequences which are convergent to a solution. We note that we

require contact condition $\mathbf{MN}^k \geq \mathbf{0}$ in (5.24) rather than $\mathbf{N}^k \geq \mathbf{0}$ in (5.22), in order to show full convergence.

LEMMA 5.11. *If $\mathbf{MN}^k \geq \mathbf{0}$ and subsequence N_{h_t,h_x} weakly* converges N in the measure space $\mathcal{M}(0,T;W_{-\theta})$, then $N \in K^*$.*

Proof. Let any ξ be in $C^1(0,T;W_\theta)$. Since $N_{h_t,h_x} \rightharpoonup^* N$ as $W_{-\theta}$-valued measure, we have

$$\int_{[0,T]} \langle N_{h_t,h_x},\, \xi\rangle_{W_{-\theta}\times W_\theta}\, dt \to \int_{[0,T]} \langle N,\, \xi\rangle_{W_{-\theta}\times W_\theta}\, dt.$$

Let $\xi_{h_x}(t) = \sum_{i=1}^{\mathcal{N}_{h_x}} \xi_i(t)\psi_i \in W_\theta \cap K^{h_x}$ be the projection of ξ onto K^{h_x} with $\xi_i(t) > 0$ for all t and i. Since the convex projection is a nonexpansive mapping, ξ_{h_x} is uniformly bounded; ξ_{h_x} is also Lipschitz with the same Lipschitz constant as ξ. Then from (B2) $\xi_{h_x} \to \xi \in K$ with all t. By the standard arguments, $\|\xi_{h_x} - \xi\|_{L^\infty(0,T;W_\theta)} \to 0$ as $h_x \downarrow 0$. Therefore as $h_t \downarrow 0$ and $h_x \downarrow 0$ in suitable subsequence,

$$\int_{[0,T]} \langle N_{h_t,h_x},\, \xi_{h_x}\rangle_{W_{-\theta}\times W_\theta}\, dt \to \int_{[0,T]} \langle N,\, \xi\rangle_{W_{-\theta}\times W_\theta}\, dt. \tag{5.27}$$

Since $\mathbf{MN}^k \geq \mathbf{0}$ and $\xi_i(t) > 0$, it follows that

$$\begin{aligned}
&\int_{[0,T]} \langle N_{h_t,h_x},\, \xi_{h_x}\rangle_{W_{-\theta}\times W_\theta}\, dt \\
&= \int_{[0,T]} \left(h_t \sum_{k=0}^{T/h_t-1} \delta(t-(k+1)h_t) \sum_{j=1}^{\mathcal{N}_{h_x}} N_j^k \beta\psi_j,\, \sum_{i=1}^{\mathcal{N}_{h_x}} \xi_i(t)\beta\psi_i \right)_Z dt \\
&= h_t \sum_{k=0}^{T/h_t-1} \left(\sum_{j=1}^{\mathcal{N}_{h_x}} N_j^k \beta\psi_j,\, \sum_{i=1}^{\mathcal{N}_{h_x}} \xi_i(t_k)\beta\psi_i \right)_Z \\
&= h \sum_{k=0}^{T/h_t-1} \sum_{i=1}^{\mathcal{N}_{h_x}} \xi_i(t_k)(\mathbf{MN}^k)_i \geq 0.
\end{aligned}$$

Thus by (5.27), we have

$$0 \leq \int_{[0,T]} \langle N_{h_t,h_x},\, \xi_{h_x}\rangle_{W_{-\theta}\times W_\theta}\, dt \to \int_{[0,T]} \langle N,\, \xi\rangle_{W_{-\theta}\times W_\theta}\, dt.$$

Thus $\int_{[0,T]} \langle N,\, \xi\rangle\, dt \geq 0$. This implies that $N \in K^*$ in the W'-valued measure. ∎

We note that $du_{h_t,h_x}/dt(t) = \left(v_{h_t,h_x}(t) - v_{h_t,h_x}(t-h_t)\right)/2$ with $v_{h_t,h_x}(t) = v^0$ for $-h_t \leq t < 0$ from numerical formulation (5.2). Before we see the next Lemma, we notice that $u_{h_t,h_x} \rightharpoonup^* u$ in $L^\infty(0,T;V_\sigma)$ with $\sigma \leq 1$ and $v_{h_t,h_x} \rightharpoonup^* v$ in $L^\infty(0,T;H)$, since $L^\infty(0,T;V_\sigma) \simeq \left(L^1(0,T;V_\sigma)\right)^*$ and $L^\infty(0,T;H) \simeq \left(L^1(0,T;H)\right)^*$.

LEMMA 5.12. *Suppose that dual cone K^* is strongly pointed and $\sigma < 1$ and $\theta < 1$. Then the solution $(u,\, N)$ obtained from numerical trajectories $(u_{h_t,h_x},\, N_{h_t,h_x})$ satisfies the complementarity condition* (5.24) *in the weak sense.*

Proof. By the energy bounds in Lemma 5.3 and the similar argument to Lemma 5.6, we have

$$\begin{aligned}\|u_{h_t,h_x}(s)-u_{h_t,h_x}(t)\|_{V_\sigma} &\le \|u_{h_t,h_x}(s)-u_{h_t,h_x}(t)\|_H^{1-\sigma}\,\|u_{h_t,h_x}(s)-u_{h_t,h_x}(t)\|_V^{\sigma}\\ &\le C_\sigma\,\|(v_{h_t,h_x}(t)-v_{h_t,h_x}(t-h_t))\|_H^{1-\sigma}\\ &\quad\times\left(\|u_{h_t,h_x}(s)\|_V+\|u_{h_t,h_x}(t)\|_V\right)^{\sigma}\\ &\le C_\sigma\,|s-t|^p\,,\end{aligned}$$

where $p=1-\sigma$ and $\sigma<1$. By the Arzela–Ascoli Theorem, there is a subsequence $u_{h_t,h_x}\to u$ in $C(0,T;V_\sigma)$, as $h_t,\,h_x\downarrow 0$. Since $\beta: V_\sigma\to W_\theta$ (see Lemma 5.9) is a bounded linear operator, by (5.24) we obtain

$$\begin{aligned}&\int_{[0,T]}\langle N_{h_t,h_x},\beta u_{h_t,h_x}-\varphi_{h_x}\rangle_{W_{-\theta}\times W_\theta}\,dt\\ &=\int_{[0,T]}\left(h_t\sum_{k=0}^{T/h_t-1}\delta(t-(k+1)h_t)\sum_{j=1}^{\mathcal{N}_{h_x}}N_j^k\beta\psi_j,\ \sum_{i=1}^{\mathcal{N}_{h_x}}g_i^{k+1}\beta\psi_i\right)_Z dt\\ &=h_t\sum_{k=0}^{T/h_t-1}\left(\sum_{j=1}^{\mathcal{N}_{h_x}}N_j^k\beta\psi_j,\ \sum_{i=1}^{\mathcal{N}_{h_x}}g_i^{k+1}\beta\psi_i\right)_Z\\ &=h_t\sum_{k=0}^{T/h_t-1}(\mathbf{g}^{k+1})^T\mathbf{M}\mathbf{N}^k=0.\end{aligned}$$

Since $N_{h_t,h_x}\rightharpoonup^* N\in K^*$ and $\beta u_{h_t,h_x}-\varphi_{h_x}\to\beta u-\varphi\in K$ as measure with $h_t,\,h_x\downarrow 0$ from Lemma 5.11, taking the limit in the subsequence, it follows that

$$0=\int_{[0,T]}\langle N_{h_t,h_x},\beta u_{h_t,h_x}-\varphi_{h_x}\rangle_{W_{-\theta}\times W_\theta}\,dt\to\int_{[0,T]}\langle N,\beta u-\varphi\rangle_{W_{-\theta}\times W_\theta}\,dt.$$

The proof is complete. ∎

6. Conclusions. We have been able to show the existence of solutions for a general class of impact problems of the form

$$\begin{aligned}u_{tt}&=-Au+\beta^*N(t)+f,\\ u(0)&=u^0,\\ u_t(0)&=v^0,\\ K\ni\beta u(t)-\varphi\perp N(t)\in K^*&\quad\text{for all } t\in[0,T],\end{aligned}$$

with the last condition interpreted in a weak sense: N is a measure on $[0,T]$ with values in K^* ($\int_B N\in K^*$ for any Borel set B), $\beta u(t)-\varphi\in K$ for all t, and

$$\int_{[0,T]}\langle\beta u(t)-\varphi,\,N(t)\rangle\,dt=0,$$

with the integrand identified as a measure.

The most important assumption that was made was (A6) which is a strong pointedness assumption. The usual way in which we guarantee this is by requiring that W is a Sobolev space $H^m(\Omega_1)$ with Ω_1 a d-dimensional manifold or open set in $\mathbb{R}^d$,

where $m > d/2$ in order to get an embedding $H^{m-\epsilon}(\Omega_1) \subset C(\overline{\Omega_1})$ for some $\epsilon > 0$. This means that either m is large or d is small. Thus dynamic obstacle problems for the wave equation or the equations of elasticity in two or three dimensions (or more) cannot be treated by these methods. The incorporation of viscosity to create a visco-elastic system of equations can be used to get greater regularity in the solution u and the velocity $u_t = v$ [16]. However, it is not clear what happens to the normal contact force in such cases.

The solutions whose existence is shown are quite weak, and we cannot in general expect to prove either uniqueness or conservation of energy from such a weak solution. However, further work should improve on the regularity of the solutions (perhaps with, perhaps without, visco-elasticity) and we hope to apply the results in [30] to get conservation of energy, or at least an energy balance in which no energy losses or gains are attributed to contact forces. This is a matter for future research.

REFERENCES

[1] J. Ahn and D. E. Stewart, *Euler–Bernoulli beam in impact: Existence*, Technical report, University of Iowa, Iowa City, Iowa, 2005.

[2] J. Ahn and D. E. Stewart, *Euler–Bernoulli beam with dynamic contact: Discretization, convergence, and numerical results*, SIAM J. Numer. Anal., 43 (2005), pp. 1455–1480.

[3] J. Ahn and D. E. Stewart, *Frictionless dynamic contact in linear viscoelasticity*, preprint.

[4] L. Amerio and G. Prouse, *Study of the motion of a string vibrating against an obstacle*, Rend. Mat., 2 (1975), pp. 563–585.

[5] K. T. Andrews, M. Shillor, and S. Wright, *On the dynamic vibrations of an elastic beam in frictional contact with a rigid obstacle*, J. Elasticity, 42 (1996), pp. 1–30.

[6] C. Baiocchi and A. Capelo, *Variational and Quasivariational Inequalities: Applications to Free Boundary Problems*, Wiley, Chichester, New York, 1984.

[7] J. H. Bramble and X. Zhang, *The Analysis of Multigrid Methods*, Handb. Numer. Anal. VII, North-Holland, Amsterdam, 2000.

[8] V. G. Čeban and I. V. Russu, *A numerical method of solution of the dynamical problem of the elastic impact of a thin rectangular plate on a rigid obstacle*, Prikl. Mat. i Programmirovanie Vyp., 4 (1971), pp. 94–104.

[9] M. Cocu, E. Pratt, and M. Raous, *Constructive aspects of functional analysis for the treatment of frictional contact*, Math. Comput. Modelling, 28 (1998), pp. 109–120.

[10] M. Cocu and J.-M. Ricaud, *Analysis of a class of implicit evolution inequalities associated to viscoelastic dynamic contact problems with friction*, Internat. J. Engrg. Sci., 38 (2000), pp. 1535–1552.

[11] J. Diestel and J. J. Uhl, Jr., *Vector Measures*, Math. Surveys Monogr., Amer. Math. Soc., Providence, RI, 1977.

[12] N. Dunford and J. T. Schwartz, *Linear Operators. Part* I: *General Theory*, Wiley Interscience, New York, 1957.

[13] L. Hörmander, *The Analysis of Linear Partial Differential Operators I: Distribution theory and Fourier Analysis*, in Grundlehren Math. Wiss. 256, Springer-Verlag, Berlin, Heidelberg, New York, 1983.

[14] J. U. Kim, *A boundary thin obstacle problem for a wave equation*, Commun. Partial Differential Equations, 14 (1989), pp. 1011–1026.

[15] K. Kuttler, *Modern Analysis*, CRC Press, Boca Raton, FL, 1998.

[16] K. Kuttler and M. Shillor, *Dynamic contact with Signorini's condition and slip rate dependent friction*, Electron. J. Differential Equations, (2004).

[17] K. L. Kuttler and M. Shillor, *Vibrations of a beam between two stops*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, 8 (2001), pp. 93–110.

[18] S. Lang, *Real and Functional Analysis*, 2nd ed., Graduate Texts in Math. 142, Springer, New York, 1993.

[19] G. Lebeau and M. Schatzman, *A wave problem in a half-space with a unilateral constraint at the boundary*, J. Differential Equations, 53 (1984), pp. 309–361.

[20] C. Marchionna and S. Panizzi, *On the Timoshenko beam vibrating under an obstacle condition*, Meccanica, 32 (1997), pp. 101–114.

[21] J.-J. Moreau, *Bounded variation in time*, in Topics in Nonsmooth Mechanics, J. J. Moreau, P. D. Panagiotopoulos, and G. Strang, eds., Birkhäuser, Basel, Switzerland, 1988, pp. 1–74.

[22] A. Petrov and M. Schatzman, *Viscoélastodynamique monodimensionnelle avec conditions de Signorini*, C. R. Math Acad. Sci. Paris Sér. I, 334 (2002), pp. 983–988.
[23] A. Petrov and M. Schatzman, *A pseudodifferential linear complementarity problem related to a one dimensional viscoeleastic model with Signorini conditions*, Arch. Ration. Mech. Anal., to appear.
[24] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Texts Appl. Math. 13, Springer-Verlag, New York, Berlin, Heidelberg, 1993.
[25] E. J. Routh, *A Treatise on the Dynamics of a System of Rigid Bodies*, Macmillan, London, 1860.
[26] M. Schatzman, *A hyperbolic problem of second order with unilateral constraints: The vibrating string with a concave obstacle*, J. Math. Anal. Appl., 73 (1980), pp. 138–191.
[27] M. Schatzman, *Un problème hyperbolique du 2ème ordre avec contrainte unilatérale: La corde vibrante avec obstacle ponctuel*, J. Differential Equations, 36 (1980), pp. 295–334.
[28] P. Shi, *The restitution coefficient of a linear elastic rod*, Math. Comput. Modelling, 28 (1998), pp. 427–435.
[29] D. E. Stewart, *Formulating measure differential inclusions in infinite dimensions*, Set-Valued Anal., 8 (2000), pp. 273–293.
[30] D. E. Stewart, *Differentiating complementarity problems*, Houston J. Math., to appear.
[31] M. E. Taylor, *Partial Differential Equations* 1, Appl. Math. Sci. 115, Springer, New York, 1996.
[32] S. Timoshenko, *On the correction for shear of the differential eqution for transverse vibrations of prismatic bars*, Philosophical Magazine, 41 (1921), pp. 744–746.
[33] S. Timoshenko, *On the transverse vibrations of bars of uniform cross-section*, Philosophical Magazine, 43 (1922), pp. 125–131.
[34] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North Holland, Amsterdam, 1978.
[35] J. Wloka, *Partial Differential Equations*, Cambridge University Press, London, 1987.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 64–102

BLOW-UP AND GLOBAL ASYMPTOTICS OF THE LIMIT UNSTABLE CAHN–HILLIARD EQUATION*

J. D. EVANS†, V. A. GALAKTIONOV†, AND J. F. WILLIAMS‡

Abstract. We study the asymptotic behavior of classes of global and blow-up solutions of a semilinear parabolic equation of the "limit" Cahn–Hilliard type

$$u_t = -\Delta(\Delta u + |u|^{p-1}u) \quad \text{in } \mathbf{R}^N \times \mathbf{R}_+, \quad p > 1,$$

with bounded integrable initial data. We show that in some $\{p, N\}$-parameter ranges it admits a *countable* set of blow-up similarity patterns. The most interesting set of blow-up solutions is constructed at the first critical exponent $p = p_0 = 1 + \frac{2}{N}$, where the first simplest profile is shown to be stable. Unlike the blow-up case, we show that, for $p = p_0$, the set of global decaying source-type similarity solutions is *continuous* and determine the stable mass-branch. We prove that there exists a countable spectrum of critical exponents $\{p = p_l = 1 + \frac{2}{N+l},\ l = 0, 1, 2, \dots\}$ creating bifurcation branches, which play a key role in general description of solutions globally decaying as $t \to \infty$.

Key words. semilinear parabolic equation, similarity solutions, blow-up, asymptotic behavior

AMS subject classifications. 35K55, 35K65

DOI. 10.1137/S0036141004440289

1. Introduction.

1.1. The model and discussion. In this paper we study the blow-up and the long-time asymptotic behavior of solutions of the fourth-order semilinear parabolic equation

$$u_t = -\Delta(\Delta u + u^p) \quad \text{in } \mathbf{R}^N \times \mathbf{R}_+, \quad p > 1, \quad \text{where } u^p := |u|^{p-1}u. \tag{1.1}$$

We consider the Cauchy problem for (1.1) with initial data

$$u(x, 0) = u_0(x) \quad \text{in } \mathbf{R}^N, \quad u_0 \in L^1(\mathbf{R}^N) \cap L^\infty(\mathbf{R}^N), \tag{1.2}$$

assuming in most cases that $u_0(x)$ decays exponentially as $x \to \infty$.

Equation (1.1) is a model connected with various applications. For instance, it arises as the limit case of the phenomenological, "unstable" Cahn–Hilliard equation for $N = 1$, 2 and $p = 3$, see the references in [32] and [12],

$$u_t = -(\gamma u_{xx} - u^3 + \gamma_1 u)_{xx} - \gamma_2 u.$$

It is also a reduced model from solidification theory with $N = 1$ or 2 and $p = 2$ [31, 3]. Equations of this form arise in the theory of thermo-capillary flows in thin layers of viscous fluids with free boundaries, with an anomalous dependence of the surface tension coefficient on temperature [14, 1]. Equation (1.1) also occurs as the *limit*

*Received by the editors May 6, 2004; accepted for publication (in revised form) September 28, 2005; published electronically March 31, 2006. Research supported by TMR networks ERB FMRX CT98-0201 and RTN network HPRN-CT-2002-00274.

http://www.siam.org/journals/sima/38-1/44028.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (masjde@maths.bath.ac.uk, vag@maths.bath.ac.uk).

‡Department of Mathematics, Simon Fraser University, Burnaby, BC, V6T 1Z2, Canada (jfw@math.sfu.ca).

case as $\gamma \to 0^+$ of the Cahn–Hilliard equation with a standard double-well potential function of the form

$$
\begin{aligned}
&u_t = \nabla \cdot (\nabla (F(u) - \Delta u)), \quad \text{where} \\
&F(u) = |u|^{p-1}u - \gamma |u|^p u, \quad \text{with } \gamma > 0,
\end{aligned}
\tag{1.3}
$$

so the constant stationary state satisfies $|U_0| = \frac{1}{\gamma} \to \infty$. As usual in the similarity analysis, the present blow-up results can be applied for nonlinearities in (1.3) for sufficiently small $\gamma > 0$, i.e., for $|U_0| \gg 1$, exhibiting *intermediate* asymptotics in the sense of Zel'dovich and Barenblatt [2].

Another important class of fourth-order models related to (1.1) admitting both blow-up and decaying solutions comes from the theory of thin films and general long-wave unstable equations (see [3] and also [36]), where a typical *quasi-linear* equation takes the form

$$
u_t = -(u^n u_{xxx} + u^m u_x)_x. \tag{1.4}
$$

Equations of this form are known to admit nonnegative solutions constructed by special parabolic approximations of the nonlinear coefficients; see [4] and the references therein. Then $m = n + 2$ corresponds to the critical self-similar "conservative" case to be treated here for $n = 0$. Notice that the Cauchy problem for the uniformly parabolic equation (1.1) cannot admit compactly supported solutions and this changes some essential properties of evolution. On the other hand, one can state the same free-boundary conditions for (1.1) as for (1.4) hence admitting finite interfaces.

From the mathematical point of view, in the case of the Cauchy problem studied in the present paper, writing (1.1) in the form

$$
Pu_t = \Delta u + u^p, \tag{1.5}
$$

with the positive operator $P = (-\Delta)^{-1}$ on the right-hand side, defines a pseudo-parabolic second-order equation. Many aspects of such equations are well understood with both existence and uniqueness of local and global classical solutions and the blow-up of solutions known from the 1970's; see the first results on blow-up in [28] and the references in the surveys [15, 27].

We are mainly interested in the study of blow-up behavior of the solutions to (1.1), (1.2). In this sense (1.1) is a special model, for which (1.5) clearly indicates that, at least formally, we can expect some similarities of blow-up singularity formation phenomena to the classical semilinear heat equation from combustion theory (the solid-fuel model)

$$
u_t = \Delta u + u^p \quad \text{in } \mathbf{R}^N \times \mathbf{R}_+ \quad (p > 1). \tag{1.6}
$$

There is a huge mathematical literature, developed in the last twenty years, devoted to the study of blow-up solutions of (1.6), where in some ranges of the parameters p and N a complete description of all possible patterns has been achieved; see the references in [34, Chap. 4] and survey [15]. On the other hand, the unstable nonlinear operator in (1.1) gives the classical *porous medium equation* but posed backwards in time $u_t = -\Delta u^p$. It is not well posed and leads to blow-up of u or its derivatives in arbitrarily small times; see the concavity techniques in [29].

In standard form, (1.1) (rather than (1.5)) is a fourth-order semilinear parabolic equation. It is well known that any kind of detailed asymptotic analysis for higher-order equations is much more difficult in comparison with the second-order counterparts in view of the lack of the Maximum Principle, comparison, order-preserving semigroups and potential properties of the operators involved, etc. In this sense, we are going to show that (1.1) can be treated as an "intermediate" model between second- and fourth-order parabolic equations, where some questions of blow-up and global asymptotics, at least partially, can be studied by reasonably standard mathematical methods. Note that a completely rigorous analysis of blow-up asymptotics, comparable to that for second-order problems, is not yet available for higher-order semilinear equations such as

$$u_t = -(-\Delta)^m u + u^p, \quad m \ge 2, \tag{1.7}$$

see [6], nor for the generalized Frank–Kamenetskii equation

$$u_t = -\Delta^2 u + \mathrm{e}^u$$

with similar blow-up properties, nor for the model equation from the Semenov–Rayleigh–Benard problem with the leading operator of the form

$$u_t = -u_{xxxx} + \beta[(u_x)^3]_x + \mathrm{e}^u, \quad \beta \ge 0,$$

see [16]. The mathematical difficulties in understanding the ODE and PDE patterns increase dramatically with the order of differential operators in the equations.

1.2. The main conclusions and plan of the paper. The primary goal is to present some general principles of formation of stable generic blow-up and global asymptotics for the limit Cahn–Hilliard equation (1.1). To this end, we construct various sets of self-similar solutions of (1.1) in different ranges of the parameters p and N. We establish that, in the most interesting *conservative* critical case $p = p_0 = 1 + \frac{2}{N}$,

(i) (1.1) admits a countable discrete set of blow-up similarity solutions, and

(ii) there exists an unbounded continuous family of global similarity solutions decaying as $t \to \infty$.

We also show which of the similarity solutions are stable in a proper rescaled sense and hence describe the behavior of a wide class of more general solutions. We study the stability of the main branches of similarity solutions and discuss the sets of similarity profiles for arbitrary $p > 1$.

In the remainder of this section we discuss some basic auxiliary properties of (1.1). In section 2 we introduce the blow-up and global similarity solutions and perform a local asymptotic analysis of the corresponding ODEs. The spectral properties of the key non-self-adjoint linearized operators are described in section 3. In section 4 we present an existence analysis of blow-up similarity solutions and show that the minimal profile is an attractor for a wide set of initial data. A countable set of solutions in the critical case $p = p_0 = 1 + \frac{2}{N}$ is then constructed via a singular perturbation expansion. The non-conservative case is briefly discussed in section 5, where more general p values are considered. In section 6 we study classes of global solutions of (1.1) decaying as $t \to \infty$ and prove that, unlike the blow-up case, the family of similarity solutions at $p = p_0$ is continuous and we determine the stable branch. We also detect a sequence of critical exponents

$$p_l = 1 + \frac{2}{N + l}, \quad l = 0, 1, 2, \dots, \tag{1.8}$$

corresponding to a *transition* phenomenon between the two main classes of asymptotic patterns for the PDE (1.1). Under certain assumptions, we prove that, at each $p = p_l$, a bifurcation of similarity solutions occurs, and show that, actually, for $p \neq p_0$, p_1, the set of exponentially decaying similarity profiles is expected to be countable. Concerning possible *non-self-similar* asymptotic behavior, we show that at the same critical exponents $p = p_l$ unusual center manifold patterns occur for the limit *stable* Cahn–Hilliard equation

$$u_t = -\Delta(\Delta u - u^p), \tag{1.9}$$

where the operator on the right-hand side is monotone and coercive in $H^{-1}(\mathbf{R}^N)$, so that the Cauchy problem has a unique classical solution decaying in time, at least, for $p < \frac{N+2}{N-2}$; see [17, sect. 2].

1.3. Some preliminary results. We begin with some well-known related properties.

(a) *A potential operator and gradient system.* Equation (1.1) is uniformly parabolic with all spatial differential operators appearing in divergence form. It admits a unique, classical, local in time solution and the standard parabolic theory applies, [13]. The operator on the right-hand side of (1.1), $-\Delta(\Delta u + u^p)$, is potential in $H^{-1}(\mathbf{R}^N)$ (see [25]) and the dynamical system admits the Lyapunov function

$$E[u](t) = \frac{1}{2}\|\nabla u(t)\|_2^2 - \frac{1}{p+1}\|u(t)\|_{p+1}^{p+1}, \quad u(t) \in X = H^2(\mathbf{R}^N) \cap L^{p+1}(\mathbf{R}^N), \tag{1.10}$$

which is monotone decreasing with time on uniformly bounded orbits in X, $\frac{\mathrm{d}}{\mathrm{d}t}E[u] = -\|u_t\|_{H^{-1}}^2 \le 0$. The PDE defines a smooth gradient system in $H^{-1}(\mathbf{R}^N)$. By the gradient system theory [19], the ω-limit set of any uniformly bounded orbit,

$$\omega(u_0) = \left\{ f \in C(\mathbf{R}^N) : \exists\ \{t_k\} \to \infty \text{ such that } u(\cdot, t_k) \to\ f \ \text{ uniformly} \right\},$$

is known to consist of classical stationary solutions: $-\Delta(\Delta f + f^p) = 0$ in $\mathbf{R}^N$ for any $f \in \omega(u_0)$. If the set of stationary solutions consists of isolated equilibria, the asymptotic behavior does not essentially differ from the classical second-order theory and any bounded orbit approaches a stationary profile as $t \to \infty$. However, one can see from (1.5) that, for this problem with $p < p_S = (N+2)/(N-2)_+$ (which is the critical Sobolev exponent for the elliptic operator in (1.5)), the only admissible stationary solution is trivial, $f = 0$, so that the large-time behavior cannot be stabilization to a nonzero equilibrium. Instead, there can be blow-up or convergence to $f = 0$ with a rate as yet to be determined.

(b) *The fundamental solution and local existence.* We first need the fundamental solution of the linear fourth-order parabolic equation

$$u_t = -\Delta^2 u \quad \text{in } \mathbf{R}^N \times \mathbf{R}_+ \tag{1.11}$$

having the similarity form

$$b(x,t) = t^{-N/4}F(y), \quad y = x/t^{1/4}, \tag{1.12}$$

where the rescaled kernel F is the unique radial solution of the elliptic equation

$$\mathbf{B}F \equiv -\Delta^2 F + \frac{1}{4}y \cdot \nabla F + \frac{N}{4}F = 0 \quad \text{in } \mathbf{R}^N, \quad \text{with } \int F = 1. \tag{1.13}$$

The kernel $F = F(|y|)$ has exponential decay, oscillates as $|y| \to \infty$ and satisfies the following estimate [11, p. 47]: for some positive constants D and d,

$$|F(y)| < D\,\mathrm{e}^{-d|y|^{4/3}} \quad \text{in } \mathbf{R}^N. \tag{1.14}$$

A sharp estimate $d < 3/2^{5/2}$ is obtained from the exponential asymptotics of solutions to the ODE (1.13) as $|y| \to \infty$, such as those described in section 2.

Local existence of the unique classical solution of the Cauchy problem (1.1), (1.2) follows from the equivalent integral equation

$$u(\cdot,t) = \mathbf{M}(u) \equiv b(\cdot,t) * u_0 - \int_0^t \Delta b(\cdot,t-s) * u^p(s)\mathrm{d}s, \quad t > 0, \tag{1.15}$$

where $\mathbf{M}$ is a contraction in the metric space of continuous functions in $\mathbf{R}^N \times [0,\delta]$, $\delta > 0$ small, with the sup-norm. Using (1.12), one can see from (1.15) that any uniformly bounded solution $u(x,t)$ admits a local-in-time classical extension. Therefore, as is usual in classical parabolic theory, if blow-up of any form occurs it must do so in the L^∞-norm.

Finite-time blow-up. Despite the gradient structure, from known results on blow-up for such pseudoparabolic equations (1.5) [28], it follows that classical solutions whose initial data satisfy the energy inequality

$$E(u_0) = \frac{1}{2}\,\|\nabla u_0\|_2^2 - \frac{1}{p+1}\,\|u_0\|_{p+1}^{p+1} < 0 \tag{1.16}$$

cannot be extended beyond a finite blow-up time $T < \infty$. The problem of blow-up has been extensively studied in recent years for various higher-order quasi-linear and semilinear parabolic models, where many interesting results have been obtained; see also [32, 31, 3, 4, 5, 36] and the references therein.

2. Similarity variables for global and blow-up asymptotics. Equation (1.1) is invariant under the group of scaling transformations

$$t \mapsto \lambda t, \quad x \mapsto \lambda^{1/4} x, \quad u \mapsto \lambda^{-1/2(p-1)} u, \quad \lambda > 0.$$

This symmetry suggests the introduction of the following rescaled variables:

$$u(x,t) = [\sigma(T-t)]^{-1/2(p-1)}\theta(y,\tau), \quad y = x/[\sigma(T-t)]^{1/4}, \quad \tau = -\sigma\ln[\sigma(T-t)], \tag{2.1}$$

where σ takes the two values ± 1: $\sigma = 1$ corresponds to blow-up at the unknown blow-up time $t = T$, and $\sigma = -1$ to infinite time decay as $t \to \infty$ with a reference time T (generically, we will take $T = 0$ in this case). Without loss of generality, we suppose that the solution $u(x,t)$ blowing up at the finite time $t = T$ in the L^∞-norm, i.e.,

$$\sup_x |u(x,t)| \to \infty \text{ as } t \to T^-, \tag{2.2}$$

is such that the corresponding blow-up set $B(u_0)$ contains the origin,

$$0 \in B(u_0) = \{x \in \mathbf{R}^N : \ \exists\ \{x_n\} \to x,\ \{t_n\} \to T^- \text{ such that } u(x_n,t_n) \to \infty\}.$$

The rescaled solution $\theta(y,\tau)$ defined as in (2.1) satisfies the semilinear equation

$$\theta_\tau = \mathbf{A}(\theta) \equiv -\Delta(\Delta\theta + \theta^p) - \frac{\sigma}{4} y\cdot\nabla\theta - \frac{\sigma}{2(p-1)}\theta, \tag{2.3}$$

and we are interested in the possible asymptotic dynamics of solutions for $\tau \gg 1$.

An exact similarity solution of (1.1)

$$u_S(x,t) = [\sigma(T-t)]^{-1/2(p-1)} f(y) \tag{2.4}$$

gives an independent of τ stationary solution $f(y)$ of (2.3) leading to a fourth-order elliptic equation for the similarity profile f,

$$\mathbf{A}(f) = 0 \quad \text{in } \mathbf{R}^N. \tag{2.5}$$

It is worth mentioning that classical variational approaches do not apply to (2.5) because $\mathbf{A}$ is not a potential operator.

We restrict our attention to the case of one-dimensional or radial geometry, where (2.5) is a fourth-order ODE, and in most cases we impose symmetry conditions at the origin $y = 0$ and a suitable decay condition (possibly exponential) at infinity:

$$f'(0) = 0, \ \ f'''(0) = 0, \quad \text{and } \ f(y) \to 0 \quad \text{as } \ y \to \infty. \tag{2.6}$$

Existence of nonradial solutions to (2.5) is an interesting open problem.

2.1. Conservative similarity solutions and the first critical exponent. Under appropriate decay conditions at infinity (say, exponential), (1.1) is *conservative*:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbf{R}^N} u \, \mathrm{d}x = - \int_{\mathbf{R}^N} \Delta(\Delta u + u^p) \, \mathrm{d}x = 0. \tag{2.7}$$

Given an exact similarity solution (2.4), and assuming that $f \in L^1(\mathbf{R}^N)$, we have that

$$\int_{\mathbf{R}^N} u_S(x,t) \, \mathrm{d}x = [\sigma(T-t)]^{-1/2(p-1)+N/4} \int_{\mathbf{R}^N} f(y) \, \mathrm{d}y, \tag{2.8}$$

which satisfies (2.7) for nonzero mass, $\int f \neq 0$, only if $p = p_0$, where

$$p_0 = 1 + \frac{2}{N} \tag{2.9}$$

is the *critical exponent* in the problem. It is interesting that p_0 coincides with the Fujita exponent for the semilinear heat equation (1.6). In fact, it is the *first* critical exponent and in section 6 we show that there exists a countable sequence of further critical exponents (1.8) corresponding to special cases of globally decaying solutions.

It follows from (2.9) that in the one-dimensional case, $N = 1$, the crucial critical case corresponds to $p = p_0 = 3$, where nonzero similarity masses can be preserved (section 4). However, we will consider other cases as well; see section 5. Moreover, from (2.8) we have

$$\text{for any } \ p \neq p_0, \quad f \in L^1 \quad \Longrightarrow \quad \int f = 0, \tag{2.10}$$

i.e., any L^1-similarity profile f has zero mass. Of course the same follows from the radial ODE (2.5) by integrating over $\mathbf{R}^N$.

2.2. Local asymptotic properties of self-similar solutions. Here we describe the possible asymptotics of small solutions to (2.5) satisfying $f(y) \to 0$ as $y \to \infty$, see (2.6). Consider the linearization of (2.5) about $f = 0$,

$$-f'''' - \frac{2(N-1)}{y} f''' - \frac{\mu}{y^2} f'' + \frac{\mu}{y^3} f' - \frac{\sigma}{4} y f' - \frac{\sigma}{2(p-1)} f = 0, \tag{2.11}$$

where $\mu = (N-1)(N-3)$. To determine the balance between the leading terms f'''' and $\frac{\sigma}{4} y f'$, we set $z = y^\alpha$ with $\alpha = \frac{4}{3}$ reducing the ODE to

$$-f'''' - a_1 f' - a_2 z^{-1} f + \mathbf{D}(z) f = 0. \tag{2.12}$$

Here $a_1 = \sigma/4\alpha^3$, $a_2 = \sigma/2(p-1)\alpha^4$ and $\mathbf{D}(z) f = \sum_{j=1}^{3} \gamma_j z^{j-4} f^{(j)}$ is a linear operator with bounded coefficients as $z \to \infty$, where the first coefficient of the derivative f' is of order $O(z^{-3})$. By the perturbation theory of linear ODEs (see Chapters III–V in [9]), we have that the leading terms of exponentially decaying solutions are described by the operator in (2.12) with constant coefficients,

$$-f'''' - a_1 f' = 0. \tag{2.13}$$

Setting $f = \mathrm{e}^{pz}$, $p \neq 0$, gives the characteristic equation $p^4 + a_1 p = 0$, when

$$p^3 = -a_1 = -\sigma/4\alpha^3 \equiv \rho_0^3(-\sigma), \quad \text{where } \rho_0 > 0. \tag{2.14}$$

There exist three roots of the form

$$p_k = \rho_0 \mathrm{e}^{i\pi(4k+\sigma+1)/6}, \quad k = 0, 1, 2, \tag{2.15}$$

where the number of roots with negative real part is two for $\sigma = -1$ and only one for $\sigma = +1$. Thus, as $y \to \infty$, there exists

(2.16) a two-dimensional stable bundle for the global case $\sigma = -1$, and

(2.17) a one-dimensional stable bundle for the blow-up case $\sigma = 1$.

This difference in the dimensionality of the stable manifold about $f = 0$ at $y = \infty$ is what distinguishes the continuous spectrum of solutions for the global case from the discrete set of blow-up solutions.

On the other hand, (2.12) also admits solutions with algebraic decay (rather than exponential) as $z \to \infty$ described by the first-order operator

$$-a_1 f' - a_2 z^{-1} f = 0 \quad \Longrightarrow \quad f(z) = A z^{-2/\alpha(p-1)}, \ \ A \neq 0.$$

Existence of solutions with such decay for the perturbed equation (2.12) is established by a standard expansion analysis by calculating solutions via a Kummer-type series converging uniformly for $z \gg 1$. For the linearized equation (2.11), the leading order behavior is algebraic,

$$f(y) = A|y|^{-2/(p-1)}(1 + o(1)) \ \ \text{as} \ \ |y| \to \infty, \quad f \in L^1 \ \ \text{iff} \ \ p < p_0, \tag{2.18}$$

with arbitrary constant $A \neq 0$. For blow-up similarity solutions (2.4), the limit-time profile is then bounded for any $x \neq 0$ and has the form

$$u_S(x, T^-) = A|x|^{-2/(p-1)}, \quad x \neq 0.$$

Clearly, for $p \geq p_0$, this is not an admissible solution for the conservative case given finite initial mass, and thus in the critical case $p = p_0$ we must have $A = 0$ and exclusively exponential decay of any similarity profiles.

3. The spectral properties of the rescaled linear operators. The structure of the rescaled equation (2.3) suggests the study of the spectral properties of the linearized operator with $\sigma = -1$ (global solutions), which is conveniently represented in a form associated with the operator in (1.13),

$$\mathbf{A}'(0) = -\Delta^2 + \frac{1}{4}\, y \cdot \nabla + \frac{1}{2(p-1)}\, I \equiv \mathbf{B} + c_* I, \quad c_* = \frac{N}{4(p-1)}(p_0 - p). \tag{3.1}$$

The spectral properties of $\mathbf{B}$ and the corresponding adjoint operator $\mathbf{B}^*$ occurring for $\sigma = 1$ will play an important role in the further asymptotic analysis of the nonlinear PDE. The operators are defined in weighted L^2-spaces with the weight functions induced by the exponential estimate of the rescaled kernel (1.14).

3.1. The point spectrum of the non-self-adjoint operator B. Note that $\mathbf{B}$ is not symmetric and does not admit a self-adjoint extension. We consider $\mathbf{B}$ in the weighted space $L^2_\rho = L^2_\rho(\mathbf{R}^N)$ with the exponentially growing weight function

$$\rho(y) = \mathrm{e}^{a|y|^\alpha} > 0 \quad \text{in } \mathbf{R}^N, \quad \text{where } \alpha = \frac{4}{3}, \tag{3.2}$$

and $a \in (0, 2d)$ (d is as in (1.14)) is a sufficiently small constant. We ascribe to $\mathbf{B}$ the domain H^4_ρ being a weighted Hilbert space with the norm

$$\|v\|^2 = \int \rho(y) \sum_{k=0}^{4} |D^k v(y)|^2 \, \mathrm{d}y,$$

induced by the corresponding inner product. From [10] we have the following result.

LEMMA 3.1. *$\mathbf{B} : H^4_\rho \to L^2_\rho$ is a bounded linear operator with the real spectrum*

$$\sigma(\mathbf{B}) = \left\{ \lambda_l = -\frac{l}{4},\ l = 0, 1, 2, \dots \right\}. \tag{3.3}$$

The eigenvalues λ_l have finite multiplicity with eigenfunctions

$$\psi_\beta(y) = \frac{(-1)^{|\beta|}}{\sqrt{\beta!}}\, D^\beta F(y), \quad \textit{with any } |\beta| = l, \tag{3.4}$$

where F is the rescaled kernel in (1.12)*, and the set of eigenfunctions $\Phi = \{\psi_\beta,\ |\beta| = 0, 1, 2, \dots\}$ is complete in L^2_ρ.*

Lemma 3.1 gives the center and stable subspaces of $\mathbf{B}$, $E^c = \mathrm{Span}\{\psi_0 = F\}$ and $E^s = \mathrm{Span}\{\psi_\beta,\ |\beta| > 0\}$.

3.2. The polynomial eigenfunctions of the adjoint operator $\mathbf{B}^*$. Consider the operator adjoint to $\mathbf{B}$,

$$\mathbf{B}^* = -\Delta^2 - \frac{1}{4}\, y \cdot \nabla, \tag{3.5}$$

which is related to $\mathbf{A}'(0)$ in (2.3) for $\sigma = 1$ (the blow-up case). We consider $\mathbf{B}^*$ in $L^2_{\rho^*}$ with the exponentially decaying weight function $\rho^*(y) = 1/\rho(y) \equiv \mathrm{e}^{-a|y|^\alpha} > 0$.

LEMMA 3.2. *$\mathbf{B}^* : H^4_{\rho^*} \to L^2_{\rho^*}$ is a bounded linear operator with the same spectrum,* (3.3)*. The eigenfunctions $\psi^*_\beta(y)$ with $|\beta| = l$ are lth order polynomials*

$$\psi^*_\beta(y) = \frac{1}{\sqrt{\beta!}} \left[y^\beta + \sum_{j=1}^{\lfloor |\beta|/4 \rfloor} \frac{1}{j!} (-\Delta)^{2j} y^\beta \right], \tag{3.6}$$

and the set $\{\psi^*_\beta\}$ is complete in $L^2_{\rho^*}$.

With this definition of the adjoint eigenfunctions, the orthonormality condition

$$\langle \psi_\beta, \psi^*_\gamma \rangle = \delta_{\beta,\gamma} \tag{3.7}$$

holds, where $\langle \cdot, \cdot \rangle$ denotes the standard (dual) L^2 inner product.

4. Blow-up similarity profiles for $p = 3$ in one dimension. The dynamics simplify for the conservative case $p = p_0 = 1 + \frac{2}{N}$ as (2.5) can be integrated once. The particular critical case

$$p = p_0 = 3 \quad \text{with} \ \ N = 1 \tag{4.1}$$

is of key importance in our analysis and highlights the typical techniques required to describe a countable set of self-similar blow-up patterns. Throughout this section we assume that (4.1) holds.

4.1. Preliminaries. In the case (4.1), the ODE (2.5) with $\sigma = 1$ for the similarity profile can be integrated once to give

$$f''' + \frac{1}{4} y f + (f^3)' = 0 \quad \text{for} \ \ y > 0, \quad f'(0) = 0, \tag{4.2}$$

where we have set the constant of integration to zero on the right-hand side. This removes the algebraically decaying mode (2.18), $f(y) = Ay^{-1}(1 + o(1))$, by looking for L^1-solutions satisfying the conservation of mass condition (2.7). It follows from (2.17) that we are left with a *one-parameter* family of exponentially decaying functions satisfying

$$f(y) = Cy^{-1/3} \mathrm{e}^{-\beta y^{4/3}} (1 + o(1)) \quad \text{as} \ y \to +\infty, \quad C \geq 0, \quad \beta = 3/4^{4/3}, \tag{4.3}$$

where the extra algebraic factor $y^{-1/3}$ is detected by a standard refined asymptotic WKBJ expansion according to the ODE theory, [9]. This asymptotic expansion fully corresponds to the linear part of the operator in (4.2) and does not depend on the cubic nonlinear term. In order to construct a solution to (4.2) we will use a shooting type argument starting from $y = \infty$. Thus, all admissible similarity profiles $f(y)$ have the asymptotic behavior (4.3) and also should satisfy the symmetry condition at the origin.

Let us now study the behavior of solutions on the manifold (4.3) parameterized by C. Denoting $f(y; C)$ as the function which satisfies (4.2) with decay from the bundle (4.3), the goal is to find the set of C such that

$$S(C) \equiv f'(0; C) = 0. \tag{4.4}$$

Because the function $S(C)$ in (4.4) is analytic in C (see below), (4.4) has at most a countable set of roots, which can accumulate at $C = \infty$ only and this actually happens. We begin with the global existence for the family $\{f(y; C)\}$.

LEMMA 4.1. *For any $C > 0$, the solution $f(y; C)$ to* (4.2), (4.3) *is well defined for all $y \in \mathbf{R}$.*

Proof. This follows from the local properties of the operator in (4.2) which are close to those for the second-order case $f'' + f^3 = 0$. Obviously, it does not admit blow-up of solutions at finite y. Integrating (4.2), $f'' = -f^3 - \frac{1}{4} \int f y \, \mathrm{d}y$, multiplying

by f' and integrating again over a sufficiently small interval near any fixed point y_0 yields

$$\begin{aligned}\frac{1}{2}\,f'^2(y) &= -\frac{1}{4}\,f^4(y) - \frac{1}{4}\int_{y_0}^{y} f'(z)\int_{y_0}^{v} f(v)v\,\mathrm{d}v\mathrm{d}z + \mathrm{const}\\ &\le \frac{1}{4}\int_{y_0}^{y} |f'(z)|\int_{y_0}^{v} |f(v)v|\,\mathrm{d}v\mathrm{d}z + \mathrm{const},\end{aligned} \tag{4.5}$$

where, by Lagrange's formula of finite increments, $|f(v)| \le |f(y_0)| + (\sup_v |f'(v)|)|v - y_0|$, the right-hand side is not more than quadratic in f'. Therefore, $f(y)$ cannot blow-up at a finite point y_* along a sequence since (4.5) guarantees that $f'(y_*)$ is finite. Obviously, this analysis applies for all $p > 1$. ∎

4.2. Existence of the first monotone blow-up similarity pattern. The proof that there exists a first, minimal $C = C_1 > 0$ such that (4.4) holds involves three steps. First, we show that in the limit as $C \to 0^+$ solutions $f(y;C)$ are strictly monotone decreasing on $[0,\infty)$ and, second, that in the other limit $C \to +\infty$ all solutions cannot be monotone. Then, by a standard continuity argument, we have that there must exist an intermediate $C = C_1 > 0$ corresponding to an admissible monotone solution.

PROPOSITION 4.2. *For all $0 < C \ll 1$, solutions $f(y;C)$ are strictly monotone decreasing in $y \ge 0$ and $f'(0;C) < 0$.*

Proof. Rescaling by setting $f = Cg$, we have

$$g''' + \frac{1}{4}yg + C^2(g^3)' = 0 \quad \text{in } \mathbf{R}_+. \tag{4.6}$$

In view of the behavior at infinity (4.3), by standard results on continuous dependence for ODEs [9], it follows that, as $C \to 0^+$,

$$f(y;C) = C(\phi_0(y) + o(1)) \quad \text{uniformly in } \mathbf{R}_+, \tag{4.7}$$

where ϕ_0 solves the linear ODE $\phi''' + \frac{1}{4}\,\phi' y = 0$ and satisfies (4.3) with $C = 1$. All the derivatives of f converge similarly. Let us now show that $\phi_0(y)$ is strictly monotone decreasing. Assume that y_i is the first (from $y = \infty$) local maximum point of ϕ_0, $\phi_0'(y_i) = 0$ and $\phi_0'(y) < 0$ on (y_i,∞). Integrating over (y_i,∞), we obtain the contradiction, $\phi_0''(y_i) = \int_{y_i}^{\infty} s\phi_0(s)\,\mathrm{d}s > 0$, thus observing an elementary feature of the Maximum Principle for such third-order equations. ∎

PROPOSITION 4.3. *There exists a $C_* > 0$ such that $f(y;C_*)$ has a local maximum point at some $y_* \ge 0$ and $f(y;C_*) > 0$ on $[y_*,\infty)$.*

Proof. Assume for contradiction that $f(y;C)$ is strictly monotone decreasing in $\mathbf{R}_+$ for all $C > 0$, and then $f(0;C) > 0$. One can see from the rescaled ODE (4.6) that the solutions $f(y;C)$ cannot be bounded for $y \in \mathbf{R}_+$ uniformly in $C > 0$. Therefore, there exists a sequence $\{C_k\} \to \infty$ such that $a_k = f(0;C_k) \to \infty$. Performing the scaling

$$f(y) = a_k g_k(z), \quad y = z/a_k, \tag{4.8}$$

we arrive at a perturbed ODE for the sequence $\{g_k(z)\}$,

$$g_k''' + (g_k^3)' = -\frac{1}{4a_k^4}\,zg_k \quad \text{for } z > 0, \quad g_k(0) = 1, \tag{4.9}$$

where $0 < g_k(z) \le 1$ and $g_k(z)$ is monotone decreasing in z. Since $\{g_k\}$ is a uniformly bounded sequence of solutions of the asymptotically perturbed ODE (4.9) with regular coefficients, by the Ascoli–Arzelá theorem and standard ODE estimates [9], we have that along a subsequence, $g_k \to \bar{g}$ uniformly on compact subsets in z, where $\bar{g}$, $0 \le \bar{g} \le 1$, must be a monotone decreasing solution of the unperturbed ODE

$$(\bar{g}'' + \bar{g}^3)' = 0, \quad \bar{g}(0) = 1. \tag{4.10}$$

Since all such solutions of (4.10) are oscillatory, which is easily checked by integrating twice, this leads to a contradiction. Therefore, there exists a sufficiently large $C_* > 0$ such that $f(y; C_*)$ is not strictly monotone in $\mathbf{R}_+$ and has a local maximum point. □

THEOREM 4.4. *There exists a constant $C_1 \in (0, C_*]$ such that (4.4) holds for $C = C_1$ and $f_1(y) = f(y; C_1)$ is a strictly monotone decreasing symmetric positive similarity profile.*

Proof. Introducing the set

$$W_1 = \{\mu > 0 : \ f(y; C) \text{ is strictly monotone decreasing in } \mathbf{R}_+ \text{ for all } C \in (0, \mu)\},$$

we have that $W_1 \ne \emptyset$ by Proposition 4.2 and W_1 is bounded above by Proposition 4.3. Hence, there exists

$$C_1 = \sup W_1 \le C_*, \tag{4.11}$$

where, by construction, $f_1(y) = f(y; C_1)$ is monotone decreasing for $y \ge 0$. By the definition of supremum in (4.11), one can see that $f_1'(y)$ must vanish at the origin, i.e., (4.4) holds. □

We now briefly describe a countable set of the similarity profiles satisfying (4.2) and (4.3) with some $C > C_1$. This construction is similar to that for the second-order ODEs from blow-up reaction-diffusion theory; see [34, pp. 190–195] and related references therein. Similar to Proposition 4.3 we show that as C increases, the function $f(y; C)$ becomes more and more oscillatory for $y > 0$. Indeed, performing the scaling (4.8) and passing to the limit $C = C_k \to \infty$, we obtain a bounded solution $\bar{g}$ satisfying (4.10) admitting oscillatory solutions only. Hence, $f(y; C)$ can have an arbitrarily large number of oscillations for $y > 0$ with $C \gg 1$. As in the proof of Theorem 4.4, for any $k = 2, 3, \dots$, we define the sets

$$W_k = \{\mu > C_{k-1} : \ f(y; C) \text{ has at most } k \text{ local extrema for } y \ge 0 \, \forall \, C \in [C_{k-1}, \mu)\}.$$

Then, once C_{k-1} is known, starting with $k = 2$, we have that $W_k \ne \emptyset$ and by the oscillatory behavior for $C \gg 1$, W_k is bounded from above. Hence, we define

$$C_k = \sup W_k > C_{k-1}, \quad k = 2, 3, \dots, \tag{4.12}$$

and by construction, $f_k(y) = f(y; C_k)$ satisfies the symmetry condition at the origin (otherwise, it is not the supremum in (4.12)).

Such a construction, while giving an infinite sequence of similarity profiles, does not describe the important properties of the functions $f_k(y)$, such as positivity (or at least "positivity dominance") and the actual distribution of local extrema and inflection points of the profiles. This will be done by a more delicate and partially formal asymptotic matching procedure.

4.3. Asymptotic construction of a countable set of similarity profiles. Here we will use the method of matched asymptotics expansions to describe the solution structure of the similarity profiles $\{f_k(y)\}$ for large k, i.e., in the limit $C_k \to \infty$. We will show that the similarity profiles will be composed of three regions. The two primary regions are an *outer region* localized near the origin, where the mass of the solution is concentrated with oscillations around a parabolic profile (see (4.20) below), and a *far-field region*, in which the solution is described by the asymptotic bundle (4.3). Joining these two regions is a narrow transition or *inner* region. The details containing the scalings and corresponding matching are given below. Because the characteristics of the family are determined by the far-field behavior, we will begin by considering refined asymptotics as $y \to \infty$.

The far-field behavior. To determine the asymptotic behavior as $y \to \infty$ in greater detail, we produce an asymptotic description of solutions from the exponential bundle (4.3). Specifically, the solution is given by the formal power series

$$f(y;C) = \sum_{n=0}^{\infty} C^{2n+1}\phi_n(y), \tag{4.13}$$

where ϕ_0 is given in (4.7) and the rest of the terms are obtained from the relation

$$\mathbf{B}^*\phi_n \equiv \phi_n''' + \frac{1}{4}\,y\phi_n = -\sum_{\substack{1\le i,j,k\le n\\ i+j+k=n+1}} (\phi_i\phi_j\phi_k)', \quad n = 2,3,\dots, \tag{4.14}$$

with the condition $\phi_n(y) = o(\phi_0(y))$ as $y \to \infty$. Using the expansion (4.3), it can be shown by direct calculation that for $y \gg 1$,

$$\phi_n(y) = \left[\gamma_n + o(y^{-1})\right]\, y^{-(4n+1)/3}\mathrm{e}^{-(2n+1)\beta y^{4/3}}, \tag{4.15}$$

with suitable constants $|\gamma_n| \le 1$. Moreover, the right-hand side of (4.15) with a sufficiently large constant γ_n gives uniform estimates of ϕ_n in $\mathbf{R}_+$, which establishes the uniform convergence of (4.13) on subsets $\{y \ge y_0\}$ with $y_0 \gg 1$. By the Weierstrass theorem, the solution $f(y;C)$ in (4.13) obtained as a uniformly converging series of analytic functions is analytic in C for $y \ge y_0$. Therefore, on extension to $y \in [0, y_0)$, as a solution of an ODE with analytic coefficients and analytic dependence on C in the Dirichlet boundary condition at $y = y_0$, it is analytic in C for any $y \ge 0$ (cf. typical analyticity results in the classical ODE theory, [9, section 8, Chapter I]). Hence, (4.4) is an analytic function having isolated zeros only and we arrive at the following conclusion.

PROPOSITION 4.5. *The problem* (4.2), (4.3) *has at most a countable set of solutions.*

The discrete nature of the function $S(C)$ will be made evident by close examination of the inner and transition regions.

A singular perturbation problem. We begin by looking for possible solutions localized in a neighborhood of the origin $y = 0$. We rescale the ODE (4.2) as (cf. (4.8))

$$f(y) = ag(z), \quad y = az, \quad \text{where } a = f(0;C), \tag{4.16}$$

with a an as yet unspecified function of C to be determined for the similarity profiles $f_k(y)$. Under the assumption that $a(C) \to \infty$ as $C \to \infty$, possibly along a

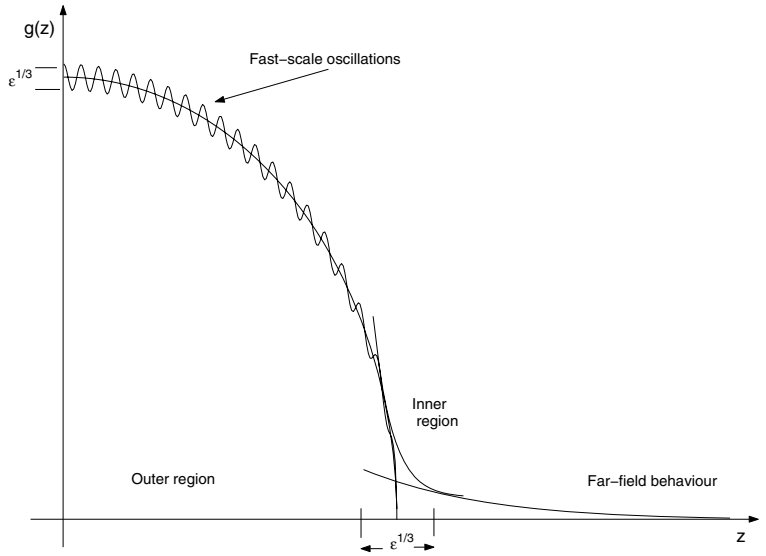


FIG. 1. *Sketch of the regions for construction of asymptotic profiles of* (4.17).

subsequence, we define

$$\varepsilon = a^{-4} \ll 1 \quad \text{for } C \gg 1.$$

Under the rescaling (4.16), (4.2) leads to a singularly perturbed ODE

$$\varepsilon g''' + \frac{1}{4} z g + (g^3)' = 0. \tag{4.17}$$

We need to describe a two-parameter family of solutions of (4.17) defined in a suitable neighborhood of the origin $z = 0$, which can be matched with the exponential bundle (4.3). In order to describe such a family, we use the method of matched asymptotic expansions and supplement the ODE (4.17) with two standard conditions

$$g(0) = 1, \quad g'(0) = 0. \tag{4.18}$$

In addition, we have

$$g''(0) = b\varepsilon^{-m}(1 + o(1)), \tag{4.19}$$

where the constant b and parameter m are as yet unspecified. The third condition (4.19) is introduced for convenience only and we show in what follows that both b and m are determined later by the matching procedure. However, their inclusion here is to emphasize that in the outer region, a faster scale is necessary in order to resolve the additional oscillatory structure as described in the proof of Proposition 4.3 by the perturbed problem (4.9).

We will show that the solution to (4.17) has a three layer structure, schematically illustrated in Figure 1, comprising

(i) an *outer region* $0 \le z < \sqrt{12}$ in which $g = O(1)$,

(ii) an *inner region* $z = \sqrt{12} + O(\varepsilon^{1/3})$ in which $g = O(\varepsilon^{1/6})$, and finally

(iii) a *third region* $z > \sqrt{12}$, where g is exponentially small (and belongs to the exponential bundle described in section 2).

The outer region. Let $g_0(z)$ be the solution of the unperturbed problem (4.17), namely,

$$\frac{1}{4} z g_0 + (g_0^3)' = 0 \quad \text{with} \quad g_0(0) = 1 \implies g_0(z) = \sqrt{1 - \frac{1}{12}\, z^2}. \tag{4.20}$$

Close to $z = 0$, we perform the linearization

$$g(z) = g_0(z) + G(Z), \tag{4.21}$$

with the fast variable $Z = z/\varepsilon^{1/2}$. Then G satisfies the equation

$$G''' + (3 g_0^2 G)' + \frac{1}{4} \varepsilon Z G + (3 g_0 G^2 + G^3)' + \varepsilon^{3/2} g_0''' = 0, \tag{4.22}$$

where $g_0^2(Z) = 1 - \frac{\varepsilon}{12}\, Z^2$. It follows that on compact subsets $0 \le Z \le c$, the principal part of (4.22) has constant coefficients,

$$G''' + 3G' + \mathcal{N}_1 = 0, \quad G(0) = G'(0) = 0, \tag{4.23}$$

where $\mathcal{N}_1 = O(\varepsilon G + G^2 + \varepsilon^2)$ on bounded smooth solutions. Therefore, by standard perturbation ODE theory [9], recalling the third condition in (4.18), the solution satisfies the following expansion as $\varepsilon \to 0$ uniformly on subsets $Z \in [0, c]$:

$$G(Z) = \theta \left(1 - \cos\left(\sqrt{3}\, Z\right)\right) + o(\theta), \tag{4.24}$$

where $\theta = \frac{1}{3}\, b \varepsilon^{1-m}$. This rigorously determines the asymptotic behavior in the region $\{z = O(\varepsilon^{1/2})\}$. For matching reasons, we need to extend the expansion (4.24) to larger subsets. This is done by introducing the new independent variable

$$t = t(Z) \equiv \int_0^Z \left(1 - \varepsilon \frac{u^2}{12}\right)^{1/2} \mathrm{d}u = \frac{1}{\varepsilon^{1/2}} \left(\frac{z}{2} \left(1 - \frac{z^2}{12}\right)^{1/2} + \sqrt{3} \sin^{-1}\left(\frac{z}{2\sqrt{3}}\right)\right), \tag{4.25}$$

and setting $G(Z) = P(t)/g_0(Z)$. This gives, similar to (4.23), an equation with constant coefficients in the principal part

$$P''' + 3P' + \mathcal{N}_2 = 0, \quad P(0) = P'(0) = 0, \tag{4.26}$$

where the perturbation $\mathcal{N}_2$, in addition to the same small terms as above from (4.23), now contains differential operators with coefficients depending on the derivatives of $g_0 = g_0(z(t))$ satisfying

$$\frac{\mathrm{d}g_0}{\mathrm{d}t} = \frac{g_0'}{t'} = O(\varepsilon^{1/2}), \quad \frac{\mathrm{d}^2 g_0}{\mathrm{d}t^2} = O(\varepsilon), \quad \frac{\mathrm{d}^3 g_0}{\mathrm{d}t^3} = O(\varepsilon^{3/2})$$

uniformly on compact subsets in $z \in [0, c]$ for any $c < \sqrt{12}$. Therefore, upon returning to the original variables $\{z,\, g\}$, the outer expansion takes the form

$$g(z) = g_0(z) + \theta \frac{1}{g_0(z)} \left(1 - \cos\sqrt{3} t\right) + \ldots \equiv g_0 + \theta g_1 + \ldots, \quad \theta = \frac{1}{3} b \varepsilon^{1-m}. \tag{4.27}$$

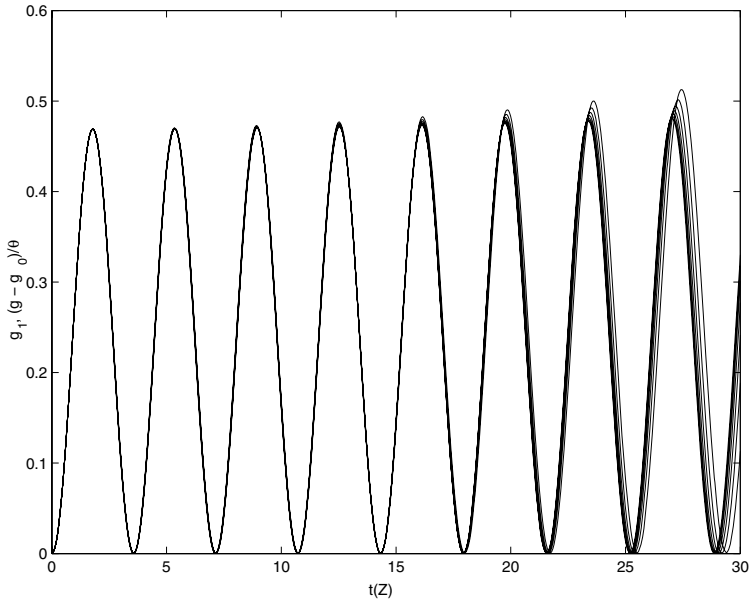


FIG. 2. *Convergence as* $\varepsilon \to 0$ *of numerical solutions of* (4.17) *to* $g_1(z,t)$ *given in* (4.27)*;* $\varepsilon \in [7.51 \cdot 10^{-4}, 1.08 \cdot 10^{-1}]$.

Note that we cannot guarantee that this expansion is valid uniformly on subsets $z \in [0, c]$ since though the principal part in (4.26) is exactly the same as in (4.23), equation (4.26) is then considered on expanding subsets in $t \in [0, c\varepsilon^{-1/2}]$ becoming unbounded as $\varepsilon \to 0$. It is worth remarking that the crucial fast scale variable (4.25) is determined by the requirement that (4.26) has constant coefficients, which fixes it to the order considered here. For first-order matching, it suffices to prescribe only the rate of deviation $O(g_0^{-1})$ of $g(z)$ from the unperturbed profile $g_0(z)$ as correctly described by the second term in (4.27). Figure 2 shows the accuracy of expansion (4.27), where we have already fixed $m = \frac{2}{3}$ (to be determined later) so that $\theta \sim \varepsilon^{1/3}$. Plotted is the first-order term g_1 from (4.27) with the full numerical solution for g scaled in the form $(g - g_0)/\varepsilon^{1/3}$.

The inner region. The outer expansion (4.27) does not reflect the far-field behavior (4.3). Thus, in order to specify the behavior for $z \approx \sqrt{12}$, we introduce the inner variables $\bar{z}$ and $R(\bar{z})$ by

$$z = \sqrt{12} + \varepsilon^{1/3}\bar{z}, \qquad g = \varepsilon^{1/6} R, \quad \text{where } R \text{ solves} \tag{4.28}$$

$$R''' + \frac{1}{4}\left(\sqrt{12} + \varepsilon^{1/3}\bar{z}\right) R + \left(R^3\right)' = 0. \tag{4.29}$$

The unperturbed equation

$$R_0''' + \frac{1}{2}\sqrt{3} R_0 + \left(R_0^3\right)' = 0 \tag{4.30}$$

admits a one-parameter family of suitable exponentially decaying solutions

$$R_0(\bar{z}) \simeq A_0 \mathrm{e}^{-3^{1/6}\bar{z}/2^{1/3}} \qquad \text{as } \ \bar{z} \to +\infty, \tag{4.31}$$

for some constant A_0, and a three-parameter family of slowly growing solutions as $\bar{z} \to -\infty$. Solutions from the bundle (4.31) satisfy

$$R_0(\bar{z}) \simeq \left(-\frac{\bar{z}}{3^{1/2}}\right)^{1/2} + \frac{\bar{a}_1}{(-\bar{z})^{1/2}}\left[1 - \bar{a}_2 \cos\left(c(-\bar{z})^{3/2}\right) - \bar{a}_3 \sin\left(c(-\bar{z})^{3/2}\right)\right], \tag{4.32}$$

where $c = 2/3^{3/4}$ and $\bar{a}_1$, $\bar{a}_2$, $\bar{a}_3$ are suitable constants. This expansion is determined in the standard way by setting $R_0 = \tilde{R}_0 + \tilde{R}_1$, where $\tilde{R}_0 = (-\bar{z})^{1/2}/3^{1/4}$ under the assumption that $|\tilde{R}_1| \ll |\tilde{R}_0|$. Then the leading linear equation for $\tilde{R}_1$ is

$$\tilde{R}_1''' + \frac{\sqrt{3}}{2}\tilde{R}_1 - \frac{1}{\sqrt{3}}\left(\bar{z}\tilde{R}_1\right)' = 0, \tag{4.33}$$

which has an exact solution in terms of the Airy functions Ai and Bi,

$$\tilde{R}_1 = c_1 Ai^2(\bar{z}) + c_2 Ai(\bar{z})Bi(\bar{z}) + c_3 Bi^2(\bar{z}), \quad c_1, c_2, c_3 \in \mathbf{R}. \tag{4.34}$$

This fact is important for the reliable numerical integration of (4.30).

Matching. To determine the sequence $\{C_k,\, k \gg 1\}$ and the corresponding values $f(0; C_k) = a(C_k)$, we now match the inner and outer solutions. Expanding the components of the outer solution (4.27) near the transition point $z = \sqrt{12}$, we have

$$g_0 \simeq \varepsilon^{1/6}\left(-\frac{\bar{z}}{3^{1/2}}\right)^{1/2}\left(1 + \varepsilon^{1/3}\frac{\bar{z}}{8 \cdot 3^{1/2}}\right), \tag{4.35}$$

(4.36)

$$g_1 \simeq \varepsilon^{-1/6}\frac{b}{3}\left(-\frac{3^{1/2}}{\bar{z}}\right)^{1/2}\left[1 - \cos\frac{\sqrt{3}t_0}{\varepsilon^{1/2}}\cos\left[c(-\bar{z})^{3/2}\right] - \sin\frac{\sqrt{3}t_0}{\varepsilon^{1/2}}\sin\left[c(-\bar{z})^{3/2}\right]\right],$$

where $t_0 = \int_0^{\sqrt{12}} \sqrt{1 - v^2/12}\,\mathrm{d}v = \frac{\sqrt{3}}{2}\pi$. Comparing the outer expansion (4.27), using (4.35) and (4.36) with the inner expansion $g = \varepsilon^{1/6}R_0 + \ldots$, where R_0 is given by (4.32), yields the matching condition $1 - m - \frac{1}{6} = \frac{1}{6}$, if full balance is to be obtained, i.e.,

$$m = \frac{2}{3}, \tag{4.37}$$

together with the coefficients

$$\bar{a}_1 = 3^{-3/4}b, \quad \bar{a}_2 = \cos\frac{3^{1/2}t_0}{\varepsilon^{1/2}} \quad \text{and} \quad \bar{a}_3 = \sin\frac{3^{1/2}t_0}{\varepsilon^{1/2}} \quad \left(t_0 = 3\sqrt{\pi}/2\right). \tag{4.38}$$

It follows that (4.37) and the first equality in (4.38) are established rigorously since we have used the "envelope" characteristic in expansion (4.27) without specifying the highly oscillatory component. In particular, (4.38) demands

$$\bar{a}_2^2 + \bar{a}_3^2 = 1. \tag{4.39}$$

To determine these values, we observe that (4.30) is translation invariant, as is the asymptotic behavior for $\bar{z} \to \infty$, but not (to leading order) as $\bar{z} \to -\infty$. To see this, we consider the solution to (4.30) as a shooting problem in the single parameter A_0.

However,

$$R_0(\bar{z}; \hat{A}_0) = R_0(\bar{z} - \bar{z}_0; A_0), \quad \text{where} \quad \hat{A}_0 = A_0 \mathrm{e}^{3^{1/6}2^{-1/3}\bar{z}_0}.$$

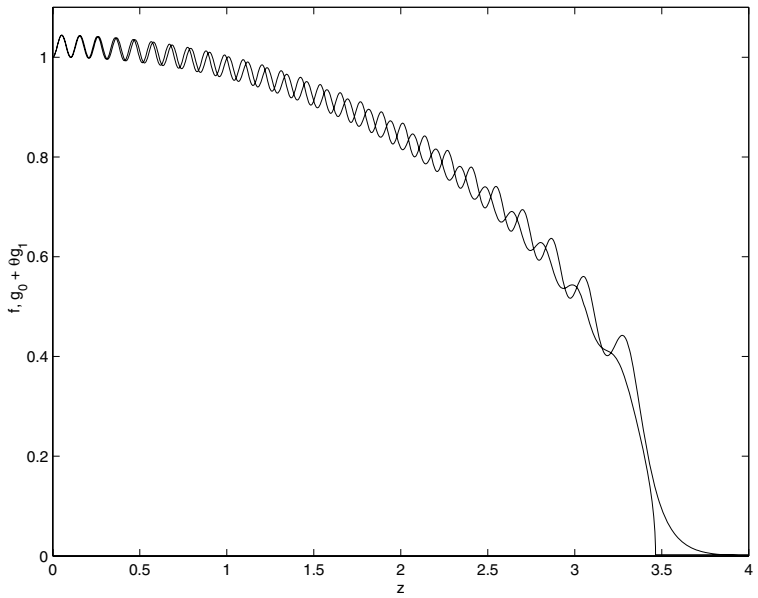


FIG. 3. *Comparison of numerical and asymptotic* (4.27) *solutions of the ODE* (4.17). *The numerical profile* $g(z)$ *exhibits the exponential decay for* $y \approx 4$.

Clearly this translation does not change the ODE although it will affect the coefficients in the far-field behavior (4.32). Denoting the new coefficients by $\hat{a}_i$, $i = 1, 2, 3$, we have to leading order that

$$\hat{a}_1 = \bar{a}_1 + \bar{z}_0/2 \cdot 3^{1/4}, \quad \hat{a}_2 = \bar{a}_2, \quad \hat{a}_3 = \bar{a}_3,$$

i.e., the translation only affects the coefficient $\bar{a}_1$ in (4.32). Thus,

$$\hat{A}_0 = A_0 \mathrm{e}^{3^{5/12} 2^{2/3} (\hat{a}_1 - \bar{a}_1)}. \tag{4.40}$$

We now seek A_0 such that the form (4.32) satisfies the constraint (4.39). From the translation argument we have that $\bar{a}_1$ is monotone decreasing in A_0, and it follows from (4.33), (4.32) that there are precisely two possible solutions which are the negatives of each other, i.e., $\bar{a}_1$ and $-\bar{a}_1$. Taking $\hat{a}_1 = -\bar{a}_1$, (4.40) gives

$$\hat{A}_0 = A_0 \mathrm{e}^{-2^{5/3} 3^{5/12} \bar{a}_1}. \tag{4.41}$$

Numerical computation gives

$$A_0 \approx 0.448\ldots, \quad \bar{a}_1 = 0.341\ldots$$

and from (4.41) we deduce the other set of values

$$A_0 \approx 0.081\ldots, \quad \bar{a}_1 = -0.341\ldots.$$

Finally, (4.38) may be used to determine the two values of b. Using the value $\bar{a}_1 \approx 0.341\ldots$ in the asymptotic expression (4.27) gives very good agreement to full numerical solutions near $z = 0$, as seen in Figure 3. In particular, the amplitude of the fast-scale oscillation is excellent.

The third region. It follows from (4.3) and (4.16) that

$$g(z) \simeq C\varepsilon^{1/3} z^{-1/3} \mathrm{e}^{-\beta z^{4/3}/\varepsilon^{1/3}} \tag{4.42}$$

is the appropriate behavior satisfying (4.17) for $z > \sqrt{12}$. Writing (4.42) in terms of the inner variable $\bar{z}$ from (4.28) gives the leading order expression

$$g(z) \simeq C\varepsilon^{1/3} 12^{-1/6} \mathrm{e}^{-\beta 12^{2/3}/\varepsilon^{1/3}} \mathrm{e}^{-3^{1/6}\bar{z}/2^{1/3}}$$

after using $\varepsilon^{1/3}\bar{z} = O(1)$. Comparing to the inner behavior (4.31) and using (4.28) gives the matching condition

$$A_0 \simeq C 12^{-1/6} \varepsilon^{1/6} \mathrm{e}^{-\beta 12^{2/3}/\varepsilon^{1/3}}. \tag{4.43}$$

Since both values of A_0 are of order $O(1)$, then

$$\varepsilon \sim \frac{3^5}{2^4}\frac{1}{(\ln C)^3} + \frac{3^6}{2^5}\frac{\ln\ln C}{(\ln C)^4} + \frac{3^6}{2^4}\frac{\ln(3^{-2/3}2A_0)}{(\ln C)^4} \quad \text{as } C \to \infty.$$

Recall that all the estimates obtained without using the highly oscillatory tails in the singular layer given by (4.27) are rigorous and asymptotically sharp. The fact that there are two values of A_0 and thus two relationships between C and ε corresponds to the two families of solutions, those with $g''(0) > 0$ or $g''(0) < 0$. Most importantly, given that asymptotically $\bar{a}_2 = \cos(3\pi/2\varepsilon^{1/2})$ and $\bar{a}_3 = \sin(3\pi/2\varepsilon^{1/2})$ in (4.38) are *unique* (up to a change in sign in both coefficients), we have that

$$3\pi/2\varepsilon^{1/2} = k\pi \pm \phi\pi(1 + o(1)), \tag{4.44}$$

for an unknown phase shift ϕ (this is a formal conclusion). Using this in the definition of $\varepsilon = a^{-4}$ gives

$$a_k = \sqrt{2k/3}(1 + o(1)) \quad \text{as } k \to \infty.$$

Also, from (4.43) this fixes the corresponding values of the expansion coefficients $C = C_k$ in (4.3) in the outer region

$$C_k \sim k^{1/3} \mathrm{e}^{3k^{2/3}/2^{2/3}} \quad \text{for } k \gg 1.$$

Using the computed values of the coefficients and expansions, we present a comparison of the numerical and asymptotic solutions in Figure 3. The slow phase-shift can be corrected by further terms in the expansion but, even to this order, we have captured the correct number of fast-scale oscillations. With $\varepsilon = 0.00083\ldots$, using only the Z variable instead of the $t(Z)$ variable given in (4.25) over-predicts the number of oscillations seen in Figure 3 by about one half.

4.4. Numerical methods.

Solution of the full ODEs. In (4.4) we introduced the shooting function S whose zeroes correspond to admissible blow-up profiles. By numerically integrating the ODE (4.2) with the far-field behavior (4.3), we have been able to approximate this function, as seen in Figure 4. An asymptotic form of the function $S(C)$ for $C \gg 1$ can be constructed by a similar matching argument by putting an extra term $\tilde{b}\sin(\sqrt{3}t)/g_0(z)$ into (4.27) ($\tilde{b} = 0$ iff $S(C) = 0$). This leads to a harder multi-parameter matching

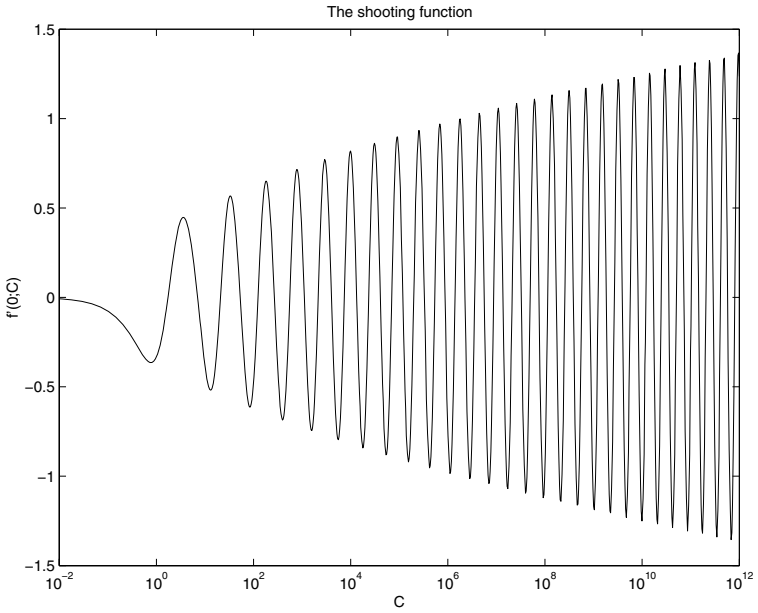


FIG. 4. *Numerical approximation to* $S(C)$ *in* (4.4).

procedure and will not be considered any further. Labeling, as above, the solutions to (4.2) corresponding to the nth zero of $S(C)$ as f_n, we find that $f_n(y)$ has precisely n maxima on $\mathbf{R}$. In Figure 5 we present the first profile f_1 given in Theorem 4.4, which has the simplest bell-shaped form. It is evolutionary stable for a broad range of initial data (see the next section). For visual clarity we separately present the profiles $\{f_{1+4n}\}$ and $\{f_{2+4n}\}$ for $n = 0, \dots, 15$.

The shooting function was approximated numerically for $0 < C \le 10^{12}$ by solving the ODE (4.2) subject to the Dirichlet condition $f(L) = C\phi_0(L)$ with L such that $f(L) = 10^{-10}$. The zeros of $S(C)$ then were used as initial guesses for a boundary value solver where now (4.2) was solved on $y \in (0, 150)$ subject to $f'(0) = 0, f(150) = 0, f'(150) = 0$. The right-hand conditions are satisfied by exponentially decaying solutions within error tolerance.

Numerical solution in the inner region. Because of the fast oscillation in (4.32) and the slow convergence to this profile (the next terms are $\mathrm{O}(1/\bar{z})$), the accurate numerical computation of the coefficients (4.43) is not straightforward; in many ways it is the most difficult numerical ODE problem in this paper. In fact, it is known that for second-order equations of the type

$$y''(t) + g(t)y(t) = 0, \quad \text{with } g(t) \to \infty \text{ as } t \to \infty$$

(of which the Airy functions are particular solutions), standard Runge–Kutta or linear multistep methods are not only poorly suited, but are incapable of producing reliable solutions in double precision arithmetic for large t [23]. As such, the values were computed using a modified magnus method employing Filon quadratures [22, 24].

Numerical solution of the PDE. The possibility of finite-time blow-up suggests that adaptive strategies in both time and space are required for reliable numerical solution to (1.1). The geometric features of this problem are also key, both the scale invariance and the preservation of mass. As such we use a scale-invariant moving-mesh

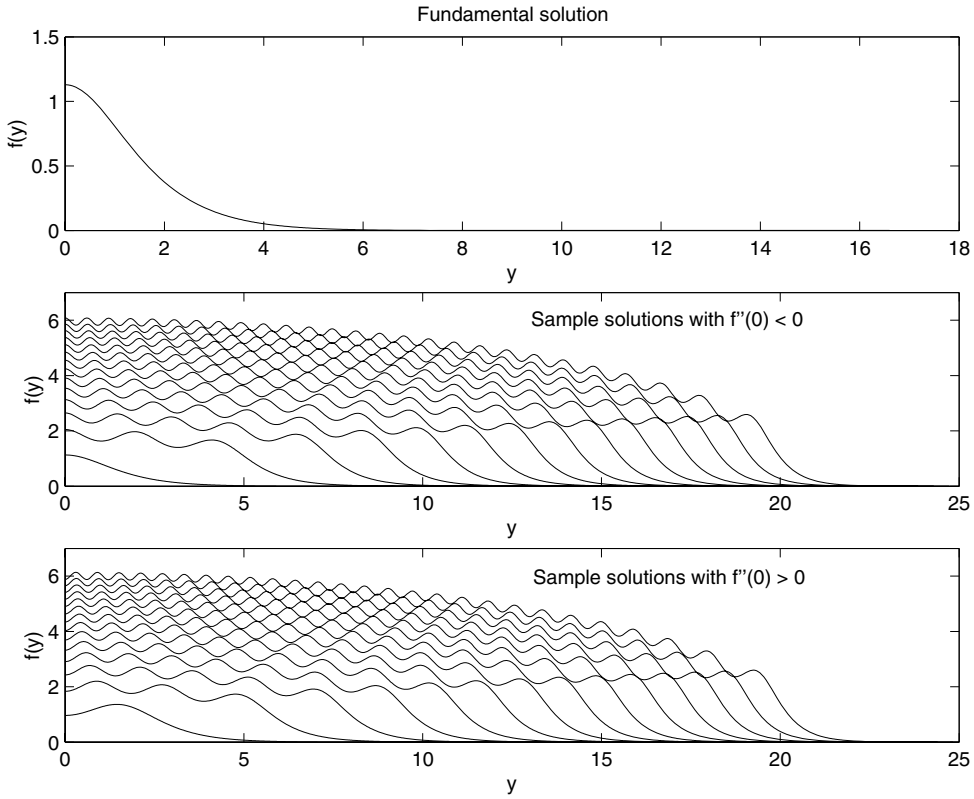


FIG. 5. *Various numerical solutions to* (4.2).

strategy which dynamically clusters the grid points where a solution indicator, in this case $|u|^3$, is large [20, 7, 21, 33, 8]. Additionally, we rescale dynamically in time and compute the physical time as part of the solution procedure [8]. This strategy does not assume any particular solution structure, but instead follows the scaling structure of the full PDE. This allows the computational grid points to move along level sets of any (emerging) similarity variable. We have also used a *conservative collocation* discretization [20, 33] which preserves the mass of the solution even into blow-up while on a moving grid.

This code is fully implicit in time and designed to solve general problems of the form $f(t, x, u, u_x, u_{xx}, u_{xxx}) = \frac{\mathrm{d}}{\mathrm{d}x} g(t, x, u, u_x, u_{xx}, u_{xxx})$ and has been used for examining many higher-order equations [6, 16, 33]. The spatial adaptive strategy has been shown to be reliable on problems for which static regridding codes have failed. For details of the code we refer to [33], while the adaptive strategy is described in [21, 7, 8].

4.5. Exponential asymptotic stability of the first blow-up pattern with profile f_1: Numerical evidence. Theorem 4.4 together with the above matching analysis for $k \gg 1$ imply that there is a discrete family of solutions to the similarity ODE (4.2), and hence a discrete set of admissible masses for the final time profiles

$$M_k = \int f_k(y)\,\mathrm{d}y, \quad k = 1, 2, \ldots . \tag{4.45}$$

We now justify that the first blow-up pattern with profile f_1 is the only stable one.

Returning to the rescaled PDE (2.3), which for $N = 1$ and $p = 3$ takes the form

$$\theta_\tau = \mathbf{A}(\theta) \equiv -\theta_{yyyy} - \frac{1}{4}(\theta y)_y - (\theta^3)_{yy}, \tag{4.46}$$

with initial data θ_0, we study the asymptotic behavior of the global orbit $\{\theta(\cdot,\tau)\}$. First we establish the following natural (but not straightforward) property of its ω-limit set $\omega(\theta_0)$.

PROPOSITION 4.6. *With the definition* (2.1) *of the rescaled solution,*

$$0 \notin \omega(\theta_0). \tag{4.47}$$

Proof. We have from (4.46) that $\mathbf{A}'(0) = \mathbf{B}^* - \frac{1}{4}\,I$ with the known spectral and other properties described in section 3.2. By Lemma 3.1, $\sigma(\mathbf{A}'(0)) = \{-\frac{k+1}{4},\ k \ge 0\}$. The principle of linearized stability [30, Chapt. 9] implies that any sufficiently small solution $\theta(y,\tau)$ decays as $\tau \to \infty$ exponentially fast and that, for some constant $C_0 > 0$,

$$|\theta(y,\tau)| \le C_0 \mathrm{e}^{-\tau/4} \quad \text{uniformly in } \mathbf{R}.$$

Then one can see from scaling (2.1) with $p = 3$ that $|u(x,t)| \le C_0$ for all $t \approx T^-$, i.e., u does not blow-up at $t = T$ contradicting the choice of blow-up time T. □

The linear part of the operator $\mathbf{A}$, $\mathbf{B}^* - \frac{1}{4}\,I$, is not self-adjoint and we do not expect that $\mathbf{A}$ is a potential operator nor that (4.46) is a gradient system. Therefore, the stability properties of the first similarity profile f_1 are studied via the spectrum of the linearized operator

$$\mathbf{A}'(f_1) = \mathbf{B}^* - \frac{1}{4}\,I - 3\frac{\mathrm{d}^2}{\mathrm{d}^2 y}(f_1^2 I), \tag{4.48}$$

which is posed in a functional setting similar to that for $\mathbf{B}^*$ introduced in section 3.

PROPOSITION 4.7. *$\mathbf{A}'(f_1) : H^4_\rho \to L^2_\rho$ is a bounded linear operator with the discrete spectrum $\sigma(\mathbf{A}'(f_1)) = \{\mu_l\}$.*

Proof. We recall that by Theorem 4.4, $f_1(y)$ has exponential decay as $y \to \infty$ so that (4.48) is a lower-order perturbation with smooth, bounded and exponentially decaying coefficients of the operator (3.1) and hence $\mathbf{A}'(f_1)$ is a bounded operator by Lemma 3.2; see [18]. In view of the known spectrum of $\mathbf{B}^*$ having compact resolvent $(\mathbf{B}^* - \lambda I)^{-1}$ in L^2_ρ [10], and by taking $(\mathbf{A}'(f_1) - cI)^{-1}$ with a large constant c, we have that the additional terms in (4.48) form a compact perturbation of the integral operator. Hence, the spectrum of $\mathbf{A}'(f_1)$ is discrete. □

Note that the two first real positive eigenvalues of $\mathbf{A}'(f_1)$ are easily calculated explicitly:

$$\mu_0 = 1,\ \ \phi_0 = (y f_1(y))' \quad \text{and} \quad \mu_1 = \frac{1}{4},\ \ \phi_1 = f_1'(y), \tag{4.49}$$

corresponding to the invariance of the original PDE (1.1) under the group of translations in t and x, respectively. As is usual in blow-up problems, since the blow-up scaling (2.1) does not admit such translations (the time T and the blow-up point $x = 0$ are fixed), these unstable modes are not available for the rescaled PDE (4.46).

A meaningful estimate of the real part of the remainder of the eigenvalues $\{\mu_l,\ l \ge 2\} \subset \mathbb{C}$ of $\mathbf{A}'(f_1)$ is nontrivial. Instead, we present a numerical calculation, from

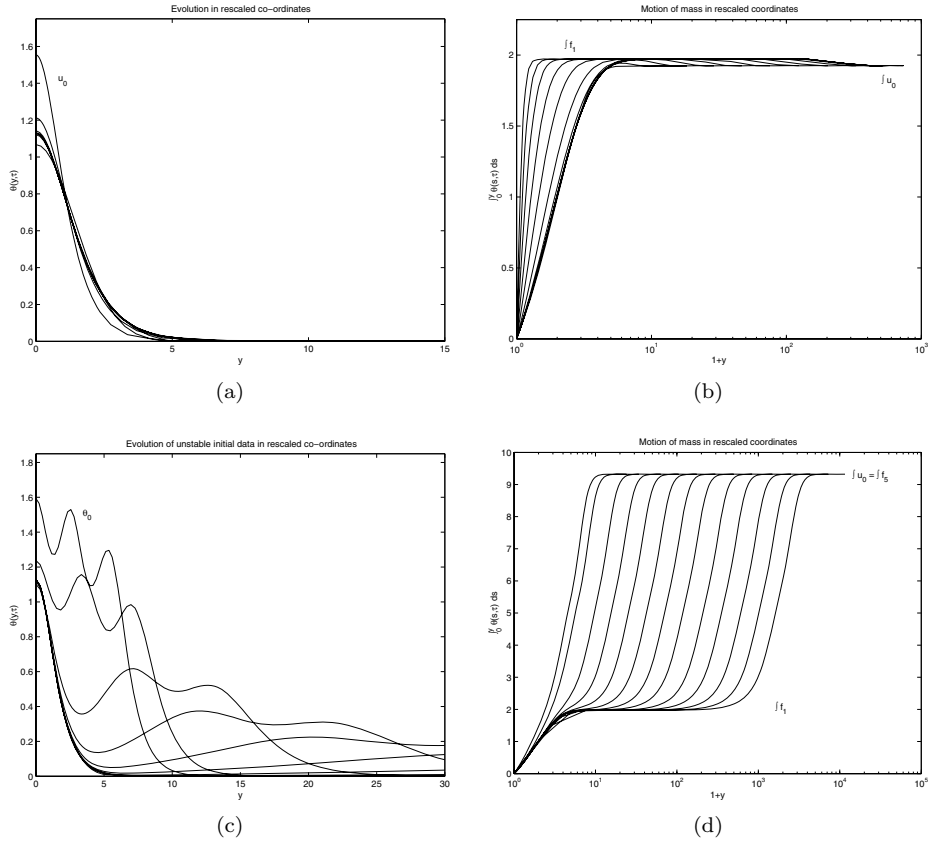


FIG. 6. *Evolution of solutions to* (1.1). (a) *Convergence to* f_1 *in rescaled coordinates.* (b) *Motion of the mass for initial mass* $\int u_0 < \int f_1$. (c) *Instability of* f_5. (d) *Motion of the mass for initial mass* $\int u_0 = \int f_5 > \int f_1$. *The rescaled coordinates are determined through the transformation* (2.1) *with* T *being taken from the numerical approximation.*

directly simulating the PDE, in Figure 6(a), where we rescale the PDE solutions according to the blow-up scaling (4.4) with $\sigma = 1$ and carefully computed blow-up time T. We have chosen symmetric monotone decreasing initial data having a smaller mass than the similarity solution,

$$\int u_0 < M_1 = \int f_1.$$

Figure 6(b) shows the time-evolution of the mass distribution ($\int_0^y \theta(s,\tau)\,\mathrm{d}s$) of this solution concentrated on intervals $[0,y]$ for $\tau \gg 1$, so that the total mass is achieved for $y = \infty$ (for $y \gg 1$ in numerical experiments), establishing convergence to M_1 (our numerical scheme was designed to preserve mass for problems of this form). In order to gain an extra positive mass to converge to f_1 uniformly on compact subsets, the rescaled solution forms two small negative humps, which disappear at $y \to \pm\infty$ as $\tau \to \infty$; see further comments below. By considering this distribution of mass one can see firstly that the numerical scheme preserves the total mass and also how the rescaled solution converges to the first similarity profile on compact sets of the rescaled spatial coordinate y.

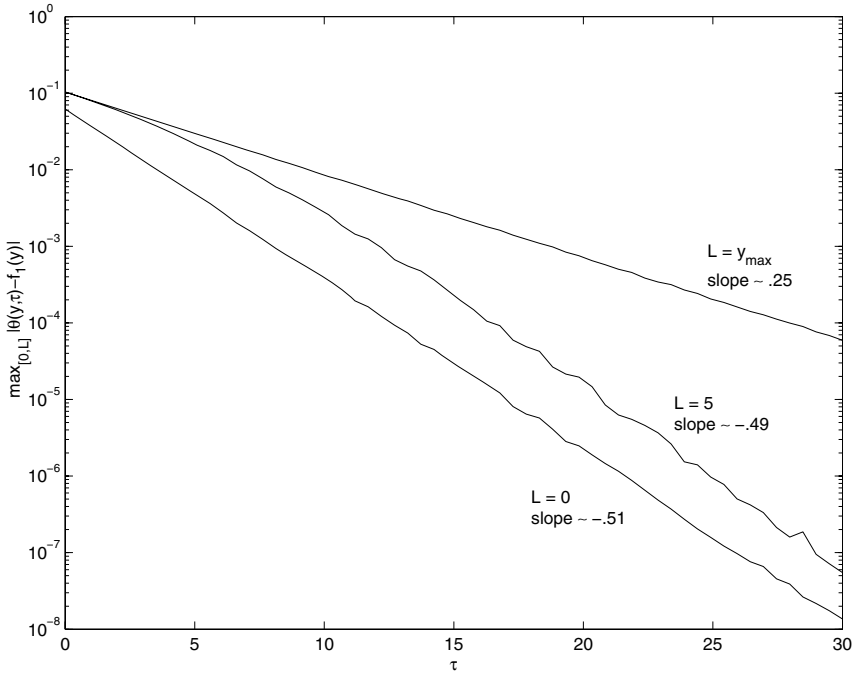


FIG. 7. *Rate of convergence to the similarity profile for $p = 3$, $N = 1$.*

From a variety of numerical experiments, we conjecture that the next eigenvalue of the linearized operator (4.48) satisfies

$$\boxed{\operatorname{Re}\mu_2 = -0.51\ldots.} \tag{4.50}$$

This eigenvalue ensures the exponential stability of the first blow-up profile f_1. We are not aware of a proof of the inequality $\operatorname{Re}\mu_2 < 0$ for operator (4.48), and, as is often happens in the operator and stability theory, we initially rely on rather delicate numerical results. This calculation is based on carefully measuring the rate of the L^∞-convergence in the rescaled coordinates over compact subsets in the similarity variable y. This is of crucial importance, since measurement over all of x shows another (wrong) estimate $\operatorname{Re}\mu_2 \simeq -\frac{1}{4}$ instead. This is understood analogously to Proposition 4.6. Indeed, solutions with mass different from the final time profile cannot, because of conservation, lose that mass, but, similarly, it cannot contribute to the final-time profile in the rescaled coordinates. This extra mass (positive or negative) is damped to zero in the rescaled coordinates as described in Proposition 4.6 and hence with exponential rate $\mathrm{e}^{-\tau/4}$. However, this is not a property of the linearized operator about the first similarity profile and such an incorrect estimate follows from the scaling (2.1), where $\frac{1}{2(p-1)} = \frac{1}{4}$ for $p = 3$. This discrepancy is indicated in Figure 7. This type of estimate can of course be made by numerical approximation of the similarity ODE and numerical evaluation of the spectrum of the numerical linearization about such solutions. However, this gives no information about more general convergence behavior.

Numerical simulation of the full PDE with stationary profiles other than f_1 shows them to be exponentially unstable and the linearized operator

$$\mathbf{A}'(f_k) = \mathbf{B}^* - \frac{1}{4}\, I - 3\frac{\mathrm{d}^2}{\mathrm{d}y^2}(f_k^2 I) \tag{4.51}$$

has eigenvalues with positive real parts. Ordering the sequence of eigenvalues $\{\mu_l\}$ such that their real parts are nonincreasing means that

$$\operatorname{Re}\mu_2 > 0 \quad \text{for all } \ k \geq 2. \tag{4.52}$$

This has been checked numerically in a number of PDE experiments with $k = 2, 3, 4$ and 5. For instance, Figures 6(c) and 6(d) displays the unstable evolution for the initial function $u_0(x)$ given by the rescaled profile $\theta_0 \simeq f_5(y)$ (with the addition of a small positive perturbation) with the larger mass $\int \theta_0 = \int f_5 > \int f_1$. Then to compensate such a positive mass-defect, two positive humps appear and move fast to infinity in the rescaled coordinates showing convergence to f_1 on compact subsets of the similarity variable y. This and similar other numerical experiments with very precise approximation clearly suggest that the similarity profiles $f_2 - f_5$ are exponentially unstable.

5. General *p*'s. The case of general p is more complicated than for the conservative case as without conservation the ODE (2.5) remains truly fourth order.

5.1. Moment conservative case $p = 2$. One seeming exception to this is the second critical exponent $p_1 = 2$, for which one can also derive a third-order ODE. These are the solutions that conserve the first moment. For simplicity, we consider only $N = 1$. Multiplying by x in the PDE (1.1) and integrating by parts, we have

$$0 = -\int_{\mathbf{R}} (u_{xx} + u^p)_{xx}\, \mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbf{R}} ux\, \mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}\left[(T-t)^{-1/2(p-1)+1/2}\int_{\mathbf{R}} \theta(y,t) y\, \mathrm{d}y\right].$$

Thus, $p = 2$ defines the exponent, for which the first moment of solutions is conserved. Generically, in $\mathbf{R}^N$ this leads to the second critical exponent $p_1 = 1 + \frac{2}{N+1}$ in the sequence (1.8) and similar to (2.10), we conclude that

$$\text{for any } \ p \neq p_1, \quad yf \in L^1 \quad \Longrightarrow \quad \int yf = 0. \tag{5.1}$$

Multiplying the ODE (2.5) in $\mathbf{R}$ with y and integrating once, we have

$$f'''y - f'' + \frac{1}{4}y^2 f + (f^2)'y - f^2 = -f''(0) - f^2(0) \equiv \frac{1}{4}A, \tag{5.2}$$

where the constant $A \in \mathbf{R}$ determines the rate of algebraic decay at infinity. Recall that from (2.18) with $p = 2$ the asymptotic algebraic behavior is

$$f(y) = Ay^{-2}(1 + o(1)) \quad \text{as } \ y \to \infty. \tag{5.3}$$

Unlike the critical case $p = 3$ (cf. (4.2)), regardless of the fact that (5.2) is a third-order ODE, we need to keep two symmetry conditions at the origin

$$f'(0) = f'''(0) = 0. \tag{5.4}$$

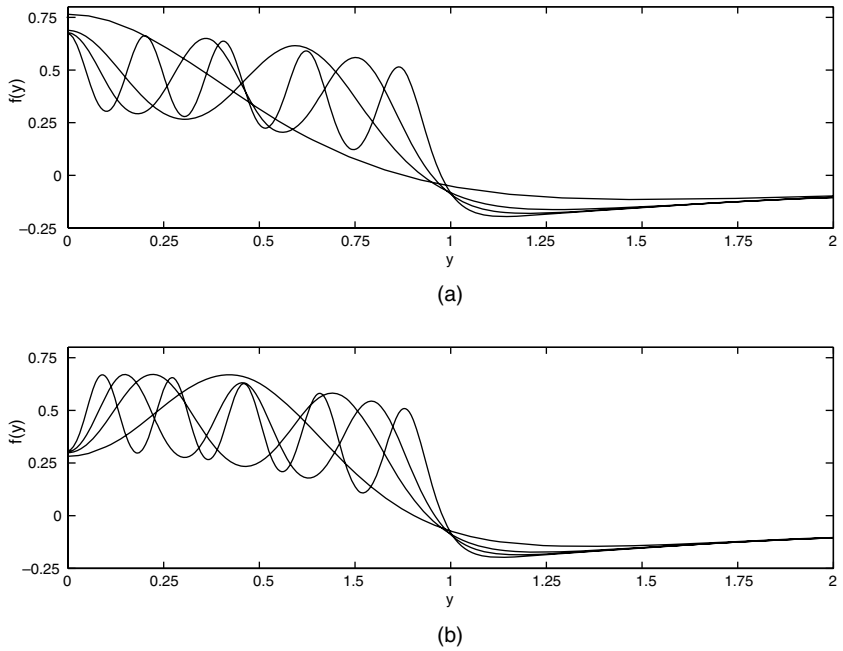


FIG. 8. (a) *Examples from the family of algebraically decaying solutions of* (5.2) *with* $f''(0) < 0$. (b) *Examples from the family of algebraically decaying solutions of* (5.2) *with* $f''(0) > 0$. *(All solutions are shown with* f *rescaled with* $(-A)^{-1/2}$*.)*

For the case $A = 0$, (5.2) may be rewritten as

$$f'' + (f^2)' + \frac{1}{4}\, yF = 0, \quad F = \int_0^y f(s)\, \mathrm{d}s.$$

Then the far-field behavior is purely exponential with a one-dimensional (1D) bundle there given in (2.17). This 1D bundle is not enough to "shoot" two boundary conditions (5.4), and we have got reliable numerical and analytical evidence that there is no exponentially decaying solution for $p = 2$.

For $A \neq 0$, solutions can again be constructed by continuation in A, but now the condition of zero mass must be enforced creating, once again, a fourth-order system. For all $A \neq 0$, we solve the ODE (5.2), (5.3) imposing the condition $\int_{\mathbf{R}} f = 0$ by setting $f(y) = F'(y)$, where F is odd, $F(0) = 0$ and $F(y) \to 0$ as $y \to \infty$, finding A such that $f'(0; A) = 0$. In Figure 8 we present example solutions with algebraic decay. These solutions are presented with f rescaled with $(-A)^{-1/2}$. The main difference from the case $p = 3$ is that now the shooting from $y = \infty$ is two-dimensional including the parameter A in (2.18) and an additional $C \in \mathbf{R}$ from the corresponding exponential bundle (2.17). This 2D asymptotic bundle is sufficient to match with two symmetry conditions (5.4) and to generate a countable set of similarity profiles $\{f_k\}$ shown in Figure 8. This 2D matching problem is more difficult than the 1D one for $p = 3$ and already exhibits typical "multidimensional" features of higher-order ODEs.

The details of a matched asymptotic construction are given in the appendix, which are now more involved than the $p = 3$ case. Figure 8 illustrates that the inner solutions have oscillations whose amplitudes are not asymptotically small (cf. (4.27)). However, we expect that an existence result similar to that in Theorem 4.4 as well

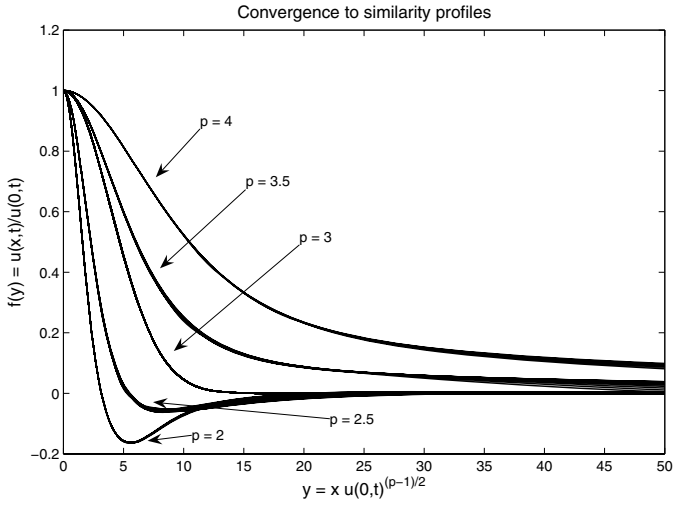


FIG. 9. *Asymptotic blow-up similarity behavior of rescaled solutions of the* $1D$ *PDE* (1.1) *for* $p \in [2, 4]$. *For each value of* p *solutions are plotted with* $u(0, t)$ *varying over 5 orders of magnitude.*

as the multiplicity analysis that follows can still be performed but become essentially more delicate.

5.2. On arbitrary $p > 1$. Finally, by direct simulation of the full PDE (1.1) with various values of $p \neq 3$, we conjecture that stable similarity profiles satisfying the ODE (2.5) govern the blow-up dynamics for all p, not just in the conservative case; see Figure 9. For instance, for $p = 2$ solutions converge exponentially fast in τ on compact subsets of y to the first fundamental *algebraically* decaying solution in Figure 8(a). Recall that, for all $p \neq 3$, precisely two symmetry conditions (5.4) should be taken into account, so we always deal with a 2D shooting by using a 2D algebraic-exponential bundle as $y \to \infty$ defined by a combination of (4.3) and (2.18),

$$f(y) = Ay^{-2/(p-1)} + \cdots + Cy^{-1/3}\mathrm{e}^{-\beta y^{4/3}} + \cdots \quad \text{as } y \to +\infty,$$

where A and C are arbitrary shooting parameters. Then the set of similarity profiles is always discrete and the first stable profile is isolated. Moreover, for such p's, a kind of singular perturbation technique similar to that in section 4 can be applied detecting, as a typical feature, countable sets of blow-up similarity profiles. A detailed similarity analysis for $p \neq 3$ and in the "singular" case $p < 1$ (with finite propagation and solutions of changing sign) will be presented in a forthcoming paper.

6. Global similarity and approximate similarity patterns. We now turn our attention to the better mathematically tractable case of global solutions. Taking $\sigma = -1$ and $T = 0$ in (2.1) yields the rescaling of global-in-time solutions. We study the asymptotic behavior as $\tau \to +\infty$ of solutions satisfying the parabolic PDE

$$\theta_\tau = \mathbf{A}(\theta) \equiv -\Delta(\Delta\theta + \theta^p) + \frac{1}{4}\, y \cdot \nabla\theta + \frac{1}{2(p-1)}\,\theta. \tag{6.1}$$

6.1. Similarity patterns for the one-dimensional equation with $p = 3$. As in the case of blow-up in section 4, we begin with the analysis of global self-similar

solutions

$$u_S(x,t) = t^{-1/4} f(y), \quad y = x/t^{1/4},$$

where f satisfies the ODE obtained from (2.5), (2.3) by integration

$$f''' - \frac{1}{4} yf + (f^3)' = 0, \quad \text{for } y > 0, \quad f'(0) = 0. \tag{6.2}$$

We are looking for profiles f with exponential decay at infinity so that these are L^1-solutions satisfying the conservation law (2.7). Recall that, unlike the blow-up case (4.2), the ODE (6.2) admits a two-dimensional exponential bundle as $y \to \infty$; see (2.16). This essentially simplifies the existence analysis and implies a continuous set of similarity profiles.

THEOREM 6.1. *The ODE problem* (6.2) *has an unbounded continuous family of exponentially decaying solutions.*

Proof. Step 1: monotonicity of solutions as $y \to \infty$. By a standard local analysis, one can check that, in addition to the two-dimensional exponential bundle of solutions mentioned in (2.16) and those with the algebraic decay (2.18), (6.2) has a three-dimensional bundle of asymptotically monotone growing solutions as $y \to \infty$

$$f(y) = \frac{y}{\sqrt{12}} + \frac{1}{y}\left(C_1 \cos\left(\frac{y^{3/2}}{3} \right) + C_2 \sin\left(\frac{y^{3/2}}{3} \right) + C_3 \right) + \cdots, \tag{6.3}$$

where C_1, C_2, C_3 are arbitrary parameters.

Step 2: the shooting argument. For a fixed constant $a \ge 0$, denote by $f(y;C)$ the solution of (6.2) with the conditions

$$f(0) = a, \quad f''(0) = C.$$

As in section 4, one can show that $f(y;C)$ is globally defined. Using the stability of the three-dimensional bundle (6.3) concentrated around the stable explicit solution

$$f_*(y) = \frac{y}{\sqrt{12}} \to \infty, \quad y \to \infty, \tag{6.4}$$

we have that there exists a sufficiently large $C_1 > 0$ such that $f(y;C)$ belongs to the bundle in (6.3) for all $C \ge C_1$. On the other hand, via symmetry by reflection, for all $-C \gg 1$, $f(y)$ approaches as $y \to \infty$ the bundle around the explicit profile $-f_*(y)$. Introducing the set

$$W = \{\mu < C_1 : \ \ f(y;C) \ \text{ belongs to (6.3) for all } C \in (\mu, C_1)\},$$

we have that there exists a finite $C = \inf W$, and by construction, $C = C(a)$ provides us with a profile $f(y;C(a))$ which belongs to the exponential bundle (2.16) as $y \to \infty$. □

Note that this implies that there exists a solution with all $f(0) = a \in \mathbf{R}$. In Figure 10 we present the bifurcation diagram with respect to mass of this continuous family of similarity profiles. As we already know, there exists another case $p = 2$, $N = 1$, where the fourth-order ODE (2.5) reduces to a simpler third-order one. The existence results here are quite similar and the family of solutions is also continuous. However, continuous families for $p = p_0$ and $p = p_1$ (quite special conservative cases associated with (1.1) in divergence form) are exceptional and these are not the generic situation for arbitrary p; see section 6.3, where discrete sets will be detected (cf. the VSSs in [17]).

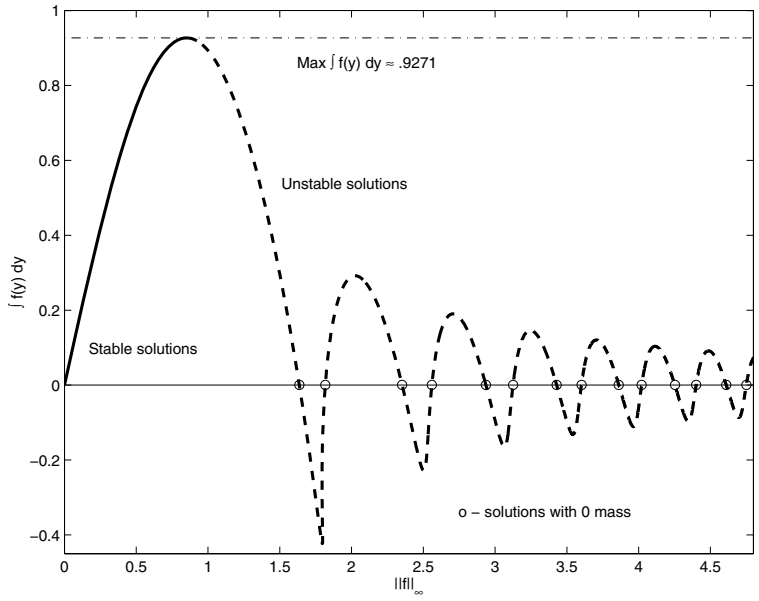


FIG. 10. *Mass bifurcation diagram.*

6.2. The minimal mass-branch is evolutionary stable. It follows from Figure 10 that, for any fixed initial mass $m_0 = \int u_0 \neq 0$, there exists a finite, or empty if

$$\boxed{|\int u_0| > m^* = 0.9271\dots},$$

set of similarity solutions with the given mass. Moreover, for $m_0 = 0$, besides $f(y) \equiv 0$, a countable family of such solutions is expected to exist that is clearly shown in Figure 10 as intersection points with the horizontal axes.

The crucial problem for such a case of multiple solutions is the stability of those solutions in the PDE sense. Let us show that the *minimal branch* in Figure 10 corresponding to the limit

$$f \to 0 \quad \text{as} \quad m_0 \to 0 \tag{6.5}$$

is stable at least for all small masses $m_0 > 0$ (or $m_0 < 0$ replacing $f \mapsto -f$). As a typical example, we perform such computations for the N-dimensional case bearing in mind that the first critical exponent is now $p = p_0 = 1 + \frac{2}{N}$. Consider the corresponding elliptic equation

$$\mathbf{A}(f) \equiv -\Delta^2 f + \frac{1}{4}\, y \cdot \nabla f + \frac{N}{4} f - \Delta f^p = 0, \quad \int f = m_0 > 0. \tag{6.6}$$

First, as an example, we establish a local result on the existence of such similarity patterns in general dimension. For $N = 2$ such a local existence was earlier established in [1] by a different technique.

PROPOSITION 6.2. *Let $p = p_0$. For any sufficiently small $m_0 > 0$,* (6.6) *admits a solution satisfying*

$$f = m_0 F + O(m_0^p) \qquad (\textit{with } F \textit{ is as in } (1.13)). \tag{6.7}$$

Proof. Setting $f = m_0 v$ and introducing operator $\mathbf{C} = \mathbf{B} - I$ with the strictly negative point spectrum, $\sigma(\mathbf{C}) = \{\lambda_l = -1 - \frac{l}{4}\}$, we consider the equivalent integral equation

$$v = \mathbf{D}(v) \equiv -\mathbf{C}^{-1} v + |m_0|^{p-1} \mathbf{C}^{-1} \Delta v^p, \tag{6.8}$$

with condition $\int v = 1$. Since $\mathbf{B}$ has compact resolvent in $L^2_\rho(\mathbf{R}^N)$ [10], we have that after a suitable smooth truncation of the nonlinearity v^p for $|v| \gg 1$ (see further comments below), $\mathbf{D}$ is a compact Hammerstein operator [25]. Since $\mathbf{D}'(0) = \mathbf{C}^{-1}$ has the simple eigenvalue $\lambda_0 = 1$, bifurcation occurs at $m_0 = 0$ and by the Lyapunov–Schmidt method [25], the solution can be represented in the form $v = F + w$, where $w \in \mathcal{L}^\perp\{F\}$. On substitution into (6.8) we obtain that $w = O(|m_0|^{p-1}) \in L^2_\rho$ is small for $|m_0| \ll 1$. Hence, w is small in H^{2m}_ρ and by the known asymptotic properties of the ODE under consideration, w is small uniformly. This means that the truncation of the unbounded nonlinearity v^p does not affect the solution bifurcating at $m_0 = 0$. Finally, we obtain the representation (6.7) in H^{2m}_ρ and uniformly in $\mathbf{R}^N$. See more details in [17, sect. 6]. □

To study the stability of the minimal branch, we consider the linearized operator

$$\mathbf{A}'(f)Y = \mathbf{B}Y - p\Delta(|f|^{p-1}Y) \equiv \mathbf{B}Y - |m_0|^{2/N} p \Delta(|F|^{p-1}Y) + \cdots . \tag{6.9}$$

For small m_0, $\mathbf{A}'(f)$ is a perturbation of $\mathbf{B}$ with the known spectrum as given in (3.3). In 1D and in the radial geometry all the eigenvalues are simple and we will denote the eigenfunctions by ψ_l instead of ψ_β with $l = |\beta|$. Similarly to Proposition 4.7, we have that $\mathbf{A}'(f)$ has a discrete spectrum which is a perturbation of that of $\mathbf{B}$ [18]. Recall that, from the orthogonality condition (3.7), $\psi_0 = F$ is the only eigenfunction of $\mathbf{B}$ with nonzero mass; see (1.13). Thus, requiring that the mass of profiles be preserved, we have to take into account perturbations of eigenvalues corresponding to perturbed eigenfunctions from $\mathcal{L}^\perp\{F\}$, i.e., of eigenvalues of $\mathbf{B}$ for $l = 1, 2, \dots$ ($l = 0$ not included). By direct calculation from (6.9), one can estimate this point spectrum given, to leading order, by

$$\sigma(\mathbf{A}'(f)) = \left\{ -\frac{l}{4} + |m_0|^{2/N} p \left\langle \Delta(|F|^{p-1}\psi_l, \psi_l^* \right\rangle + \cdots , \ l = 1, 2, \dots \right\}.$$

This guarantees that, for all small $|m_0|$, the real parts of the eigenvalues are bounded away from zero from above and that this branch is exponentially stable. Note that this expansion is not valid near the subsequent zeros of mass (other than 0) seen in Figure 10, as the L^∞-norm of such solutions is not zero for $m_0 = 0$ and hence expansion (6.7) does not hold. Therefore, other continuous monotone increasing or decreasing sub-branches are not stable.

Hence, for almost all initial data with $|\int u_0| \le m^*$, there exists the stable branch of similarity solutions that can attract the solution as $t \to \infty$. If $|\int u_0| > m^*$, then, as is seen from Figure 10, self-similar asymptotic behavior is not available and the solution is expected to blow-up in finite time.

6.3. The *p*-bifurcation diagram and asymptotic behavior. To describe the global bifurcation diagram of similarity profiles, we first determine the spectrum of critical exponents as bifurcation points. Writing the elliptic equation (2.5) in the form

$$\mathbf{B}f + c_* f = \Delta f^p \quad \text{in } \mathbf{R}^N, \quad f(y) \to 0 \text{ exponentially fast as } y \to \infty, \tag{6.10}$$

with $c_* = N(p_0 - p)/4(p-1)$, we obtain the following result.

PROPOSITION 6.3. *Let, for an integer $l \geq 0$, the eigenvalue $\lambda_l = -\frac{l}{4}$ of the operator* (3.1) *be of odd multiplicity. Then the critical exponent p_l given in* (1.8) *is a bifurcation point for the problem* (6.10).

Proof. After performing, as above, a smooth truncation of the nonlinearity f^p, the proof follows analogously to that of Proposition 6.2 (cf. Proposition 6.1 in [17]). By using the explicit representation of the resolvent of $\mathbf{B}$ [10], this differential equation reduces to an integral one with compact Hammerstein operators to which the classical bifurcation techniques apply [25]. □

Using standard bifurcation theory, we briefly describe the results calculated directly from the differential equation. By Lemma 3.1, the linear operator in (6.10) has the spectrum $\sigma(\mathbf{B}+c_*I) = \{c_* - \frac{l}{4},\, l \geq 0\}$. Therefore, any $p = p_l$ for which $c_* - \frac{l}{4} = 0$ gives the critical exponents (1.8), determines a bifurcation point for problem (6.10) provided that $\lambda_l = -\frac{l}{4}$ is of odd multiplicity (e.g., this is always true for $N = 1$ or in the radial geometry where the eigenvalues are always simple). In order to describe the local behavior of the bifurcation branches at $p \approx p_l$, we fix an l and set

$$p = p_l + \varepsilon, \quad \text{with } p_l = 1 + \frac{2}{N+l}.$$

Since $c_* = \frac{l}{4} - \mu_l\varepsilon + O(\varepsilon^2)$, with $\mu_l = \frac{1}{8}(N+l)^2$, (6.10) takes the form

$$\left(\mathbf{B} + \frac{l}{4}I\right) f - \mu_l\varepsilon f + O(\varepsilon^2) = \Delta f^p. \tag{6.11}$$

Using the Lyapunov–Schmidt method ([25, Chap. 8]) by setting $f = C\psi_l + w$, where $w \in \mathcal{L}^{\perp}\{\psi_l\}$, one can find the algebraic equation for C. Solving this, one finds that near the bifurcation point the solution takes the form

$$f = [\nu_l(p-p_l)]^{1/(p-1)}\psi_l + \cdots \quad \text{as } p \to p_l^+, \text{ where } \nu_l = -(N+l)^2/8\kappa_l \text{ and} \tag{6.12}$$

$$\kappa_l = \langle \Delta\psi_l^{p_l}, \psi_l^* \rangle = \langle \psi_l^{p_l}, \Delta\psi_l^* \rangle < 0. \tag{6.13}$$

The expansion (6.12) is posed under the crucial assumption that (6.13) holds. The strict inequality (6.13) has been checked numerically for various $l \geq 2$. It is worth mentioning that an analytical proof of (6.13) is not straightforward even in the simplest case $l = 2$, $N = 1$, where $\psi_2 = F''(y)/\sqrt{2}$, $(\psi_2^*)'' \equiv \sqrt{2}$ and $p_2 = \frac{5}{3}$, though the positivity dominance of the rescaled kernel F in the sense that $\int F = 1$ directly suggests that $\kappa_2 < 0$.

Notice however, that

$$\kappa_0 = \kappa_1 = 0 \quad \text{for all } \ N \geq 1,$$

as $\psi_0^* = 1$ and $\psi_1^*(y)$ is linear in y. Thus, the corresponding branches leave these bifurcation points vertically, suggesting a continuous family of solutions for these critical exponents described above. For $p = 3$, this has been proved in Proposition 6.2. For $p = 2$, this follows from the ODE integrated once after multiplying by y (cf. (5.2) with $A = 0$),

$$f'''y - f'' - \frac{1}{4}y^2 f + (f^2)'y - f^2 = 0. \tag{6.14}$$

Therefore, we need a single condition of odd symmetry at the origin $f(0) = 0$ since the second one $f''(0) = 0$ follows from the ODE (6.14).

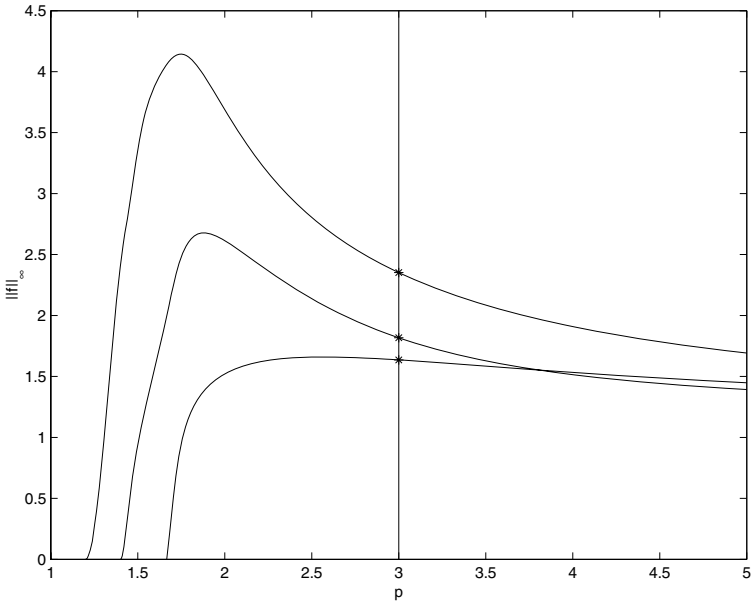


FIG. 11. *Bifurcation diagram of* (6.10) *for* $N = 1$ *with respect to* p *with even symmetry conditions at* $y = 0$. *There is a continuous branch of solutions from* $p = 3$ *and* 2 *but only a discrete set of solutions for all* $p \neq 3, 2$. *Another vertical branch of solutions bifurcates from* $p = 2$ *but with odd symmetry about* $y = 0$.

In Figure 11 we see that the remaining branches bifurcate with increasing p which is guaranteed by (6.13). Then, bearing in mind the countability of bifurcation points (1.8), Figure 11 clearly suggests that for any $p > 1$, $p \neq p_0$, p_1, there exists a countable set of solutions to (6.10). For noncoercive smooth potential operators, this is a typical result from Lusternik–Schnirel'man theory (see the classical formulation in [25, p. 381] and [26, 35] on variational methods in weighted Sobolev spaces like L^2_ρ for problems in $\mathbf{R}^N$), though problem (6.10) is not variational. The countability of solutions is not necessarily associated with the existence of a variational formulation.

It follows from (2.4) that, for $p \neq p_0$, any solution of (6.10) must have zero mass, see (2.10). The similarity profiles on all the bifurcation branches in Figure 11 satisfy (2.10). Lastly, recall from Figure 10 that there is a discrete set of μ_k such that $f(0) = \mu_k$ and $\int f_{\mu_k} = 0$. Clearly, in Figure 11 this countable set of points corresponds to intersections of the bifurcation branches with the vertical one $\{p = 3\}$. To investigate problem (6.10) numerically, we have solved the fourth-order system

$$f''' = \frac{1}{4} fy + c_* F - |f|^{p-1} f', \quad F' = f,$$

coupled with the boundary conditions

$$f'(0) = 0, \quad F(0) = 0, \quad \lim_{y \to \infty} f(y) = 0, \quad \lim_{y \to \infty} F(y) = 0,$$

as a continuation problem in $p < 3$ and $p > 3$ starting from those already known profiles at the intersection points with $\{p = 3\}$.

Figure 11 shows branching of solutions of (6.10), i.e., similarity profiles with exponential decay at infinity. Concerning possible solutions with the algebraic decay

(2.18), one can see that such functions satisfy $f \in L^1(\mathbf{R}^N)$ for $p < p_0$ so that the blow-up rate (2.4) implies the zero mass condition (2.10). We expect that there exists a continuous family of such solutions with certain domain of attraction which do not play any role for exponentially decaying initial data.

On stability of the p_2-branch for $p \in (p_2, p_0)$. Let us show that the first nonvertical branch, from $p = p_2$, is stable for $p \approx p_2^+$. This is done by estimating the real parts of the eigenvalues of the linearized operator given in (6.10), which for $l = 2$ has the form

$$\mathbf{A}'(f) = \mathbf{B} + \frac{1}{2}\, I - \varepsilon \mathbf{C} + o(\varepsilon), \quad \mathbf{C} = \mu_2 I + p_2 \nu_2 \Delta(|\psi_2|^{p_2 - 1} I), \tag{6.15}$$

where we substitute the expansion of f given by (6.12) with $l = 2$. By $\{\bar\lambda_l = \bar\lambda_l(\varepsilon)\}$, $\varepsilon = p - p_2 > 0$, we denote the discrete spectrum of $\mathbf{A}'(f)$.

For $N = 1$, due to the conservation of mass (for any $p \le p_0 = 3$) and of the first moment (for any $p \le p_1 = 2$), the first modes with positive unperturbed eigenvalues of $\mathbf{B} + \frac{1}{2} I$ ($\varepsilon = 0$), $\bar\lambda_0(0) = \frac{1}{2}$ and $\bar\lambda_1(0) = \frac{1}{4}$ are not taken into account and we need to check the perturbed third eigenvalue $\bar\lambda_2(\varepsilon)$. The same holds for $N > 1$ in the radial setting deleting all eigenvalues corresponding to nonsymmetric eigenfunctions and hence $\bar\lambda_1$. Since this unperturbed eigenvalue vanishes, $\bar\lambda_2(0) = 0$, by perturbation theory [18], the eigenvalue expansion takes the form $\bar\lambda_2(\varepsilon) = \varepsilon\rho + o(\varepsilon)$ with the eigenfunction given by $\bar\psi_2(\varepsilon) = \psi_2 + \varepsilon\varphi + o(\varepsilon)$. Substituting these approximations into the eigenvalue equation, we find

$$\mathbf{A}'(f)\bar\psi_2(\varepsilon) = \bar\lambda_2(\varepsilon)\bar\psi_2(\varepsilon),$$

and using (6.15), we obtain from the equation at $O(\varepsilon)$ that the expansion coefficients ρ and φ satisfy

$$\left(\mathbf{B} + \frac{1}{2}\, I\right)\varphi = \rho\psi_2 + \mathbf{C}\psi_2.$$

Hence, by the orthogonality condition,

$$\rho = -\langle \mathbf{C}\psi_2, \psi_2^* \rangle = -\mu_2 - \frac{1}{8}\, p_2(N+2) = -\frac{1}{8}\,(N+2)(N+3),$$

and, for small $\varepsilon = p - p_2 > 0$,

$$\bar\lambda_2(\varepsilon) = -\frac{1}{4}\,(p - p_2)(N+2)(N+3) + o(p_2 - p), \quad \text{i.e., } \operatorname{Re}\lambda_2(\varepsilon) < 0. \tag{6.16}$$

This means the exponential asymptotic stability of the similarity patterns on the p_2-bifurcation branch for all $p - p_2 > 0$ sufficiently small. We expect that the whole branch remains stable for all $p \in (p_2, p_0)$. Note that for $p = p_0$ it is not stable as Figure 10 suggests since this solution has $\|f\|_\infty \approx 1.6$ and the stable solutions have smaller L^∞-norm.

Stability of the fundamental solution for $p > p_0$. It follows from (6.10) that $c_* < 0$ for $p > p_0$ so that 0 becomes stable in the rescaled PDE (6.1). Therefore, we apply to (1.1) the scaling corresponding to the fundamental solution (1.12),

$$u(x,t) = (1+t)^{-N/4} g(y,\tau), \quad y = x/(1+t)^{1/4}, \quad \tau = \ln(1+t) \ge 0, \tag{6.17}$$

to get an exponentially perturbed equation

$$g_\tau = \mathbf{B}g - \mathrm{e}^{\gamma\tau}\Delta g^p, \tag{6.18}$$

where $\gamma = \frac{1}{4}N(p_0 - p) < 0$, if $p > p_0$. Recall that by Lemma 3.1, $\mathbf{B}$ has a one-dimensional center subspace with the eigenfunction F; see (1.13). Therefore, we arrive at a typical result in the supercritical range $p > p_0$: for all sufficiently small initial data with $m_0 = \int u_0 \neq 0$, the solution of (1.1) is global and has the same asymptotic behavior as the fundamental solution (1.12) with identical mass,

$$u(x,t) = m_0 t^{-N/4}\left[F\left(x/t^{1/4}\right) + o(1)\right] \quad \text{as } t \to \infty.$$

See rigorous analysis in [10] for (1.7) in the range $p > p_0 = 1 + \frac{2m}{N}$, which can be applied to the PDE (6.18).

6.4. Transitional non-self-similar patterns at critical exponents for the limit stable Cahn–Hilliard equation. Our analysis shows that at the bifurcation points (1.8) there occurs a transition from the linearized patterns corresponding to the stable manifold of $\mathbf{A}$ with the linearization (3.1) to the essentially nonlinear similarity patterns described in Figure 11. This is quite a general phenomenon in nonlinear parabolic equations observed in various higher-order models, [6, 10, 16, 17].

It remains to determine what happens at the critical values $p = p_l$ for the original PDE (1.1). Here one can expect some nontrivial center manifold patterns in the rescaled (6.1). It turns out that this center manifold behavior is unstable (possibly, with blow-up) for (1.1) but leads to nonsimilarity patterns for the stable (1.9). The construction of global self-similar solutions (2.4), $\sigma = -1$, is quite the same as that already considered in section 6.3 but instead of (6.10) we arrive at the equation

$$\mathbf{B}f + c_* f = -\Delta f^p. \tag{6.19}$$

Therefore, in the bifurcation analysis similar to Proposition 6.3 we obtain expansions (6.12), where $\nu_l = (N+l)^2/8\kappa_l < 0$, i.e., the branches bifurcate at $p = p_l^-$ concentrating at $p = 1^+$, and hence p is expected to decrease along them (no rigorous proof of nonlocal properties of branches is still available). A similar p-diagram occurs for very singular similarity solutions for the semilinear parabolic equations with absorption [17].

It is important to understand how such bifurcations can affect the evolution orbits of the original PDE (1.9). Consider the corresponding rescaled (6.1), which at $p = p_l$ takes the form

$$\theta_\tau = \left(\mathbf{B} + \frac{l}{4}I\right)\theta + \Delta\theta^p \quad \text{for } \tau \gg 1. \tag{6.20}$$

We fix $l \geq 1$ so that, by orthogonality, ψ_l has zero mass. By (3.3), $\mathbf{B} + \frac{l}{4}I$ has a nontrivial kernel, and therefore, by the general invariant manifold theory (see, e.g., [30]), it is natural to verify the center manifold behavior. This is expressed by looking for solutions in the form

$$\theta(\tau) = a_l(\tau)\psi_l + w(\tau), \quad \text{with } w(\tau) \in \mathcal{L}^\perp\{\psi_l\}, \tag{6.21}$$

where the center manifold dominance in the asymptotic behavior means that $w(\tau) = o(a_l(\tau))$ as $\tau \to \infty$. Substituting this expansion into (6.20) and multiplying by the adjoint eigenfunction ψ_l^* in $L^2(\mathbf{R}^N)$ yields $\dot{a}_l = \kappa_l a_l^p + o(a_l^p)$, where κ_l is as given in (6.13). Since $\kappa_l < 0$, integrating this asymptotic ODE gives

$$a_l(\tau) = [(p-1)|\kappa_l|\tau]^{-1/(p-1)}(1 + o(1)). \tag{6.22}$$

In the original $\{x,t,u\}$-variables, this defines the following asymptotic patterns for the PDE (1.1) at $p=p_l$, $l\geq 2$ as $t\to\infty$:

$$u(x,t)=C_l(t\ln t)^{-(N+l)/2}\left[\psi_l\left(y/t^{1/4}\right)+o(1)\right],\quad \text{where } C_l=[2|\kappa_l|/(N+l)]^{-(N+l)/2}.$$

Obviously, using (3.1) as the linearization in the rescaled (6.1), one can see that, in the supercritical range $p>p_l$, there exist stable manifold patterns with the behavior

$$\theta(y,\tau)=C\mathrm{e}^{-(N+l)(p-p_l)\tau/4(p-1)}[\psi_l(y)+o(1)]\quad \text{as } \tau\to\infty\quad (C\neq 0) \tag{6.23}$$

(cf. (6.21), (6.22) for $p=p_l$). The patterns (6.21), (6.22) are transitional ones at $p=p_l$ describing the exchange between the linearized stable manifold ones (6.23) and the nonlinear similarity patterns studied above. This transition phenomenon is similar to that detected in [17] for the very singular similarity solutions of the nonconservative higher-order parabolic equations $u_t=-(-\Delta)^m u-u^p$.

These transition phenomena are very difficult to detect numerically as it is only a slow logarithmic correction of an asymptotic algebraic decay rate. Moreover, it occurs only at a discrete set of nonlinearities corresponding to critical exponents $\{p_l\}$. Regardless, understanding such phenomena is crucial for the description of evolutionary completeness of patterns for this problem.

7. Summary and final conclusions. This paper has considered both blow-up and global solutions to the fourth-order semilinear parabolic equation (1.1). We have termed such an equation the limit unstable Cahn–Hilliard equation since it can arise as a limiting case of the more familiar and standard Cahn–Hilliard equation with a double-well potential function (1.3). Our attention has focused on describing the eventual patterns of similarity solutions to this equation. In the case of *blow-up*, these occur as the finite blow-up time is approached, while for the *global* solutions these occur as a large-time behavior. As such, in section 2 we presented the similarity scalings and derived the governing ODE equations for both the blow-up and global cases together with the critical value of the exponent $p=p_0=1+\frac{2}{N}$ if mass is to be conserved. Since we are interested in solutions with noncompact support (notice that a free-boundary formulation for (1.1) can give finite interfaces but we do not consider such problems concentrating on the canonical Cauchy problem in $\mathbf{R}^N\times\mathbf{R}_+$), we described in section 2.2 the possible far-field behaviors. In section 3, further preliminary results are given concerning the spectral theory of the linearized operator, which is used in the stability analysis of blow-up profiles in section 4 and plays a key role in bifurcation analysis of global similarity patterns considered in section 6.

In section 4 we provided a detailed analysis of the possible blow-up similarity profiles for the conservative case $p=p_0=1+\frac{2}{N}$ in one space dimension $N=1$, i.e., $p=p_0=3$. We conveniently parameterized the possible profiles using the free constant C that appears in the far-field behavior (4.3), and in section 4.2 gave existence results that C in fact takes an ordered countable set of values $\{C_k\}$ with $C_k>C_{k-1}$. A striking feature of the corresponding similarity profiles $\{f_k\}$ is that they posses a number of local maxima that can be related to the "index" of f_k, which increases as k increases (this can be formally associated with the Morse index or the Sturm's number of zeros of a related "eigenfunction"). In section 4.3, the method of matched asymptotic expansions was used to describe the structure of this family of profiles in the limit when C_k becomes large. A novel two-scale outer expansion was used to the capture the oscillatory behavior associated with the local maxima of the profiles satisfying a singular perturbed ODE. Importantly the asymptotics derived an explicit

expression for the values of C_k with k large. When appropriate we compared the asymptotic solution with the corresponding full numerical solution of the ODE problem to confirm its accuracy, and also in section 4.5 was described a numerical scheme for solution of the original PDE (1.1) in order to confirm convergence to our stable similarity profiles from arbitrary generic initial data. As such in section 4.5, numerical evidence was provided for the exponential stability of the first blow-up pattern f_1.

In section 5 we briefly considered blow-up in cases $p \neq p_0$. We noted the second critical exponent $p = p_1 = 1 + \frac{2}{N+1}$, in which the first moment of solutions is conserved. In this case analytical and numerical progress can be made and we again noted that, similar to the critical case $p = p_0$, we can classify the family of blow-up similarity solutions as a countable set. Matched asymptotic expansions were again used to determine the solution structure for the limiting members of this set, which is relegated to the appendix.

In section 6 we turned our attention to *global* (source-type) similarity solutions. Unlike the blow-up patterns, our main conclusion is that (1.1) admits a continuous set of similarity patterns, which can be parameterized by the mass. We proved existence and exponential stability of solutions on the minimal branch by using the classical bifurcation theory for non-self-adjoint operators, and derive a countable sequence of bifurcation exponents $\{p_l\}$, which lead to a finite number of similarity profiles for noncritical p's.

Appendix. A countable set of blow-up patterns for $p = 2$, $N = 1$. In this appendix we carry out formal asymptotic computations for the ODE (5.2), which are similar to those presented in section 4. While we again have an integrable fourth-order ODE, the calculation is now more involved. The limit $A \to -\infty$ may be considered by recasting (5.2) with (5.3) as a singular perturbation problem in a manner similar to that developed in section 4.3 for $p = 3$. Introducing the small parameter $\varepsilon > 0$ by

$$A = -\varepsilon^{-1/2},$$

the rescaling $y = \varepsilon^{-1/4}\bar{y}$ gives the following problem for $f(\bar{y})$:

$$\varepsilon \left(\frac{f''}{\bar{y}} \right)' + \varepsilon^{1/2} \left(\frac{f^2}{\bar{y}} \right)' + \frac{1}{4} f = -\frac{1}{4\bar{y}^2}, \quad \text{with} \tag{A.1}$$

$$f'(0) = 0, \qquad f''(0) + \frac{f^2(0)}{\varepsilon^{1/2}} = \frac{1}{4\varepsilon}, \quad \text{and} \tag{A.2}$$

$$f = -\frac{1}{\bar{y}^2}(1 + o(1)) \quad \text{as} \quad \bar{y} \to +\infty. \tag{A.3}$$

Here $'$ denotes $\frac{\mathrm{d}}{\mathrm{d}\bar{y}}$. In the limit $\varepsilon \to 0^+$, we obtain a four-layer structure illustrated schematically in Figure 12. We have an outer region $\bar{y} = O(1)$, in which f takes the far-field behavior (A.3), and an inner region $\bar{y} = O(\varepsilon^{1/8})$, in which $f = O(\varepsilon^{-1/4})$. This inner region is partitioned into two distinct regions, an inner 1, where $\bar{y} > \varepsilon^{1/8}\hat{a}$, and an inner 2, where $\bar{y} < \varepsilon^{1/8}\hat{a}$. The partition is through an inner inner region located at $\bar{y} = \varepsilon^{1/8}\hat{a}$, which is of width $O(\varepsilon^{3/8})$. The constant $\hat{a}$ will be shown below to be close to the location of the first singularity of a Painlevé transcendent-type equation.

The outer region $\bar{y} = O(1)$. At leading order (A.1) gives

$$f = -\frac{1}{\bar{y}^2}, \tag{A.4}$$

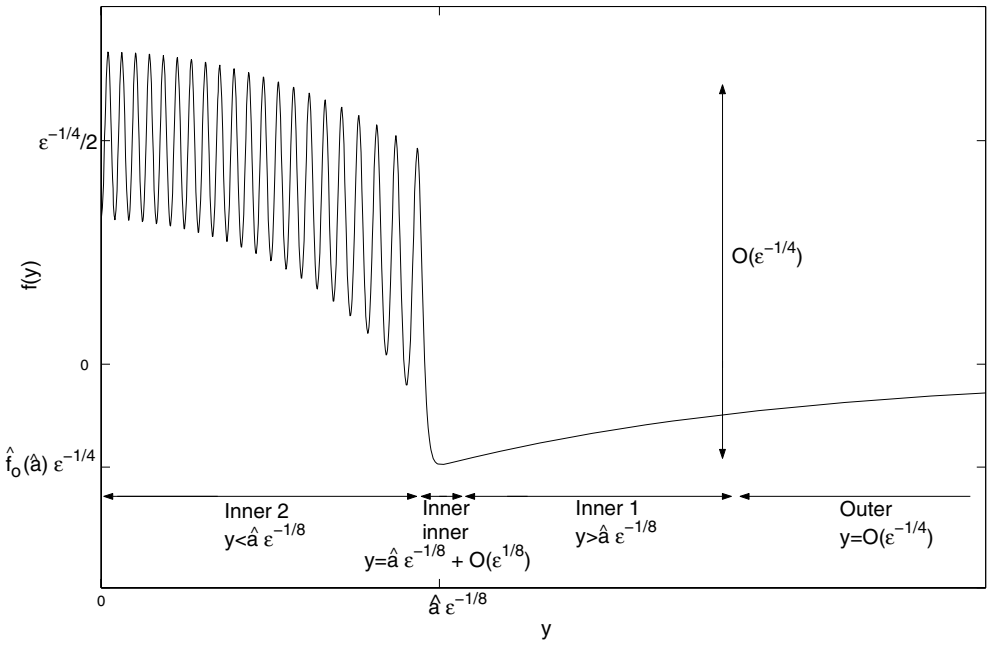


FIG. 12. *Schematic illustration of the asymptotic regions for the singular perturbation problem* (A.1)–(A.3) *expressed on the* y *variable scale.*

which obviously satisfies (A.3) but not (A.2). This gives rise to the consideration of the following inner region, in which terms involving the first derivatives of f are retrieved.

The inner 1 region $\bar{y} > \varepsilon^{1/8}\hat{a}$. Introducing the scalings

$$\bar{y} = \varepsilon^{1/8}\hat{y}, \qquad f = \varepsilon^{-1/4}\hat{f}, \quad \text{so (A.1) becomes} \tag{A.5}$$

$$\varepsilon^{1/2}\left(\frac{\hat{f}''}{\hat{y}}\right)' + \left(\frac{\hat{f}^2}{\hat{y}}\right)' + \frac{1}{4}\hat{f} = -\frac{1}{4\hat{y}^2}, \tag{A.6}$$

where $'$ now denotes $\frac{\mathrm{d}}{\mathrm{d}\hat{y}}$. Posing $\hat{f} \sim \hat{f}_0$ gives the leading order problem

$$\left(\frac{\hat{f}_0^2}{\hat{y}}\right)' + \frac{1}{4}\hat{f}_0 = -\frac{1}{4\hat{y}^2}, \quad \text{with the matching condition} \tag{A.7}$$

$$\hat{f}_0 \sim -\frac{1}{\hat{y}^2} \quad \text{as} \quad \hat{y} \to +\infty. \tag{A.8}$$

We next consider a region located at $\hat{y} = \hat{a}$ so that

$$\hat{f}_0 \sim \hat{f}_0(\hat{a}) \quad \text{as} \quad \hat{y} \to \hat{a}^+. \tag{A.9}$$

The inner inner region $\bar{y} = \varepsilon^{1/8}\hat{a} + O(\varepsilon^{3/8})$. Introducing the new variables

$$\hat{y} = \hat{a} + \varepsilon^{1/4}\hat{Y}, \qquad \hat{f} = \hat{F}, \tag{A.10}$$

we have from (A.6) that

$$\left(\frac{\hat{F}''}{(\hat{a}+\varepsilon^{1/4}\hat{Y})}\right)' + \left(\frac{\hat{F}^2}{(\hat{a}+\varepsilon^{1/4}\hat{Y})}\right)' + \frac{\varepsilon^{1/4}}{4}\left(\hat{F} + \frac{1}{(\hat{a}+\varepsilon^{1/4}\hat{Y})^2}\right) = 0, \tag{A.11}$$

where $'$ is now $\frac{\mathrm{d}}{\mathrm{d}\hat{Y}}$. Posing $\hat{F} \sim \hat{F}_0$ gives, at leading order,

$$\hat{F}_0'' + \hat{F}_0^2 = \left(\hat{f}_0(\hat{a})\right)^2, \tag{A.12}$$

after matching to (A.9). This gives the constant solution $\hat{F}_0 = \hat{f}_0(\hat{a})$ at leading order within this region.

The inner 2 region $\bar{y} < \varepsilon^{1/8}\hat{a}$. We now consider the scalings,

$$\bar{y} = \varepsilon^{1/8}\hat{y}, \qquad f = \varepsilon^{-1/4}F, \tag{A.13}$$

where $F = F(\hat{y}, Y)$ is also taken to be a function of the fast scale $\hat{y} = \varepsilon^{1/4}Y$. Defining the multiple-scales operator $\mathcal{L}$ by $\mathcal{L} = \frac{\partial}{\partial Y} + \varepsilon^{1/4}\frac{\partial}{\partial \hat{y}}$, (A.1) becomes

$$\mathcal{L}\left(\frac{1}{\hat{y}}\mathcal{L}^2 F\right) + \mathcal{L}\left(\frac{F^2}{\hat{y}}\right) + \frac{1}{4}\left(F + \frac{1}{\hat{y}^2}\right) = 0. \tag{A.14}$$

Posing $F \sim F_0 + \varepsilon^{1/4}F_1$, we have, at leading order,

$$\frac{\partial^2 F_0}{\partial Y^2} + F_0^2 = A_0(\hat{y}), \tag{A.15}$$

for some function $A_0(\hat{y})$ and at the next order,

$$\frac{\partial}{\partial Y}\left(\frac{\partial^2 F_1}{\partial Y^2} + 2F_0F_1\right) - 2\frac{\partial F_0^2}{\partial \hat{y}} + \frac{\hat{y}}{4}F_0 + 3\hat{y}^{1/3}\frac{\partial}{\partial \hat{y}}\left(\frac{A_0}{\hat{y}^{1/3}}\right) + \frac{1}{4\hat{y}} = 0. \tag{A.16}$$

The secular terms involving Y in (A.16) are removed if $3\hat{y}^{1/3}\frac{d}{d\hat{y}}\left(\frac{A_0}{\hat{y}^{1/3}}\right) = -\frac{1}{4\hat{y}}$, which gives $A_0(\hat{y}) = \frac{1}{4} - B_0\hat{y}^{1/3}$ for some constant B_0 (assumed positive). Thus F_0 satisfies

$$F_0'' + F_0^2 = \frac{1}{4} - B_0\varepsilon^{1/12}Y^{1/3}, \tag{A.17}$$

where $' = \frac{\mathrm{d}}{\mathrm{d}Y}$ and is subject to (A.2), which at leading order in the current variables becomes

$$F_0'(0) = 0, \quad F_0''(0) + F_0^2(0) = \frac{1}{4}. \tag{A.18}$$

The second condition in (A.18) is satisfied by F_0 as seen from (A.17). The equation (A.17) is noted to have movable singularities that are second-order poles. Denoting the first (i.e., smallest) positive pole location by $\hat{a}_0$, we have the local behavior,

$$F_0 \sim -\frac{6\varepsilon^{1/2}}{(\hat{a}_0 - \hat{y})^2} \quad \text{as} \quad \hat{y} \to \hat{a}_0^-. \tag{A.19}$$

Matching F_0 with $\hat{F}_0$ from the inner region requires $F_0(\hat{a}) = \hat{f}_0(\hat{a})$, where $\hat{a} < \hat{a}_0$.

It is worth remarking that the leading order solution $\hat{f}_s(\hat{y})$ in this inner 2 region on the slow scale $\hat{y}$ satisfies (A.7). As $\hat{y} \to 0^+$, the only bounded positive solution satisfying $\hat{f}_s'(0) = 0$ has the behavior,

$$\hat{f}_s \sim \frac{1}{2} - \frac{1}{8}\,\hat{y}^2 \quad \text{as} \quad \hat{y} \to 0^+.$$

For the solution satisfying this behavior, we have that $\hat{f}_s = 0$ at $\hat{y} \approx 1.6$ with $\hat{f}_s'$ unbounded which gives a first approximation for $\hat{a}$. We now have from the numerical solution of (A.7) and (A.8) the estimate $\hat{f}_0(\hat{a}) \approx -0.2$ for $\hat{a} = 1.6$. Further, imposing this condition on the two-scale approximation (A.17) suggests taking $B_0 \approx 1/4\hat{a}^{1/3}$.

Full numerical solution of (5.2) gives the values $y \approx 1.5\varepsilon^{-1/8}, f \approx -0.21\varepsilon^{-1/4}$ for the location and value of the minimum of f when $\varepsilon = 6.6 \times 10^{-9}$. This is in very good agreement with the values $\hat{a} = 1.6$, $\hat{f}_0(\hat{a}) = -0.2$ from the above asymptotics.

There are two sets of solutions to (A.17) satisfying (A.18), which are distinguished by the their values for $F(0)$. Full numerical solution of (5.2) gives the estimates $F_0(0) \approx 0.5 \pm 0.167$.

Acknowledgments. The authors would like to thank K. Promislov for several useful discussions, A. Novick-Cohen for suggestions concerning the literature, and an anonymous referee for considerable editorial input.

REFERENCES

[1] O. V. Admaev and V. V. Pikhnachev, *Self-similar solutions of the equation* $u_t + \Delta^2 u + \Delta(u^2) = 0$, Preprint No. 3, Comp. center, Krasnoyarsk, Russia 1997.

[2] G. I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge University Press, Cambridge, 1996. Updated version of *Similarity, Self-Similarity, and Intermediate Asymptotics*, Consultants Bureau, New York, 1979.

[3] A. J. Bernoff and A. L. Bertozzi, *Singularities in a modified Kuramoto-Sivashinsky equation describing interface motion for phase transition*, Phys. D, 85 (1995), pp. 375–404.

[4] A. L. Bertozzi and M. C. Pugh, *Long-wave instabilities and saturation in thin film equations*, Comm. Pur. Appl. Math., LI (1998), pp. 625–651.

[5] A. L. Bertozzi and M. C. Pugh, *Finite-time blow-up of solutions of some long-wave unstable thin film equations*, Indiana Univ. Math. J., 49 (2000), pp. 1323–1366.

[6] C. J. Budd, V. A. Galaktionov, and J. F. Williams, *Self-similar blow-up in higher-order semilinear parabolic equations*, SIAM J. Appl. Math., 64 (2004), pp. 1775–1809.

[7] C. J. Budd, W. Huang, and R. D. Russell, *Moving mesh methods for problems with blow-up*, SIAM J. Sci. Comput., 17 (1996), pp. 305–327.

[8] C. J. Budd and J. F. Williams, *Optimal grids and uniform error estimates for PDEs with finite-time singularities*, in preparation.

[9] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York/London, 1955.

[10] Yu. V. Egorov, V. A. Galaktionov, V. A. Kondratiev, and S. I. Pohozaev, *Asymptotic behaviour of global solutions to higher-order semilinear parabolic equations in the supercritical range*, Adv. Differential Equations, 9 (2004), pp. 1009–1038.

[11] S. D. Eidelman, *Parabolic Systems*, North-Holland, Amsterdam/London, 1969.

[12] C. Elliott and Z. Songmu, *On the Cahn-Hilliard equation*, Arch. Ration. Mech. Anal., 96 (1986), pp. 339–357.

[13] A. Friedman, *Partial Differential Equations*, Robert E. Krieger, Malabar, FL, 1983.

[14] J. Funada, *Nonlinear Diffusion Equation for Interfacial Waves Generating by the Marangoni Effect*, Notes Res. Inst. Math. Sci. Kyoto Univ. 510, 1984.

[15] V. A. Galaktionov and J. L. Vazquez, *The problem of blow-up in nonlinear parabolic equations*, Discrete Contin. Dyn. Syst., 8 (2002), pp. 399–433.

[16] V. A. Galaktionov and J. F. Williams, *Blow-up in a fourth-order semilinear parabolc equation from explosion-convection theory*, European J. Appl. Math., 14 (2003), pp. 745–764.

[17] V. A. Galaktionov and J. F. Williams, *On very singular similarity solutions of a higher-order semilinear parabolic equation*, Nonlinearity, 17 (2004), pp. 1075–1099.

[18] I. C. Gohkberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.
[19] J. K. Hale, *Asymptotic Behavior of Dissipative Systems*, AMS, Providence, RI, 1988.
[20] W. Huang and R. D. Russell, *A moving collocation method for solving time dependent partial differential equations*, Appl. Numer. Math., 20 (1996), pp. 101–116.
[21] W. Huang and R. D. Russell, *Adaptive mesh movement—the mmpde approach and its applications. Numerical analysis*, J. Comput. Appl. Math., 1-2 (2001), pp. 383–398.
[22] A. Iserles, *On the global error of discretization methods for highly-oscillatory ordinary differential equations*, BIT, 42 (2002), pp. 561–599.
[23] A. Iserles, *Think globally, act locally: Solving highly-oscillatory ordinary differential equations*, Appl. Numer. Math., 43 (2002), pp. 145–160.
[24] A. Iserles, *On the numerical quadrature of highly-oscillating integrals* I: *Fourier transforms*, IMA J. Numer. Anal., 24 (2004), pp. 365–391.
[25] M. A. Krasnosel'skii and P. P. Zabreiko, *Geometrical Methods of Nonlinear Analysis*, Springer-Verlag, Berlin/Tokyo, 1984.
[26] J. C. Kurtz, *Weighted Sobolev spaces with applications to singular nonlinear boundary value problems*, J. Differential Equations, 49 (1983), pp. 105-123.
[27] H. A. Levine, *The role of critical exponents in blow-up problems*, SIAM Rev., 32 (1990), pp. 262–288.
[28] H. A. Levine, *Some nonexistence and instability theorems for solutions of formally parabolic equations of the form* $Pu_t = -Au + \mathcal{F}(u)$, Arch. Ration. Mech. Anal., 51 (1973), pp. 371–386.
[29] H. A. Levine and L. A. Payne, *Nonexistence theorems for the heat equation with nonlinear boundary conditions and for the porous medium equation backward in time*, J. Differential Equations, 16 (1974), pp. 319–334.
[30] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel/Berlin, 1995.
[31] A. Novick-Cohen, *Blow-up and growth in the directional solidification of dilute binary alloys*, Appl. Anal., 47 (1992), pp. 241–257.
[32] A. Novick-Cohen and L. A. Segel, *Nonlinear aspects of the Cahn–Hilliard equation*, Phys. D, 10 (1984), pp. 277–298.
[33] R. D. Russell, J. F. Williams, and X. Xu, *Movcol4: A high resolution moving collocation scheme for evolutionary equations*, SIAM J. Sci. Comp., 2005, submitted.
[34] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Blow-up in quasi-linear Parabolic Equations*, Walter de Gruyter, Berlin/New York, 1995.
[35] F. B. Weissler, *Rapidly decaying solutions of an ordinary differential equation with application to semilinear elliptic and partial differential equations*, Arch. Ration. Mech. Anal., 91 (1986), pp. 247–266.
[36] T. P. Witelski, A. J. Bernoff, and A. L. Bertozzi, *Blow-up and dissipation in a critical-case unstable thin film equation*, European J. Appl. Math., 15 (2004), pp. 223–256.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 103–115

GLOBAL CONTINUATION IN SECOND-GRADIENT NONLINEAR ELASTICITY*

ANITA MARENO† AND TIMOTHY J. HEALEY‡

Abstract. We consider three-dimensional elastic bodies characterized by a general class of stored-energy functions dependent upon the first and second gradients of the deformation. We assume that the dependence on the higher-order term ensures strong ellipticity. With only modest assumptions on the lower-order term, we use the Leray–Schauder degree to prove the existence of global solution continua to the Dirichlet problem. With additional, physically reasonable restrictions on the stored-energy function, we then demonstrate that our global solution branch is unbounded.

Key words. nonlinear elasticity, second gradient, global continuation

AMS subject classifications. 35G30, 65H20, 74G25, 74N99

DOI. 10.1137/050626065

1. Introduction.

In this work we consider the analysis of models for three-dimensional elastic bodies characterized by a general class of stored-energy functions dependent upon the first and second gradients of the deformation. We are motivated in part by the common use of second-gradient models to overcome the lack of regularity in nonlinear elasticity. Since the classical works of van der Waals [25], [29] and Cahn and Hilliard [6], the addition of small "interfacial energy" in multiwell problems associated with phase transitions is well known; see, e.g., [5], [22], [26], and references therein. Even problems within the setting of classical nonlinear elasticity (strongly elliptic) provide ample motivation for the use of higher-gradient models. For example, the potential failure of the complementing condition at the (smooth) boundary is a real impediment to existence [9], [11], [13].

We also mention the works [10] and [19]. Within the confines of one-dimensional, multiwell elasticity, these papers treat the existence of global solution branches. In particular, generalized solutions in the limit of vanishing interfacial energy are rigorously obtained. While this last difficult step is presently out of reach for three-dimensional problems, we establish a first important step in that process here, viz., the existence of unbounded solution branches for a general class of second-gradient models.

The outline of the paper is as follows: In section 2 we formulate the Dirichlet problem for a very general class of "forced" second-gradient models. We assume that the higher-order term of the stored energy function yields uniform strong ellipticity with only modest assumptions on the lower-order term. In section 3 we provide a global-continuation analysis via the Leray–Schauder degree, akin to that in [8] for classical first-gradient problems. The results are quite general, characterized by the usual two "Rabinowitz alternatives" [23] accompanied by the possibility of a bounded

*Received by the editors March 5, 2005; accepted for publication (in revised form) September 15, 2005; published electronically April 12, 2006. This work was supported in part by the National Science Foundation through grants DMS-0072514 and DMS-0406161, which is gratefully acknowledged. We thank Stefan Krömer for pointing out a gap in an earlier version of this work.
http://www.siam.org/journals/sima/38-1/62606.html
†Center for Applied Mathematics, Cornell University, Ithaca, NY 14853 (anitamareno@aol.com).
‡Theoretical and Applied Mechanics and Center for Applied Mathematics, Cornell University, Ithaca, NY 14853. Current address: Department of Mathematics and Computer Science, Dickinson College, Carlisle, PA 17013 (tjh10@cornell.edu).

solution branch that "terminates" due to the loss of local injectivity of solutions. Section 4 is the heart of the paper. We assume that the stored-energy function depends upon an interfacial-energetic term, quadratic in the second gradient of the deformation; we adopt the physically reasonable hypotheses from [12] for the lower-order term. Employing a recent uniqueness result [20, 21] for second-gradient systems and generalizing arguments from [12], we eliminate two of the "alternatives" from the general result of section 3, yielding the existence of an unbounded branch of classical injective solutions.

Notation. Throughout this work, the inner product and tensor product of $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$ are denoted as usual by $\mathbf{a} \cdot \mathbf{b}$ and $\mathbf{a} \otimes \mathbf{b}$, respectively. For second-order tensors we define the inner product by $\mathbf{A}{:}\mathbf{B} = \operatorname{tr}\left(\mathbf{A}^T\mathbf{B}\right) = A_{ij}B_{ij}$ (in Cartesian components), where tr denotes the trace and ^{T} denotes the transpose. Similarly, for third-order tensors $\mathbf{G}$ and $\mathbf{H}$ we define the inner product by $\mathbf{G} \vdots \mathbf{H} = G_{ijk}H_{ijk}$. For the most part, the third-order tensors encountered in this work are symmetric in their last two indices. Accordingly we define $V = \{\mathbf{G} : G_{ijk} = G_{ikj}\}$. For $\mathbf{G} \in V$ with components G_{ijk}, we define $\mathbf{G}^T$ to be the third-order tensor with components G_{jik}.

For Banach spaces X and Y, we let $L(X, Y)$ denote the space of bounded linear transformations from X to Y. (We let $L(X, X) \equiv L(X)$.) Then we define $GL\left(\mathbb{R}^3\right) = \{\mathbf{A} \in L(X) : \mathbf{A} \text{ is bijective}\}$, $GL^+\left(\mathbb{R}^3\right) = \left\{\mathbf{A} \in GL\left(\mathbb{R}^3\right) : \det \mathbf{A} > 0\right\}$, and $SO(3) = \left\{\mathbf{A} \in GL^+\left(\mathbb{R}^3\right) : \mathbf{A}^T = \mathbf{A}^{-1}\right\}$.

For the value of $T \in L(X, Y)$ at $x \in X$, we write $T[x]$. For a second-order tensor $\mathbf{A} \in L(\mathbb{R}^3)$, we write $\mathbf{A}\mathbf{x}$ instead of $\mathbf{A}[\mathbf{x}]$. For fourth-order tensors $\mathbf{C}$, $\mathbf{C}[\mathbf{A}]$ denotes the value of $\mathbf{C} \in L(L(\mathbb{R}^3))$ at $\mathbf{A} \in L(\mathbb{R}^3)$. Similarly, for sixth-order tensors $\mathbf{B}$, we write $\mathbf{B}[\mathbf{H}]$ for the value of $\mathbf{B}$ at a third-order tensor $\mathbf{H}$.

2. Problem formulation. We assume once and for all that an origin and an orthonormal basis have been chosen in three-dimensional Euclidean space, which we identify with the space $\mathbb{R}^3$. In this work, the reference configuration is chosen to coincide with a natural or stress-free configuration. The region occupied in the reference configuration is the closure $\bar{\Omega}$ of a domain Ω of $\mathbb{R}^3$, and we call $\bar{\Omega}$ itself the elastic body. We further assume that $\partial\Omega$ is locally of class C^5.

Deformations of the body are defined by mappings $\mathbf{f} : \bar{\Omega} \mapsto \mathbb{R}^3$ such that $\mathbf{f} \in C^2(\bar{\Omega}, \mathbb{R}^3)$ and $\nabla\mathbf{f}(\mathbf{x}) \in GL^+(\mathbb{R}^3)$. For a given deformation, we define the displacement field by $\mathbf{u}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{x}$ and the deformation gradient by $\mathbf{F} = \nabla\mathbf{f}(\mathbf{x}) = \mathbf{I} + \nabla\mathbf{u}(\mathbf{x}) \in GL^+(\mathbb{R}^3)$. The second gradient of $\mathbf{f}$ is denoted by $\nabla^2\mathbf{f}(\mathbf{x}) = \nabla\mathbf{F}(\mathbf{x})$, which is a third-order tensor field.

We assume that the body is subjected to a live body force $\mathbf{b}(\lambda, \mathbf{x}, \mathbf{u}, \nabla\mathbf{u})$ in Ω and a prescribed displacement $\mathbf{d}(\lambda, \mathbf{x})$ on $\partial\Omega$, where $\lambda \in \mathbb{R}$ is a loading parameter. We assume that at least one of these two fields is not identically zero and that

$$\mathbf{b}(0, \mathbf{x}, \mathbf{u}, \nabla\mathbf{u}) = \mathbf{d}(0, \mathbf{x}) \equiv \mathbf{0}. \tag{1}$$

The elastic body is presumed homogeneous and its material response is described by a stored energy function $W : V \times GL^+(\mathbb{R}^3) \to \mathbb{R}$, denoted $W(\mathbf{G}, \mathbf{F})$, such that the total internal potential energy is given by

$$E(\mathbf{f}) = \int_\Omega W(\nabla^2\mathbf{f}, \nabla\mathbf{f})\, dV. \tag{2}$$

The stored energy W is required to satisfy the principle of material objectivity:

$$W(\mathbf{Q}\mathbf{G}, \mathbf{Q}\mathbf{F}) = W(\mathbf{G}, \mathbf{F}) \text{ for all } \mathbf{Q} \in SO(3), \mathbf{G} \in V,\ \mathbf{F} \in GL^+\left(\mathbb{R}^3\right). \tag{3}$$

By a *natural* reference configuration we mean

$$\frac{\partial W(\mathbf{0},\mathbf{I})}{\partial \mathbf{G}} = \mathbf{0} \text{ and } \frac{\partial W(\mathbf{0},\mathbf{I})}{\partial \mathbf{F}} \equiv \mathbf{0}. \tag{4}$$

Furthermore, we make the following smoothness assumptions:

$$W \in C^4(V \times GL^+\left(\mathbb{R}^3\right), \mathbb{R}),$$

$$\mathbf{b} \in C^3(\mathbb{R} \times \bar{\Omega} \times \mathbb{R}^3 \times L(\mathbb{R}^3), \mathbb{R}^3),$$

$$\mathbf{d} \in C^5(\mathbb{R} \times \bar{\Omega}, \mathbb{R}^3). \tag{5}$$

We impose the geometric boundary conditions

$$\mathbf{u} = \mathbf{d} \text{ and } \frac{\partial \mathbf{u}}{\partial \mathbf{n}} = \frac{\partial \mathbf{d}}{\partial \mathbf{n}} \text{ on } \partial\Omega. \tag{6}$$

For our fourth-order problem (cf. (8) below), the conditions in (6) correspond to ideal "hard" loading. The displacement condition $(6)_1$ is common in the classical first-gradient theory, while $(6)_2$ represents the limiting case where the loading device also controls the normal derivative of the applied displacement. In plate theory, the conditions in (6) correspond to clamped edges [14].

To obtain the equilibrium equations, we calculate the first variation of $E(\mathbf{f})$ (cf. (2)) and equate it to the virtual work of the body force field $\mathbf{b}\left(\lambda, \mathbf{x}, \mathbf{u}, \nabla\mathbf{u}\right)$ acting through an admissible variation. After integrating by parts, we find the strong form of the equilibrium equations:

$$-\nabla \cdot \left(\nabla \cdot \frac{\partial W(\mathbf{G},\mathbf{F})}{\partial \mathbf{G}} - \frac{\partial W(\mathbf{G},\mathbf{F})}{\partial \mathbf{F}} \right) + \mathbf{b} = 0 \text{ in } \Omega. \tag{7}$$

Expanding (7) we obtain

$$-\mathbf{A}(\nabla^2\mathbf{u}, \mathbf{I} + \nabla\mathbf{u})[\nabla^4\mathbf{u}] + \mathbf{B}(\nabla^2\mathbf{u}, \mathbf{I} + \nabla\mathbf{u})[\nabla^3\mathbf{u}] + \mathbf{C}(\nabla^2\mathbf{u}, \mathbf{I} + \nabla\mathbf{u})[\nabla^2\mathbf{u}]$$

$$+ \mathbf{g}(\nabla^3\mathbf{u}, \nabla^2\mathbf{u}, \nabla\mathbf{u}) + \mathbf{b}\left(\lambda, \mathbf{x}, \mathbf{u}, \nabla\mathbf{u}\right) = 0 \text{ in } \Omega, \tag{8}$$

where (in components)

$$(\mathbf{A}(\mathbf{G},\mathbf{F})[\nabla^4\mathbf{u}])_i = \frac{\partial^2 W(\mathbf{G},\mathbf{F})}{\partial G_{ijk} \partial G_{lmn}} \frac{\partial^4 u_l}{\partial x_j \partial x_k \partial x_m \partial x_n},$$

$$(\mathbf{B}(\mathbf{G},\mathbf{F})[\nabla^3\mathbf{u}])_i = \frac{\partial^2 W(\mathbf{G},\mathbf{F})}{\partial F_{ij} \partial G_{klm}} \frac{\partial^3 u_k}{\partial x_j \partial x_l \partial x_m} - \frac{\partial^2 W(\mathbf{G},\mathbf{F})}{\partial F_{lm} \partial G_{ijk}} \frac{\partial^3 u_l}{\partial x_j \partial x_k \partial x_m},$$

$$(\mathbf{C}(\mathbf{G},\mathbf{F})[\nabla^2\mathbf{u}])_i = \frac{\partial^2 W(\mathbf{G},\mathbf{F})}{\partial F_{ij} \partial F_{kl}} \frac{\partial^2 u_k}{\partial x_j \partial x_l}, \tag{9}$$

and $\mathbf{g}(\cdot)$ is a vector-valued function of "higher order" in its first two arguments, viz.,

$$\mathbf{g}(\mathbf{U}, \mathbf{V}, \mathbf{W})/(|\mathbf{U}| + |\mathbf{V}|) \to \mathbf{0} \text{ as } |\mathbf{U}| + |\mathbf{V}| \to 0 \text{ for all } \mathbf{W}. \tag{10}$$

In particular, $\mathbf{g}(\cdot)$ makes no contribution to the linearization of (8) about $\mathbf{u} = \mathbf{0}$. Thus, its precise form is not important for our purposes here. Throughout this work we assume that the uniform strong ellipticity condition holds, i.e., that there exist positive constants c_1 and c_2 such that

$$c_1 \mid \mathbf{a} \mid^2 \mid \mathbf{b} \mid^4 \leq \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{b} \vdots \mathbf{A}(\mathbf{G}, \mathbf{F}) [\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{b}] \leq c_2 \mid \mathbf{a} \mid^2 \mid \mathbf{b} \mid^4 \tag{11}$$

for all $\mathbf{a}$,$\mathbf{b} \in \mathbb{R}^3 - \{\mathbf{0}\}$ and $(\mathbf{G}, \mathbf{F}) \in V \times GL^+ \left(\mathbb{R}^3\right)$. In addition, at the reference state, $(\mathbf{G}, \mathbf{F}) = (\mathbf{0}, \mathbf{I})$, we make the following reasonable assumptions on the coefficient tensors $\mathbf{B}, \mathbf{C}$ appearing in (8):

$$\mathbf{B}(\mathbf{0}, \mathbf{I}) = \mathbf{0}, \tag{12}$$

$$\mathbf{a} \otimes \mathbf{b} : \mathbf{C}(\mathbf{0}, \mathbf{I})[\mathbf{a} \otimes \mathbf{b}] > 0 \tag{13}$$

for all $\mathbf{a}$,$\mathbf{b} \in \mathbb{R}^3 - \{\mathbf{0}\}$.

Observe that (12) is weaker than assuming $\frac{\partial^2 W(\mathbf{0},\mathbf{I})}{\partial \mathbf{F} \partial \mathbf{G}} = \mathbf{0}$. The latter yields the physically reasonable condition that the quadratic-order coupling between $\mathbf{G}$ and $\mathbf{F}$ in $W(\mathbf{G}, \mathbf{F})$ vanishes. Of course (13) is simply the strong ellipticity condition for the classical linear elasticity tensor $\mathbf{C}(\mathbf{0}, \mathbf{I})$.

3. Existence of solution branches. For simplicity of presentation, we treat the case $\mathbf{d} \equiv \mathbf{0}$; cf. (6). The more general case is handled similarly after an appropriate change of variables; see, e.g., [8]. For $\alpha \in (0, 1)$, let $C^{m,\alpha}(\overline{\Omega}, \mathbb{R}^3)$ denote the usual Hölder space of m-times (Hölder) continuously differentiable functions, with norm $\|\cdot\|_{m,\alpha}$. The following Banach spaces play an important role in our forthcoming analysis:

$$Z = \left\{ \mathbf{u} \in C^{4,\alpha}(\bar{\Omega}, \mathbb{R}^3) : \mathbf{u} = \frac{\partial \mathbf{u}}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega \right\}; \|\mathbf{u}\|_Z \equiv \|\mathbf{u}\|_{4,\alpha},$$

$$X = \left\{ \mathbf{u} \in C^{3,\alpha}(\bar{\Omega}, \mathbb{R}^3) : \mathbf{u} = \frac{\partial \mathbf{u}}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega \right\}; \|\mathbf{u}\|_X \equiv \|\mathbf{u}\|_{\mathbf{3},\alpha},$$

$$Y = C^{0,\alpha}(\bar{\Omega}, \mathbb{R}^3); \|\mathbf{u}\|_Y \equiv \|\mathbf{u}\|_{0,\alpha}. \tag{14}$$

Since $\det \nabla \mathbf{f} > 0$ (pointwise) for a deformation $\mathbf{f} : \bar{\Omega} \to \mathbb{R}^3$, any solution of our problem (7), (6) should belong to the open subset

$$\mathcal{U} = \{\mathbf{u} \in X : \det(\mathbf{I} + \nabla \mathbf{u}) > 0 \text{ in } \bar{\Omega}\}. \tag{15}$$

In view of (1) and (4), we observe that $(\lambda, \mathbf{u}) = (0, \mathbf{0})$ is a solution (7), (6). Define $\mathcal{O}$ to be the maximal connected set in $\mathcal{U}$ containing $\mathbf{u} = \mathbf{0}$, i.e.,

$$\mathcal{O} = comp\{\mathbf{0}\} \text{ in } \mathcal{U}. \tag{16}$$

For every $\mathbf{u} \in \mathcal{O}$, we define a linear operator $L(\mathbf{u}) : Z \to Y$ via

$$L(\mathbf{u})[\mathbf{h}] \equiv -\mathbf{A}(\nabla^2 \mathbf{u}, \mathbf{I} + \nabla \mathbf{u}) \left[\nabla^4 \mathbf{h}\right] + \mathbf{B}(\nabla^2 \mathbf{u}, \mathbf{I} + \nabla \mathbf{u})[\nabla^3 \mathbf{h}]$$

$$+\,\mathbf{C}(\nabla^2\mathbf{u}, \mathbf{I}+\nabla\mathbf{u})[\nabla^2\mathbf{h}] \text{ for all } \mathbf{h}\in Z. \tag{17}$$

In view of (11), $L(\mathbf{u})$ defines a uniformly elliptic operator for any $\mathbf{u}\in\mathcal{O}$. Accordingly we have the following spectral estimate.

PROPOSITION 3.1. *There are positive constants ϵ, c_1, c_2, independent of $\mu, \mathbf{h}, \lambda, \mathbf{u}$, such that*

$$\|\mathbf{h}\|_Z \le c_1\,|\mu|^{\frac{\alpha}{4}}\,\|\,(L(\mathbf{u})-\mu I)\,[\mathbf{h}]\,\|_Y \tag{18}$$

for all $\mathbf{h}\in Z$, and for all $\mu\in\mathbb{C}$ such that $|\arg(\mu)|\le\frac{\pi}{2}+\epsilon$ and $|\mu|\ge c_2$. The mapping $I: Z\to Y$ is the identity map and $\alpha\in(0,1)$ is the Hölder exponent inherent in Z and Y.

Proof. The original L^p version of this result is due to Agmon [1]. For convenience we provide a detailed proof in the appendix for our (Hölder-space) setting (see also [30], [16]). □

By Proposition 3.1, for each $\mathbf{u}\in\mathcal{O}$, the linear operator $L(\mathbf{u})-a_oI: Z\to Y$ is injective for some sufficiently large real number $a_o > c_2$. Bijectivity of $L(\mathbf{u})-a_oI$ then follows from the Fredholm (of index zero) property. (For example, this last step is readily facilitated via a one-parameter homotopy of our operator with the vector-valued biharmonic operator, $-\Delta^2\mathbf{h}$, subject to the same homogeneous boundary conditions; cf. (6) with $\mathbf{d}\equiv\mathbf{0}$.) Next, for any $\mathbf{u}\in\mathcal{O}$, we consider the linear, uniformly elliptic boundary value problem

$$(L(\mathbf{u})-a_oI)\,[\mathbf{h}] \equiv -g(\nabla^3\mathbf{u}, \nabla^2\mathbf{u}, \nabla\mathbf{u}) - \mathbf{b}(\lambda, \mathbf{x}, \mathbf{u}, \nabla\mathbf{u}) - a_o\mathbf{u},$$

$$\mathbf{h} = \frac{\partial\mathbf{h}}{\partial\mathbf{n}} = \mathbf{0} \text{ on } \partial\Omega, \tag{19}$$

which has a unique solution, denoted $\mathbf{h}=G(\lambda,\mathbf{u})$, where $G:\mathbb{R}\times\mathcal{O}\to X$. By virtue of (5), the Schauder estimate [2] for (19), and the compact embedding $Z\to X$, we see that G is compact and continuous (*completely continuous*).

Moreover, any solution $(\lambda,\mathbf{u})\in\mathbb{R}\times\mathcal{O}$ of (7), (6) (with $\mathbf{d}\equiv\mathbf{0}$) satisfies

$$\mathbf{u} - G(\lambda,\mathbf{u}) = \mathbf{0}, \tag{20}$$

and conversely. Let $\mathcal{M}\subset\mathcal{O}$ be open and bounded. Since G is completely continuous, the Leray–Schauder degree of the mapping $\mathbf{u}\longmapsto G(\lambda,\mathbf{u})$ on $\mathcal{M}$, denoted $\deg(I-G(\lambda,\cdot),\mathcal{M},\mathbf{0})$, is well defined. Before stating the main result of this section, we need the following proposition.

PROPOSITION 3.2. *The mapping $\mathbf{u}\longmapsto G(0,\mathbf{u})$ is Frèchet differentiable at $\mathbf{u}=\mathbf{0}$, with derivative denoted by $D_uG(0,\mathbf{0})\equiv T$. The compact linear operator $T: X\to X$ is defined via the solution $\mathbf{h}=T[\mathbf{f}]$ of the linear, constant-coefficient elliptic system*

$$(L(\mathbf{0})-a_oI)\,[\mathbf{h}] \equiv -\mathbf{A}(\mathbf{0},\mathbf{I})\left[\nabla^4\mathbf{h}\right] + \mathbf{C}(\mathbf{0},\mathbf{I})\left[\nabla^2\mathbf{h}\right] - a_o\mathbf{h} \;= -a_o\mathbf{f} \text{ in } \Omega,$$

$$\mathbf{h} = \frac{\partial\mathbf{h}}{\partial\mathbf{n}} = \mathbf{0} \text{ on } \partial\Omega$$

for all $\mathbf{f}\in Y$. Moreover, for $\varepsilon>0$ sufficiently small, we have

$$\deg(I-G(0,\cdot), B_\varepsilon(\mathbf{0}), \mathbf{0}) = \pm 1, \tag{21}$$

where $B_\varepsilon(\mathbf{0}) = \{\mathbf{u} \in X : \|\mathbf{u}\| < \varepsilon\}$.

Proof. For differentiability, it suffices to show that $G(0,\mathbf{u}) - T[\mathbf{u}]$ is $o(\|\mathbf{u}\|_X)$ near $\mathbf{u} = \mathbf{0}$. We argue by contradiction, similar to the argument given in [8] (see also [24]). Namely, if the claim is not true, then there exist a positive constant α and a sequence $\mathbf{u}_j \to \mathbf{0}$ in $B_\varepsilon(\mathbf{0}) \subset X$ such that

$$\|G(0,\mathbf{u}_j) - T[\mathbf{u}_j]\| > \alpha \|\mathbf{u}_j\|_X \text{ for } j \text{ sufficiently large.} \tag{22}$$

Let $\mathbf{v}_j \equiv G(0,\mathbf{u}_j)/\|\mathbf{u}_j\|_X$ and $\mathbf{w}_j \equiv T[\mathbf{u}_j]/\|\mathbf{u}_j\|_X$, which satisfy

$$L(\mathbf{u}_j)[\mathbf{v}_j] - a_o\mathbf{v}_j = -(g(\nabla^3\mathbf{u}_j, \nabla^2\mathbf{u}_j, \nabla\mathbf{u}_j) + a_o\mathbf{u}_j)/\|\mathbf{u}_j\|_X \text{ in } \Omega,$$

and

$$L(\mathbf{0})[\mathbf{w}_j] - a_o\mathbf{w}_j = -a_o\mathbf{u}_j/\|\mathbf{u}_j\|_X \text{ in } \Omega,$$

respectively, subject to

$$\mathbf{v}_j = \mathbf{w}_j = \frac{\partial \mathbf{v}_j}{\partial \mathbf{n}} = \frac{\partial \mathbf{w}_j}{\partial \mathbf{n}} = \mathbf{0} \text{ on } \partial\Omega.$$

Clearly $\mathbf{u}_j/\|\mathbf{u}_j\|_X$ is uniformly bounded in X, and likewise, by virtue of the Schauder estimates, the sequences $\mathbf{v}_j$ and $\mathbf{w}_j$ are uniformly bounded in Z. By compact embedding, we conclude that each sequence has a convergent subsequence (not relabeled): $\mathbf{u}_j/\|\mathbf{u}_j\|_X \to \phi$ in $C^3(\bar\Omega, \mathbb{R}^3)$, $\mathbf{v}_j \to \mathbf{v}$ and $\mathbf{w}_j \to \mathbf{w}$ in $C^4(\bar\Omega, \mathbb{R}^3)$. Moreover, from (8), (9), (12), and (13), we find that $\mathbf{v}$ and $\mathbf{w}$ each satisfy the boundary value problem

$$L(\mathbf{0})[\mathbf{h}] - a_o\mathbf{h} = -a_o\phi \text{ in } \Omega,$$

$$\mathbf{h} = \frac{\partial \mathbf{h}}{\partial \mathbf{n}} = \mathbf{0} \text{ on } \partial\Omega. \tag{23}$$

By uniqueness, we conclude that $\mathbf{v} \equiv \mathbf{w}$. Hence,

$$\|\mathbf{v}_j - \mathbf{w}_j\|_X = \|G(0,\mathbf{u}_j) - T[\mathbf{u}_j]\| / \|\mathbf{u}_j\|_X \to 0 \text{ as } j \to \infty,$$

which contradicts (22).

Next, we consider the injectivity of the mapping $I - T \in L(X)$:

$$(I - T)[\mathbf{h}] = \mathbf{0} \Leftrightarrow T[\mathbf{h}] = \mathbf{h},$$

the second equation of which is equivalent to (by the definition of T; cf. also (12))

$$-\mathbf{A}(\mathbf{0},\mathbf{I})\left[\nabla^4\mathbf{h}\right] + \mathbf{C}(\mathbf{0},\mathbf{I})\left[\nabla^2\mathbf{h}\right] = \mathbf{0} \text{ in } \Omega,$$

$$\mathbf{h} = \frac{\partial \mathbf{h}}{\partial \mathbf{n}} = \mathbf{0} \text{ on } \partial\Omega. \tag{24}$$

Using (11) and (13), it is not hard to show (via the Fourier transform; cf. [28]) that the linear, *constant-coefficient*, strongly elliptic system (24) has the *unique* solution $\mathbf{h} \equiv \mathbf{0}$; i.e., $I - T$ is injective and thus bijective by the Riesz–Schauder theory. Accordingly,

$\mathbf{u} = \mathbf{0}$ is an isolated solution of (20) at $\lambda = 0$. By the linearization principle and the index formula [15], we conclude (for $\varepsilon > 0$ sufficiently small)

$$\deg(I - G(0,\cdot), B_\varepsilon(\mathbf{0}), \mathbf{0}) = \deg(I - T, B_\varepsilon(\mathbf{0}), \mathbf{0}) = \pm 1. \qquad \square$$

We also need the following characterization of the open set $\mathcal{O}$ in order to state our basic existence result: Define

$$\mathcal{O}_\delta \equiv \{\mathbf{u} \in \mathcal{O} : \det(\mathbf{I} + \nabla \mathbf{u}) > \delta \text{ in } \bar{\Omega}\}. \tag{25}$$

Observe that

$$\overline{\mathcal{O}_\delta} \subset \mathcal{O} \text{ for each } \delta > 0, \text{ and } \mathcal{O} = \cup_{\delta > 0} \mathcal{O}_\delta. \tag{26}$$

THEOREM 3.3. *Let $\Sigma \subset \mathbb{R} \times \mathcal{O}$ denote the connected component of solution pairs $(\lambda, \mathbf{u})$ of (17) containing the solution point $(0, \mathbf{0})$. Then at least one of the following holds:*

(A1) *Σ is unbounded in $\mathbb{R} \times X$;*

(A2) *$\Sigma - \{(0, \mathbf{0})\}$ is connected;*

(A3) *$\Sigma \not\subset \mathbb{R} \times \mathcal{O}_\delta$ for each $\delta > 0$.*

Proof. The proof is a straightforward generalization of a well-known argument; cf. [23]. For each fixed $\delta > 0$, we suppose that Σ is not characterized by any of the properties (A1)–(A3). We then obtain a contradiction from (21) and the homotopy invariance of the degree of $I - G$. We refer the reader to [17] for the details. $\square$

Remark 3.4. If property (A2) holds, then there are nontrivial solutions to problem (20) with null loading $\lambda = 0$. Property (A3) indicates a breakdown of local injectivity along a solution branch., viz., there is a sequence of solution points $(\lambda_j, \mathbf{u}_j) \subset \Sigma$ such that

$$\inf_{\mathbf{x} \in \bar{\Omega}} \det(\mathbf{I} + \nabla \mathbf{u}_j(\mathbf{x})) \searrow 0 \quad \text{as } j \to \infty. \tag{27a}$$

In particular, we note that Theorem 3.3 leaves open the possibility that (27a) holds without (A1) or (A2) being true. In this case the bounded branch could "terminate" at a point $(\lambda_*, \mathbf{u}_*)$ where $(\lambda_j, \mathbf{u}_j) \to (\lambda_*, \mathbf{u}_*)$ in $C^3(\bar{\Omega}, \mathbb{R}^3)$ by compact imbedding and

$$\inf_{\mathbf{x} \in \bar{\Omega}} \det(\mathbf{I} + \nabla \mathbf{u}_*(\mathbf{x})) = 0, \tag{27b}$$

which would seem to contradict (34).

Before ending this section we make an important observation: By construction, for every $(\lambda, \mathbf{u}) \in \Sigma$, the corresponding deformation $\mathbf{f}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$ satisfies $\det(\mathbf{I} + \nabla \mathbf{u}) > 0$ on $\bar{\Omega}$. Moreover, $\mathbf{f}(\mathbf{x}) = \mathbf{x} + \mathbf{d}(\lambda, \mathbf{x})$ on $\partial\Omega$. We now quote a well-known result ensuring that the deformation is injective.

COROLLARY 3.5. *Suppose that for each λ, the mapping*

$$\mathbf{x} \to \mathbf{x} + \mathbf{d}(\lambda, \mathbf{x}) \text{ is injective on } \bar{\Omega}. \tag{28}$$

Then for every $(\lambda, \mathbf{u}) \in \Sigma$, the deformation $\mathbf{f}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$ is injective on $\bar{\Omega}$.

Proof. For a proof, see [4, Theorem 5.5-2].

4. Existence of unbounded branches. In this section we make additional, physically reasonable assumptions on the stored energy function W sufficient to eliminate properties (A2) and (A3) of Theorem 3.3 on bounded solution branches, leading to the existence of unbounded solution branches to our problem. Specifically we henceforth assume that W has the form

$$W(\mathbf{F},\mathbf{G}) = \hat{W}(\mathbf{F}) + \epsilon h(\mathbf{F},\mathbf{G}) = \hat{W}(\mathbf{F}) + \frac{\epsilon}{2}\mathbf{G} \vdots \mathbf{A}(\mathbf{F})[\mathbf{G}], \tag{29}$$

where $\mathbf{G} \in V$, $\mathbf{A}(\mathbf{F})$ is a sixth-order-tensor-valued function on $GL^+(\mathbb{R}^3)$, and $\epsilon > 0$. We assume throughout that the second-gradient term is positive semidefinite, viz.,

$$h(\mathbf{F},\mathbf{G}) \geq 0 \text{ for all } \mathbf{F} \in GL^+(\mathbb{R}^3), \mathbf{G} \in V. \tag{30}$$

The specific quadratic form of the higher-gradient term in (29) is motivated, e.g., by the works of Triantafyllidis and Bardenhagen [27] and Hilgers and Pipkin [14], where ϵ is a small parameter representing a lattice length scale in the former and plate thickness in the latter. In this work ϵ plays no role. Nonetheless, we carry it along throughout as a reminder that (29) is useful in applications when ϵ is "small."

We now specialize our boundary value problem (6), (7) to (29) with null loading, i.e., $\lambda = 0$ (cf. (1)):

$$\begin{aligned} \nabla \cdot \left[-\epsilon \nabla \cdot \frac{\partial h}{\partial \mathbf{G}} + \left(\frac{\partial \hat{W}}{\partial \mathbf{F}} + \epsilon \frac{\partial h}{\partial \mathbf{F}} \right) \right] &= \mathbf{0} \text{ in } \Omega, \\ \mathbf{u} &= \mathbf{0} \text{ on } \partial\Omega, \\ \frac{\partial \mathbf{u}}{\partial n} &= \mathbf{0} \text{ on } \partial\Omega. \end{aligned} \tag{31}$$

In view of (4), observe that $\mathbf{u} \equiv \mathbf{0}$ is a solution of (31). With two additional, physically reasonable hypotheses, it turns out that $\mathbf{u} \equiv \mathbf{0}$ is the only classical solution, i.e., the only solution in $C^4(\Omega,\mathbb{R}^3) \cap C^1(\bar{\Omega},\mathbb{R}^3)$. The following is a special case of the main result in [21].

PROPOSITION 4.1. *Given* (29) *and* (30)*, assume that*

$$\hat{W}(\mathbf{I}) < \hat{W}(\mathbf{F}) \text{ for all } \mathbf{F} \in GL^+(\mathbb{R}^3) - SO(3); \tag{32}$$

i.e., $\hat{W}(\mathbf{I})$ is the global minimum of $\hat{W}(\cdot)$. Furthermore, suppose that

$$\Omega \text{ is a star-shaped domain.} \tag{33}$$

Then $\mathbf{u} \equiv \mathbf{0}$ is the only classical solution of (31).

Remark. Condition (32) ensures that $\hat{W}(\cdot)$ is quasi-convex at $\mathbf{F} = \mathbf{I}$.

An immediate consequence of Proposition 4.1 is that the solution branch $\Sigma - \{(0,\mathbf{0})\}$ has two disjoint components $\Sigma^+ \subset (0,\infty) \times X$ and $\Sigma^- \subset (-\infty,0) \times X$, separated by the hyperplane $\{0\} \times X$, which, in turn, implies the disjoint union

$$\Sigma = \Sigma^+ \cup \Sigma^- \cup \{(0,\mathbf{0})\}.$$

COROLLARY 4.2. *Assume the hypotheses of Theorem* 3.3 *and Proposition* 4.1. *Then the solution branch Σ is characterized by alternatives* (A1) *and/or* (A3) *of Theorem* 3.3.

Now we show with certain constitutive hypotheses that the branch Σ is unbounded. Here we list all of the assumptions sufficient to ensure the unboundedness of the branch.

Suppose that for some $\Psi : GL^+(\mathbb{R}^3) \to \mathbb{R}$ and $\Gamma : (0,\infty) \to \mathbb{R}$, $\hat{W}$ admits the decomposition

$$\hat{W} = \Psi(\mathbf{F}) + \Gamma(\det \mathbf{F})$$

for all $\mathbf{F} \in GL^+(\mathbb{R}^3)$,

$$\text{where } \Gamma(\det \mathbf{F}) \to \infty \text{ as } \det \mathbf{F} \searrow 0. \tag{34}$$

In addition, let

$$\Gamma \in C^3(0,\infty),\ \Psi \in\ C^3(GL^+(\mathbb{R}^3),\mathbb{R}) \cap C^2(U,\mathbb{R}),$$

$$\text{where } U = \{\mathbf{F} \in \overline{(GL^+(\mathbb{R}^3))} : ||\mathbf{F}|| < \infty\}. \tag{35}$$

We further assume that

$$\Gamma\prime(d) < 0 \text{ for } 0 < d < d_o, \text{ for some constant } d_o > 0. \tag{36}$$

We find it convenient to define the following quantities:

$$\Phi(d) = \Gamma(d) - d\Gamma^{'}(d) \text{ for all } d > 0, \tag{37}$$

$$\mathbf{P} = W\mathbf{I} - \mathbf{F}^T\left[\frac{\partial W}{\partial \mathbf{F}} - \epsilon\nabla\cdot\frac{\partial h}{\partial \mathbf{G}}\right] - \epsilon\mathbf{G}^T\frac{\partial h}{\partial \mathbf{G}}, \tag{38}$$

$$\tilde{\mathbf{P}}(\mathbf{F}) = \hat{W}\mathbf{I} - \mathbf{F}^T\frac{d\hat{W}}{d\mathbf{F}}, \tag{39}$$

$$\hat{\mathbf{P}}(\mathbf{F}) = \mathbf{F}^T\frac{d\Psi(\mathbf{F})}{d\mathbf{F}} - \Psi(\mathbf{F})\mathbf{I} \text{ for all } \mathbf{F} \in GL^+(\mathbb{R}^3). \tag{40}$$

LEMMA 4.3. *If* $(\lambda,\mathbf{u}) \in \Sigma$, *with* $J(\mathbf{x}) = \det \mathbf{F}(\mathbf{x})$ *for* $\mathbf{x} \in \bar{\Omega}$, *then*

$$\nabla(\Phi\circ J) = \nabla\cdot\hat{\mathbf{P}}(\mathbf{F}) + \mathbf{F}^T\mathbf{b}(\lambda,\mathbf{x},\mathbf{u},\nabla\mathbf{u}) + \epsilon\mathbf{F}^T\left[-\nabla\cdot\left(\nabla\cdot\frac{\partial h}{\partial \mathbf{G}}\right) + \nabla\cdot\frac{\partial h}{\partial \mathbf{F}}\right] \text{ on } \bar{\Omega}. \tag{41}$$

Proof. Using (7), (29), and (38), we see that

$$\begin{aligned}\nabla\cdot\mathbf{P} &= \nabla\cdot W\mathbf{I} - \nabla\cdot\left[\mathbf{F}^T\left(\frac{\partial W}{\partial \mathbf{F}} - \epsilon\nabla\cdot\frac{\partial h}{\partial \mathbf{G}}\right)\right] - \epsilon\nabla\cdot\left[\mathbf{G}^T\frac{\partial h}{\partial \mathbf{G}}\right]\\ &= \mathbf{F}^T\left(\epsilon\nabla\cdot\left(\nabla\cdot\frac{\partial h}{\partial \mathbf{G}}\right) - \nabla\cdot\frac{\partial W}{\partial \mathbf{F}}\right) = \mathbf{F}^T\mathbf{b}.\end{aligned}$$

By virtue of (39), this leads to

$$\nabla\cdot\mathbf{P} = \nabla\cdot\tilde{\mathbf{P}} - \epsilon\mathbf{F}^T\left[\nabla\cdot\frac{\partial h}{\partial \mathbf{F}} - \nabla\cdot\left(\nabla\cdot\frac{\partial h}{\partial \mathbf{G}}\right)\right] = \mathbf{F}^T\mathbf{b}.$$

Observing that

$$\frac{d\Gamma}{d\mathbf{F}} = \Gamma'(J)J\mathbf{F}^{-T},$$

we see using (34), (37), and (40) that $\tilde{\mathbf{P}} = -\hat{\mathbf{P}}(\mathbf{F}) + \Phi(J)\mathbf{I}$. Thus,

$$\nabla \cdot \left[-\hat{\mathbf{P}}(\mathbf{F}) + \Phi(J)\mathbf{I}\right] - \epsilon\mathbf{F}^T\left[\nabla \cdot \frac{\partial h}{\partial \mathbf{F}} - \nabla \cdot \left(\nabla \cdot \frac{\partial h}{\partial \mathbf{G}}\right)\right] = \mathbf{F}^T\mathbf{b}$$

and we obtain (41). □

Remark 4.4. Note that (38) is a second-gradient version of the elastic energy-momentum tensor. Also, (39) is the usual energy-momentum tensor for $\hat{W}$ deduced by Eshelby [7].

With these physically reasonable hypotheses in hand, we now show that Σ corresponds to an unbounded branch of solution pairs to problem (6), (8).

THEOREM 4.5. *Let the hypotheses of Theorem* 3.3*, Corollary* 3.5*, Proposition* 4.1*, and Lemma* 4.3 *hold. In addition assume the constitutive hypotheses* (34)–(36)*. Then* Σ *is a branch of globally injective solution pairs of* (6), (8) *that is unbounded in* $\mathbb{R} \times C^4(\bar{\Omega}, \mathbb{R}^3)$.

Proof. We generalize the proof in [12] in the context of our higher-gradient problem. Assume for the sake of contradiction that Σ is bounded in $\mathbb{R} \times C^4(\bar{\Omega}, \mathbb{R}^3)$. Then (A3) holds. Let $\{(\lambda_j, \mathbf{u}_j)\} \subset \Sigma$. By the Schauder estimates for (19), $\{(\lambda_j, \mathbf{u}_j)\}$ is uniformly bounded in Z (cf. (14)) and hence converges in $\mathbb{R} \times C^1(\bar{\Omega}, \mathbb{R}^3)$ by compact imbedding. We now write $\mathbf{f}_j = \mathbf{x}_j + \mathbf{u_j}$, $\mathbf{F}_j = \mathbf{I} + \nabla\mathbf{u}_j, \mathbf{G}_j = \nabla\mathbf{F}_j$, $J_j = \det\mathbf{F}_j$ on $\bar{\Omega}$. Using the triangle inequality and the conclusion of Lemma 4.3, we get

$$\|\nabla(\Phi \circ J_j)\Big\|_\infty \leq \|\frac{d\mathbf{P}(\mathbf{F}_j)}{d\mathbf{F}}\nabla^2\mathbf{u}_j\Big\|_\infty + \|\mathbf{F}_j^\top\mathbf{b}(\lambda_j, \mathbf{x}, \nabla\mathbf{u}_j, \mathbf{u}_j)\|_\infty$$

$$+ \epsilon\left\|\mathbf{F}^\top{}_j\left(\nabla \cdot \frac{\partial h(\mathbf{F}_j, \mathbf{G}_j)}{\partial \mathbf{F}} - \nabla \cdot \nabla \cdot \frac{\partial h(\mathbf{F}_j, \mathbf{G}_j)}{\partial \mathbf{G}}\right)\right\|_\infty, \tag{42}$$

where $\|\cdot\|_\infty$ denotes the supremum norm over $\bar{\Omega}$. By assumptions (35) and (37)–(40), the limit of the right-hand side of (42) is finite as $j \to \infty$. On the other hand, the left-hand side of (42) becomes unbounded as $j \to \infty$. To see this note that (27b) holds with$\|J_j\|_\infty$ bounded away from zero as $j \to \infty$. (This follows from Corollary 3.5 since $Vol(\mathbf{f}_j(\bar{\Omega})) = Vol(\bar{\Omega})$ for all j. On the other hand, if $J_j \to 0$ pointwise on $\bar{\Omega}$, then $Vol(\mathbf{f}_j(\bar{\Omega})) = \int_\Omega J_j dV \to 0$ which is a contradiction.) Consequently there are distinct points x_0 and $x_* \in \bar{\Omega}$ such that $\Phi(J_j(\mathbf{x}_o))$ remains bounded while $\Phi(J_j(\mathbf{x}_*)) \to \infty$ as $j \to \infty$. If we integrate $\nabla(\Phi \circ J_j)$ along any path in $\bar{\Omega}$ joining $\mathbf{x}_o$ and $\mathbf{x}_*$, we see from (34)–(40) that the left side of (42) grows without bound as $j \to \infty$. This is a contradiction, and we see that (A3) is possible only when Σ is unbounded. □

5. Concluding remarks. In practice (29) is often specialized further so that the sixth-order tensor-valued function $\mathbf{A}(\mathbf{F})$ is constant (i.e., independent of $\mathbf{F}$). In this case, it is not hard to show that the construction in section 3 is valid for X as defined in (14) but with the $C^{1,\alpha}(\bar{\Omega}, \mathbb{R}^3)$ topology. In particular, this leads to a strengthened version of Theorem 4.5.

Although Theorem 4.5 guarantees the existence of solutions "in the large" ("far" from the reference configuration), the absence of a priori bounds precludes global

existence of solutions. More specifically, the unbounded solution branch $\Sigma \subset \mathbb{R} \times \mathcal{O}$ may "blow up" in X for a finite value of the loading parameter $\lambda \in \mathbb{R}$. For the same reason it seems quite difficult to carry out the "singular limit of Σ" as $\epsilon \searrow 0$, as in [10], [19] in a one-dimensional setting.

Appendix. Spectral estimates.

PROPOSITION A.1. *For each* $\mathbf{u} \in \mathcal{O}$ *there exist positive constants* ϵ, c_1, c_2, *independent of* $\mu, \mathbf{h}, \lambda, \mathbf{u}$, *such that*

$$\|\mathbf{h}\|_Z \leq c_1 \mid \mu \mid^{\frac{\alpha}{4}} \| \left(A(\mathbf{u}) - \mu I\right)[\mathbf{h}] \|_Y \tag{43}$$

for all $\mathbf{h} \in Z$, *for all* $\mu \in \mathbb{C}$ *such that* $|\arg(\mu)| \leq \frac{\pi}{2} + \epsilon$ *and* $|\mu| \geq c_2$. *The mapping* $I : Z \to Y$ *is the identity map and* $\alpha \in (0,1)$ *is the Hölder exponent inherent in* Z *and* Y.

Proof. We follow the approach of Agmon [1] in the L^p setting. First we set $A(\mathbf{u}) \equiv A$, introduce a new variable t, and set $D_t = \frac{\partial}{\partial t}$. Consider the operator

$$\mathcal{L} = A - e^{i\theta} D_t^4. \tag{44}$$

For $|\theta| \leq \frac{\pi}{2} + \epsilon$, $\mathcal{L}$ is an elliptic operator of order 4 in the closure of the cylindrical domain $\Gamma = \{(\mathbf{x}, t) : \mathbf{x} \in \bar{\Omega}, t \in \mathbb{R}\}$, where ϵ depends on the constants in the definition of uniform ellipticity, and the complementing condition is satisfied for Dirichlet boundary conditions. Now we define

$$\mathbf{v}(\mathbf{x}, t) = \zeta(t) e^{i\nu t} \mathbf{h}(\mathbf{x}), \ \nu \in \mathbb{R}, \text{ and } \nu > 0,$$

and define $\Gamma_r = \bar{\Omega} \times [-r, r] \subset \mathbb{R}^4$, or the part of Γ in $\mid t \mid < r$, where $\zeta(t)$ is some fixed C^∞ function such that

$$\zeta(t) = \begin{cases} 0 & \text{for } \mid t \mid \geq 2, \\ 1 & \text{for } \mid t \mid \leq 1. \end{cases} \tag{45}$$

Applying the a priori estimate from Theorem 9.3 in [2] to $\mathbf{v}$, we obtain

$$\|\mathbf{v}\|_{4,\alpha;\Gamma_2} \leq c \left(\|\mathcal{L}\mathbf{v}\|_{0,\alpha;\Gamma_2} + \|\mathbf{v}\|_{0,\alpha;\Gamma_2} \right) \tag{46}$$

for a constant $c > 0$.

Using (44),

$$\mathcal{L}\mathbf{v} = \zeta(t) e^{i\nu t} \left(A - \nu^4 e^{i\nu t} I\right) \mathbf{h}(\mathbf{x}) + e^{i(\theta + \nu t)} \left[\sum_{k=0}^{3} \binom{4}{k} D_t^{4-k} \zeta(t) i^k \nu^k \right] \mathbf{h}(\mathbf{x}). \tag{47}$$

Since $\mathbf{v}(\mathbf{x}, t) = e^{i\nu t} \mathbf{h}(\mathbf{x})$ for $\mid t \mid \leq 1$ we obtain from (46)

(48)

$$\begin{aligned} \|e^{i\nu t} \mathbf{h}(\mathbf{x})\|_{4,\alpha;\Gamma_1} &\leq \|\mathbf{v}\|_{4,\alpha;\Gamma_2} \\ &\leq c \|\zeta(t) e^{i\nu t} \left(A - \nu^4 e^{i\nu t} I\right) \mathbf{h}(\mathbf{x})\|_{0,\alpha;\Gamma_2} \qquad (49) \\ &\quad + c \left\| e^{i(\theta + \nu t)} \left[\sum_{k=0}^{3} \binom{4}{k} D_t^{4-k} \zeta(t) i^k \nu^k \right] \mathbf{h}(\mathbf{x}) \right\|_{0,\alpha;\Gamma_2} \qquad (50) \\ &\quad + c \|\zeta(t) e^{i\nu t} \mathbf{h}(\mathbf{x})\|_{0,\alpha;\Gamma_2}. \qquad (51) \end{aligned}$$

Since $e^{i\nu t}$ is Hölder continuous, the inequality

$$\|\zeta(t)e^{i\nu t}\mathbf{h}(\mathbf{x})\ \|_{0,\alpha;\Gamma_2} \leq c_3\nu^{\alpha}\|\mathbf{h}\ \|_{0,\alpha;Y} \tag{52}$$

follows if we take ν sufficiently large.

Similarly,

$$\left\| e^{i(\theta+\nu t)} \left[\sum_{k=0}^{3} \binom{4}{k} D_t^{4-k}\zeta(t) i^k \nu^k \right] \mathbf{h}(\mathbf{x}) \right\|_{0,\alpha;\Gamma_2} \leq c_4 \sum_{k=0}^{3} \nu^{\alpha+k}\|\mathbf{h}\|_{0,\alpha;Y} \tag{53}$$

and

$$\|\zeta(t)e^{i\nu t}(A-\nu^4 e^{i\theta}I)\mathbf{h}\|_{0,\alpha;\Gamma_2} \leq c_3\nu^{\alpha}\|(A-\nu^4 e^{i\theta}I)\mathbf{h}\|_{0,\alpha;Y}. \tag{54}$$

On the other hand,

$$\|e^{i\nu t}\|_{4,\alpha;\Gamma_1} \geq \|\mathbf{h}\|_4 + c_5 \left(\sum_{k=0}^{3} \binom{4}{k} \nu^k \left[e^{i\nu t} D_x^{4-k}\mathbf{h}(\mathbf{x}) \right]_{0,\alpha;\Gamma_2} \right) \tag{55}$$

$$\geq c_6 \left(\|\mathbf{h}\|_{0,\alpha,Z} + \sum_{k=1}^{4} \nu^k \|\mathbf{h}\ \|_{4-k,\alpha,Z} \right) \tag{56}$$

$$= c_6 \sum_{k=0}^{4} \nu^k \|\mathbf{h}\|_{4-k,\alpha;Z}. \tag{57}$$

Therefore, using equations (48)−(51), (53), (54), and (55)−(57), we get

$$\sum_{k=0}^{4} \nu^k \|\mathbf{h}\|_{4-k,\alpha;Z} \leq c_1\nu^{\alpha} \left[\|(A-\nu^4 e^{i\theta}I)\mathbf{h}\|_{0,\alpha;Y} + \sum_{k=0}^{3} \nu^k \|\mathbf{h}\|_{0,\alpha;Y} \right]. \tag{58}$$

For ν sufficiently large, we may disregard the last term on the right-hand side, since it is dominated by the terms on the left-hand side. If we set $\mu = \nu^4 e^{i\theta}$, then

$$\sum_{k=0}^{4} |\mu|^{\frac{k-4}{4}} \|\mathbf{h}\|_{k,\alpha;Z} \leq c_1 |\mu|^{\frac{\alpha}{4}} \|(A-\mu I)\mathbf{h}\|_{0,\alpha;Y}$$

and (43) follows. □

REFERENCES

[1] S. Agmon, *On the eigenfunctions and on the eigenvalues of general elliptic boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 119–147.

[2] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[3] R. F. Brown, *A Topological Introduction to Nonlinear Analysis*, Birkhäuser Boston, Boston, 1993.

[4] P. G. Ciarlet, *Mathematical Elasticity Volume* I: *Three-Dimensional Elasticity*, Elsevier Science Publishers, Amsterdam, 1988.

[5] J. Carr, M. E. Gurtin, and M. Slemrod, *Structured phase transitions on a finite interval*, Arch. Ration. Mech. Anal., 86 (1984), pp. 317–351.

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

[7] J. D. Eshelby, *The elastic energy-momentum tensor*, J. Elasticity, 5 (1975), pp. 321–335.
[8] T. J. Healey, *Global continuation in displacement problems of nonlinear elastostatics via the Leray-Schauder degree*, Arch. Ration. Mech. Anal., 152 (2000), pp. 273–282.
[9] T. J. Healey, *Lectures on global continuation in three-dimensional elasticity*, in Proceedings of the NSA Regional Conference, UPR Humacao, 2002 (2006), to appear.
[10] T. J. Healey and H. Kielhöfer, *Global continuation via higher-gradient regularization and singular limits in forced one-dimensional phase transitions*, SIAM J. Math. Anal., 31 (2000), pp. 1307–1331.
[11] T. J. Healey and E. Montes, *Global bifurcation in nonlinear elasticity with an application to barreling states of cylindrical columns*, J. Elasticity, 71 (2003), pp. 33–58.
[12] T. J. Healey and P. Rosakis, *Unbounded branches of classical injective solutions to the forced displacement problem in nonlinear elastostatics*, J. Elasticity, 49 (1997), pp. 65–78.
[13] T. J. Healey and H. C. Simpson, *Global continuation in nonlinear elasticity*, Arch. Ration. Mech. Anal., 143 (1998), pp. 1–28.
[14] M. G. Hilgers and A. C. Pipkin, *Bending energy of highly elastic membranes*, Quart. Appl. Math., 50 (1992), pp. 389–400.
[15] M. A. Krasnoselkii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergammon Press, New York, 1964.
[16] H. Kielhöfer, *Existenz und Regularität von lösungen semilinearer parabolischer Anfangs-Randwertprobleme*, Math. Z., 142 (1975), pp. 131–160.
[17] H. Kielhöfer, *Bifurcation Theory*, Springer-Verlag, New York, 2004.
[18] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
[19] M. Lilli, T. J. Healey, and H. Kielhöfer, *Singular perturbation as a selection criterion for Young-Measure solutions*, 2006, to appear.
[20] A. Mareno, *Global Continuation in Higher-Gradient Nonlinear Elasticity*, Thesis, Cornell University, Ithaca, NY, 2002.
[21] A. Mareno, *Uniqueness of equilibrium solutions in second-order gradient nonlinear elasticity*, J. Elasticity, 74 (2004), pp. 99–107.
[22] S. Müller, *Variational models for microstructure and phase transitions*, in Calculus of Variations and Geometric Evolution Problems, S. Hildebrandt and M. Struwe, eds., Springer-Verlag, New York, 1999, pp. 85–210.
[23] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 478–513.
[24] P. H. Rabinowitz, *A global theorem for nonlinear problems and applications*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, London, 1971, pp. 11–36.
[25] J. S. Rowlinson, *Translation of J. D. van der Waals' "The thermodynamic theory of capillarity under the hypothesis of a continuation variation of density"*, J. Statist. Phys., 20 (1979), pp. 197–244.
[26] N. Triantafyllidis and E. Aifantis, *A gradient approach to localization of deformation* I. *Hyperelastic materials*, J. Elasticity, 16 (1986), pp. 225–237.
[27] N. Triantafyllidis and S. Bardenhagen, *Derivation of higher order gradient continuum theories in* 2, 3-*D nonlinear elasticity from periodic lattice models*, J. Mech. Phys. Solids, 42 (1994), pp. 111–139.
[28] L. Van Hove, *Sur l'extension de la condition de Legendre du calcul des variations aux intégrales mulitple à plusieurs fonctions inconnues,* Nederl. Akad. Wetensch., Proc., 50 (1947), pp. 18–23.
[29] J. D. van der Waals, *The thermodynamics theory of capillarity under the hypothesis of a continuous variation of density*, Verhandel, Konink. Akad. Weten. Amsterdam (Sect. 1) Vol. 1, No. 8 (1893) (in Dutch).
[30] W. Von Wahl, *Gebrochene Potenzen eines elliptischen Operators und parabolische Differentialgleichungen in Räumen hölderstetiger Funktionen*, Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II (1972), pp. 231–258.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 116–126

SPECTRAL ANALYSIS OF TRAVELING WAVES FOR NONLOCAL EVOLUTION EQUATIONS*

PETER W. BATES† AND FENGXIN CHEN‡

Abstract. Monotone traveling waves have been shown to exist for a broad class of nonlocal bistable evolution equations. In this note we examine the spectrum of the operator obtained by linearizing about such a traveling wave and hence show that these waves are exponentially asymptotically stable, up to translation.

DOI. 10.1137/S0036141004443968

Key words. AUTHOR MUST PROVIDE

AMS subject classifications. AUTHOR MUST PROVIDE

1. Introduction.

Perhaps the most familiar mathematical model for the evolution of phase boundaries in material science is the second order Allen–Cahn equation:

$$u_t = Du_{xx} + f(u), \tag{1.1}$$

where $D > 0$, f is a bistable function, i.e., f has exactly three roots, say, $u = \pm 1$ and $u = a \in (-1, 1)$ with $f'(\pm 1) < 0 < f'(a)$. It is well known that (1.1) has a traveling wave solution connecting $u = 1$ and $u = -1$, a solution of the form $u(x, t) = \phi(x - ct)$ for some constant c, and smooth function ϕ satisfying $\phi(\pm\infty) = \pm 1$. It is also known that the wave profile ϕ is monotone, it is unique up to translation, and it is exponentially asymptotically stable (see, e.g., [17]). This stability is in the sense that a perturbation leads to a solution converging to a translate of the wave. That result may be proved in various ways, one of which is by transforming (1.1) according to a frame moving with speed c and showing that the operator obtained by linearizing at ϕ has spectrum in the left half-plane, bounded away from the imaginary axis except for an algebraically simple eigenvalue at zero. The stability then follows from the spectral information by using abstract semigroup theory and invariant manifold/foliation results (see, e.g., [7] or [8] and [9]).

In many biological and physical settings, diffusion is not the only mechanism by which the state at one location affects the state at other locations, and it is appropriate to include in the mathematical model certain nonlocal interaction terms. Furthermore, there may be several types of long-range interactions acting at differing length scales. Thus, the reaction-diffusion equation (1.1) is replaced by a more general scalar evolution equation of the form

$$u_t - Du_{xx} - F(u, J_1 * s^1(u), \ldots, J_n * s^n(u)) = 0 \tag{1.2}$$

for all $x \in \mathbb{R}$ and $t \in \mathbb{R}^+$. Here $D \geq 0$ is a constant, $J * v(x, t) = \int_{\mathbb{R}} J(x - y)\, v(y, t)dy$ is spatial convolution, the kernels J_i are of class $C^1(\mathbb{R})$ and nonnegative,

* Received by the editors May 18, 2004; accepted for publication (in revised form) September 21, 2005; published electronically April 12, 2006.
http://www.siam.org/journals/sima/38-1/44396.html
†Department of Mathematics, Michigan State University, East Lansing, MI 48824. This author was supported in part by grants DMS 9970894 and DMS 9974340.
‡Department of Applied Mathematics, University of Texas, San Antonio, TX 78249 (feng@sphere.math.utsa.edu).

and s^i are smooth functions, $i = 1, 2, \ldots, n$. It is further assumed that $\bar{f}(u) \equiv F(u, s^1(u), \ldots, s^n(u))$ is a smooth bistable function.

Besides the Allen–Cahn equation given above, (1.2) includes its nonlocal version, which has been the subject of several recent studies (e.g., [1], [2], [3], [4], [5], [6], [18]),

$$u_t - \lambda(J * u - u) - f(u) = 0. \tag{1.3}$$

Here $f(u)$ is a bistable function, $\lambda > 0$, $J(x) \geq 0$, and $\int_{\mathbb{R}} J = 1$.

Other physical or biological models are included in (1.2). Some identified by Chen [12] and others may be of particular interest:

- continuum limit of an interacting particle system with Glauber dynamics and Kac potential (see [15]),

  $$u_t = \tanh(\beta J * u + h) - u,$$

  where $\beta > 1$ is the reciprocal of absolute temperature and h is an external field;
- excitatory neural model (see [16]),

  $$u_t = -u + J * S(u),$$

  where $S \in C^1(\mathbb{R})$ satisfies $S'(\cdot) > 0$ in $[0, 1]$, $S(0) = 0$, $S'(0) < 1$, $S(1) = 1$, and $S'(1) < 1$;
- thalamic model (see [13]),

  $$u_t = -\beta u + \alpha(1 - u)H^\epsilon(J * S(u) - \theta),$$

  where α and β are positive constants, H^ϵ is a smooth approximation of the Heaviside function, $S' > 0$, and $\theta \in (S(0), S(1))$ is a parameter.

Existence, uniqueness, and stability of a monotone traveling wave connecting the stable homogeneous states for the nonlocal Allen–Cahn equation (1.3) were established in [6] and for the general equation (1.2) in [12] (see also [11] for the nonautonomous almost periodic case). In those works comparison methods were used to show stability. However, the spectral theory for the operator obtained by linearizing (1.2) at the traveling wave, needed for perturbation results or more precise rates of convergence, is lacking. The purpose of this note is to address this issue.

Suppose $u(x, t) = \phi(x - c_0 t)$ is a traveling wave solution of (1.2) satisfying $\phi(\pm\infty) = \pm 1$. Let $\xi = x - c_0 t$. Then (ϕ, c_0) satisfies

$$D\phi'' + c_0\phi' + F(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi)) = 0 \text{ and } \phi(\pm\infty) = \pm 1 \tag{1.4}$$

for all $\xi \in \mathbb{R}$.

Let $\mathcal{L} = \mathcal{L}(\phi, c_0)$ be the linearized operator about the traveling wave ϕ, defined by

$$\begin{aligned} \mathcal{L}\psi = {} & D\psi'' + c_0\psi' + F_u(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi))\psi \\ & + \sum_{i=1}^{n} F_{p_i}(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi)) \cdot J_i * (s^i_u(\phi)\psi) \end{aligned} \tag{1.5}$$

on $D(\mathcal{L})$ to be specified in section 2. We prove that the spectrum of $\mathcal{L}$ consists of 0, the principle eigenvalue due to the translation invariance of the traveling waves, and the rest of it is located in the left half-plane bounded away from the imaginary axis.

As indicated above, an immediate consequence of this result is the exponential stability of traveling wave solutions. The conclusion derived from the spectral result contained in this paper when combined with the abstract theorems of [7], [8], and [9] gives more than exponential asymptotic stability, up to translation, of the traveling wave. In fact, one may conclude that the dynamics in a tubular neighborhood of the one-dimensional manifold formed by translates of the wave in the infinite dimensional phase space is foliated with codimension-one invariant stable manifolds. The leaves are transverse to the one-dimensional manifold and the dynamics is that of exponential decay to the base point, which is a suitable translate of the wave.

Even more can be said. Since the spectrum perturbs continuously, for interaction kernels $\tilde{J}_i$ near J_i but not necessarily positive, the same result holds. This is immediate from the theorems in [10] whereby the one-dimensional manifold persists as an invariant manifold, which will consist of stationary states if the $\tilde{J}_i$ are translation invariant but may have slow dynamics otherwise, with a neighborhood foliated with an invariant family of codimension-one stable manifolds with base point on the one-dimensional manifold. Comparison methods used in [12] are not applicable directly to obtain similar information when the interaction kernels are perturbed in this way.

2. Assumptions and main result. For f, J_i, and s, the following assumptions are made in [12]:

(A$_1$) $J_i(\cdot) \in C^1(\mathbb{R})$ and satisfies $J_i(\cdot) \geq 0$, $\int_{\mathbb{R}} J_i(y)dy = 1$, and $\int_{\mathbb{R}} |J_i'(y)|dy < \infty$, for $i = 1, 2, \ldots, n$. Furthermore, $\operatorname{supp} J_1 \bigcap (0, \infty) \neq \emptyset \neq \operatorname{supp} J_1 \bigcap (-\infty, 0)$.

(A$_2$) $F(u, p) = F(u, p_1, \ldots, p_n)$ and $s^1(u), \ldots, s^n(u)$ are smooth functions satisfying $F_{p_i}(u, p) \geq 0, s_u^i(u) \geq 0$, for all $(u, p) \in [-1 - \delta_0, 1 + \delta_0] \times \prod_{i=1}^n [s^i(-1 - \delta_0), s^i(1 + \delta_0)]$, and $i = 1, 2, \ldots, n$, where δ_0 is a positive constant. Furthermore, we assume $\bar{f}(\cdot) = F(\cdot, s^1(\cdot), \ldots, s^n(\cdot))$ is bistable, i.e., it has exactly three zeros ± 1 and an intermediate point q; and there exists an interval $[\alpha_1, \alpha_2] \subset (-1, 1)$ containing q such that $\bar{f}'(u) \geq 0$ for $u \in [\alpha_1, \alpha_2]$ and $\bar{f}'(u) < 0$ for $u \notin [\alpha_1, \alpha_2]$.

(A$_3$) Either (i) $D > 0$, or (ii) $F_u(u, p) < 0$ and $F_{p_1}(u, p)s_u^1(u) > 0$ on $[-1 - \delta_0, 1 + \delta_0] \times \prod_{i=1}^n [s^i(-1 - \delta_0), s^i(1 + \delta_0)]$ for some $i \in \{1, 2, \ldots, n\}$.

The following theorem is proved in [12] (see also [11]).

THEOREM 2.1. *Assume that* (A$_1$)–(A$_3$) *hold. Then*

(a) (1.2) *admits a traveling wave solution* (ϕ, c_0) *satisfying* $\phi(\pm\infty) = \pm 1$,

(b) ϕ *is strictly increasing and of class* C^1,

(c) *the traveling wave solution of* (1.2) *is unique modulo spatial translation,*

(d) *the traveling wave and its translates are an asymptotically stable family with asymptotic phase.*

Let (ϕ, c_0) be a traveling wave solution obtained in Theorem 2.1. We are going to study the spectrum of the linearized operator $\mathcal{L}$ about the traveling wave solution (ϕ, c_0) satisfying (1.4).

Let

$$X_0 = \left\{ u \in C(\mathbb{R}) \ : \ \lim_{|x| \to \infty} u(x) = 0 \right\}$$

and $\mathcal{L}$ be the linearized operator about the traveling wave (ϕ, c_0) as defined in (1.5) with domain $D(\mathcal{L})$, where $D(\mathcal{L}) = \{u \in X_0 \ : \ u', u'' \in X_0\}$ if $D > 0$; $D(\mathcal{L}) = \{u \in X_0 \ : \ u' \in X_0\}$ if $D = 0$ and $c_0 \neq 0$; and $D(\mathcal{L}) = X_0$ if $D = 0$ and $c_0 = 0$.

Let us define a normal point for an operator $\mathcal{L}$ on a Banach space to be any complex number which is in the resolvent set $\rho(\mathcal{L})$ or is an isolated eigenvalue of $\mathcal{L}$ of

finite multiplicity. The complement of the set of normal points is called the essential spectrum of $\mathcal{L}$ and is denoted by $\sigma_{ess}(\mathcal{L})$. The following is the main result of this section.

THEOREM 2.2.

(i) $\{\lambda \ : \ \mathrm{Re}\lambda \geq 0, \lambda \neq 0\} \subset \rho(\mathcal{L})$.

(ii) *0 is an algebraically simple eigenvalue with a positive eigenfunction* ϕ'.

(iii) $\sigma_{ess}(\mathcal{L}) \subset \{\lambda \ : \ Re\lambda < -\gamma_0\}$, *where* $\gamma_0 > 0$ *is a constant.*

(iv) *The adjoint operator* $\mathcal{L}^*$ *has a positive eigenfunction corresponding to the simple eigenvalue* 0.

An immediate consequence of this is the following (see [20]).

COROLLARY 2.3.

(i) *The traveling wave solution* $\phi(x - c_0 t)$ *and its translates is an exponentially stable family with asymptotic phase.*

(ii) *There exist positive constants* γ *and* C *such that*

$$\|e^{\mathcal{L}t}u\| \leq Ce^{-\gamma t}\|u\|.$$

for all u *in the range of* $\mathcal{L}$.

3. Proof of Theorem 2.2. We divide the proof of the theorem into several lemmas. We will apply comparison principle and perturbation theory to estimate the spectrum. First we need the following comparison principle (see also [11] and [12]). Here we will be assuming that $t_0 > \tau$ are fixed times, $D \geq 0$ and c_0 are constant, R_1 is an open set and $R_2 = \mathbb{R} \setminus R_1$ is the complement of R_1, $K_0(x,t) \in L^\infty(\mathbb{R} \times [\tau, t_0])$, $\hat{J}$ is an integral operator with nonnegative kernel $\hat{J}(x,y,t)$, i.e., $(\hat{J}u)(x,t) = \int_{\mathbb{R}} \hat{J}(x,y,t)u(y,t)dy$, and $\hat{J}(x,y,t) \geq 0$ satisfies $\operatorname{ess\,sup}_{x\in\mathbb{R}\ t\in[\tau,t_0]} \int_{\mathbb{R}} \hat{J}(x,y,t)dy < \infty$. Furthermore, J_1 is that mentioned in (A_1).

LEMMA 3.1 (comparison principle). *Assume that* $u \in C^1([\tau,t_0], L^\infty(\mathbb{R}))$ *and that* $u \in C([\tau,t_0], W^{1,\infty}(R_1))$ *if* $D = 0$ *and* $c_0 \neq 0$ *and* $u \in C^{1,2}((\tau,t_0] \times R_1)$ *if* $D \neq 0$. *Suppose that* $u(x,t) \geq 0$ *for all* $t \in [\tau,t_0]$ *and almost all* $x \in R_2$ *and* $u(x,t)$ *satisfies*

$$u_t - Du_{xx} - c_0 u_x - K_0(x,t)u - (\hat{J}u)(x,t) \geq 0 \tag{3.1}$$

for all $t \in (\tau,t_0]$ *and almost all* $x \in R_1$. *If* $u(x,\tau) \geq 0$ *for almost all* $x \in \mathbb{R}$, *then* $u(x,t) \geq 0$ *for all* $t \in [\tau,t_0]$ *and almost all* $x \in \mathbb{R}$. *Moreover, if* $\hat{J}(x,y,t) \geq c_1 J_1(x-y)$ *for some* $c_1 > 0$, *and* $u(\cdot,\tau) \neq 0$ *in* $L^\infty(\mathbb{R})$, *then* $u(x,t) > 0$ *for all* $t \in (\tau,t_0]$ *and almost all* $x \in R_1$.

Proof. We may assume $\tau = 0$. Let $\bar{u}(x,t) = u(x - c_0 t, t)$. Then $\bar{u}(x,t)$ satisfies (3.1) on $\bar{R}_1$ with $c_0 = 0$, $K_0(x,t)$ replaced by $\bar{K}_0(x,t) = K_0(x - c_0 t, t)$ and $\hat{J}(x,y,t)$ replaced by $\bar{J}(x,y,t) = \hat{J}(x - c_0 t, y - c_0 t, t)$, where $\bar{R}_1 = \{(x,t) \mid (x - c_0 t, t) \in R_1 \times [0,t_0]\}$.

By assumption, $\operatorname{ess\,inf}_{x\in\mathbb{R}} \bar{u}(x,t)$ is continuous. If the first conclusion of the lemma is not true, then there exist constants $\epsilon > 0, T \in (0,t_0]$ such that $\bar{u}(x,t) > -\epsilon e^{2Kt}$ for all $0 < t < T$ and almost all $x \in \mathbb{R}$, and $\operatorname{ess\,inf}_{x\in\mathbb{R}} \bar{u}(x,T) = -\epsilon e^{2KT}$, where

$$K = D + \|K_0\|_{L^\infty} + \operatorname*{ess\,sup}_{x\in\mathbb{R}\ t\in[0,t_0]} \int_{\mathbb{R}} \hat{J}(x,y,t)dy. \tag{3.2}$$

Therefore, there exist an interval $[a,b]$ and a set $S_0 \subset [a,b]$ with positive Lebesgue measure such that $\bar{u}(x,T) \leq -\frac{15}{16}\epsilon e^{2KT}$ for $x \in S_0$. Let $z(x)$ be a smooth function such that $z(x) = \min_{x\in\mathbb{R}} z(x) = 1$ for $x \in [a,b]$, $\sup_{x\in\mathbb{R}} z(x) = z(\pm\infty) = 3$, $|z'(x)| \leq 1$, and

$|z''(x)| \le 1$. Define $w_\sigma(x,t) = -\epsilon\left(\frac{3}{4} + \sigma\ z(x)\right) e^{2Kt}$ for $\sigma \in [0,1]$. Since $w_{\frac{1}{4}}(x,t) \le \bar{u}(x,t)$ for $0 \le t \le T$ and almost all $x \in \mathbb{R}$, and $w_{\frac{1}{8}}(x,T) > \bar{u}(x,T)$ for $x \in S_0$, there is a minimum $\sigma^* \in \left(\frac{1}{8}, \frac{1}{4}\right]$ such that $w_{\sigma^*}(x,t) \le \bar{u}(x,t)$ for $t \in [0,T]$ and almost all $x \in \mathbb{R}$. We claim that there exist $(x_n, t_n) \in \bar{R}_1$ and $(\bar{x}_0, \bar{t}_0)$ such that inequality (3.1), with $u, K_0, \hat{J}$ replaced by $\bar{u}, \bar{K}_0, \bar{J}$, respectively, holds for (x_n, t_n); $\lim_{n\to\infty}(x_n, t_n) = (\bar{x}_0, \bar{t}_0)$; $\lim_{n\to\infty}\{\bar{u}(x_n,t_n) - w_{\sigma^*}(x_n,t_n)\} = 0$, the essential infimum of $\bar{u}(x,t) - w_{\sigma^*}(x,t)$ on $\mathbb{R}\times[0,T]$; and $\lim_{n\to\infty}(\bar{u} - w_{\sigma^*})_t(x_n,t_n) \le 0$. If $D > 0$, we also have $\lim_{n\to\infty}(\bar{u} - w_{\sigma^*})_{xx}(x_n,t_n) \ge 0$.

We first prove the claim for $D = 0$. Let $v(x,t) = \bar{u}(x,t) - w_{\sigma^*}(x,t)$ and $\rho(t) = \operatorname{ess\,inf}_{x\in\mathbb{R}} v(x,t)$. Then $\rho(0) > 0$. Let $\bar{t}_0 = \max\{t \in (0,t_0] : \rho(\tau) > 0 \text{ for all } 0 \le \tau < t\}$. Note that $w_{\sigma^*}(\pm\infty, t) \le -\frac{9}{8}\,\epsilon e^{2Kt} < \bar{u}(x,t)$ and $\bar{u}(x,t) > w_{\sigma^*}(x,t)$ for $t \in (0,T]$ and almost all $x \in R_2 + c_0 t$. For each $t < \bar{t}_0$, since $\rho(\bar{t}_0) = 0 < \rho(t)$, there is a uniformly bounded set $A(\bar{t}_0, t) \subset R_1 + c_0 t$ of positive Lebesgue measure such that $v(x,\bar{t}_0) \le \rho(t)$ and inequality (3.1), with $u, K_0, \hat{J}$ replaced by $\bar{u}, \bar{K}_0, \bar{J}$, respectively, holds for all $x \in A(\bar{t}_0, t)$. Therefore, for $\bar{t}_n = \bar{t}_0 - \frac{1}{n}$, we have $0 \le v(x,\bar{t}_n) - v(x,\bar{t}_0) = \int_0^1 v_t(x, \bar{t}_0 + s(\bar{t}_n - \bar{t}_0))ds(\bar{t}_n - \bar{t}_0)$ for $x \in A(\bar{t}_0, \bar{t}_n)$. Therefore, there exist $t_n \in (\bar{t}_n, \bar{t}_0)$ and a bounded sequence $x_n \in A(\bar{t}_0, \bar{t}_n)$ such that $v_t(x_n, t_n) \le 0$. After taking a subsequence of x_n we may assume that the limit of x_n exists: call it $\bar{x}_0$. Then $v(x_n,t_n) \le |v(x_n,\bar{t}_0)| + |v(x_n,t_n) - v(x_n,\bar{t}_0)| \le |v(x_n,\bar{t}_0)| + \|v_t\|_{L^\infty(\mathbb{R})}(\bar{t}_0 - t_n) \to 0$, where we have used the fact that $|v(x_n,\bar{t}_0)| \le \rho(\bar{t}_n) \to 0$.

If $D > 0$, (x_n, t_n) can be chosen as the minimum point of $\bar{u}(x,t) - w_{\sigma^*}(x,t)$ in $\mathbb{R}\times[0,T]$. This proves the claim.

Therefore,

$$
\begin{aligned}
0 &\ge \lim_{n\to\infty}(\bar{u} - w_{\sigma^*})_t(x_n,t_n) - \lim_{n\to\infty} D(\bar{u} - w_{\sigma^*})_{xx}(x_n,t_n)\\
&\ge \liminf_{n\to\infty}(\bar{J}\bar{u})(x_n,t_n) + \epsilon e^{2K\bar{t}_0}\left[2K\left(\sigma^* z(\bar{x}_0) + \frac{3}{4}\right) + D\sigma^* z''(\bar{x}_0)\right] + \|K_0\|_{L^\infty} w_{\sigma^*}(\bar{x}_0,\bar{t}_0)\\
&\ge \epsilon e^{2K\bar{t}_0}\left[\frac{7}{4}\ K - \frac{1}{4}D - \frac{3}{2}\|K_0\|_{L^\infty} - \frac{3}{2}\operatorname*{ess\,sup}_{\substack{x\in\mathbb{R}\\ t\in[0,t_0]}}\int_{-\infty}^{+\infty}\hat{J}(x,y,t)dy\right] > 0,
\end{aligned}
$$

which is a contradiction. Therefore $u(x,t) \ge 0$ for almost all $x \in \mathbb{R}$ and $t \in [\tau, t_0]$.

For the conclusion of the last part of the lemma, if $D > 0$, it follows from the comparison principle of parabolic equations [19]. If $D = 0$, let $\bar{v}(x,t) = e^{Kt}\bar{u}(x,t)$. Then we have $\bar{v}_t(x,t) \ge c_1 J_1 * \bar{v}(x,t)$ for $(x,t) \in \bar{R}_1$ since $u(x,t) \ge 0$. Therefore, $\bar{v}(x,t) \ge c_1(t - t_*)J_1 * \bar{v}(x,t_*)$, where t_* is any number less than t such that $(x,t_*) \in \bar{R}_1$. After the Nth iteration, we have $\bar{v}(x,t) \ge \frac{(c_1(t-t_*))^N}{N!} J_1 * \cdots * J_1 * \bar{v}(x,t_*)$. We first consider the first case for x where t_* can be chosen to be 0. We have $\bar{v}(x,t) \ge \frac{(c_1(t-t_*))^N}{N!} J_1 * \cdots * J_1 * \bar{u}(x,0)$. If $u(x,0) \ne 0$ in $L^\infty(\mathbb{R})$, we can choose N large enough such that $J_1 * \cdots * J_1 * u(x,0) > 0$. Therefore, we have $\bar{v}(x,t) > 0$. The rest of the case is reduced to prove that $\bar{v}(x,t_*) > 0$ on a set in $R_1 + c_0 t_*$ with positive measure. After finite steps of reduction, with different choices of x and t_*, it can be reduced to the first case. This completes the proof. □

We are going to use perturbation theory to estimate the essential spectrum of $\mathcal{L}$. First, let ζ be a smooth function satisfying $\zeta(x) = 0$ for $x \le -1$, $\zeta(x) = 1$ for $x \ge 1$, and $\zeta'(x) > 0$ for $x \in (-1,1)$. Let

$$
h(\cdot) = \bar{f}'(-1)(1 - \zeta(\cdot)) + \bar{f}'(1)\zeta(\cdot), \tag{3.3}
$$

where $\bar{f}$ is defined in (A$_2$). Note that $\bar{f}'(\pm 1) < 0$, by assumption (A$_2$). Consider the operator $\mathcal{L}_0\psi$ on X_0 with domain $D(\mathcal{L}_0) = D(\mathcal{L})$ defined by

$$\mathcal{L}_0\psi = D\psi'' + c_0\psi' + h\psi + \sum_{i=1}^{n} k_i(J_i * \psi - \psi), \tag{3.4}$$

where

$$k_i(\cdot) = F_{p_i}(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi))s_u^i(\phi)(\cdot) \tag{3.5}$$

for $i = 1, \ldots, n$. We also define

$$k_0(\cdot) = F_u(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi))(\cdot) \tag{3.6}$$

for later use. We want to show that $\sigma_{ess}(\mathcal{L}) = \sigma_{ess}(\mathcal{L}_0)$ if $D > 0$ or $c_0 \neq 0$. First we prove the next lemma.

LEMMA 3.2. *For $\lambda \in \mathbb{C}$ satisfying $\mathrm{Re}\lambda > \max\{\bar{f}'(-1), \bar{f}'(+1)\}$, the operator $\lambda - \mathcal{L}_0 : D(\mathcal{L}_0) \to X_0$ is injective.*

Proof. Let $\lambda = \alpha + i\beta \in \mathbb{C}$ with $\alpha > \max\{\bar{f}'(-1), \bar{f}'(1)\}$ and let $u(x) = u^1(x) + iu^2(x) \in D(\mathcal{L}_0)$ be a solution of $\mathcal{L}_0 u = \lambda u$. We will prove that $u \equiv 0$.

Consider the Cauchy problem

$$v_t = \mathcal{L}_0 v - \alpha v, \tag{3.7}$$

$$v(x, 0) = u^1(x). \tag{3.8}$$

It has a unique solution $v(x,t) = u^1(x)\cos\beta t - u^2(x)\sin\beta t$.

On the other hand, since $\alpha > \max\{\bar{f}'(-1), \bar{f}'(+1)\}$, we can choose a constant $\rho_0 > 0$ such that $\alpha > \max\{\bar{f}'(-1), \bar{f}'(+1)\} + \rho_0$. Let $\bar{v}(t) = \max_{x\in\mathbb{R}} |u(x)|e^{-\rho_0 t}$. Then

$$\bar{v}_t - \mathcal{L}_0\bar{v} + \alpha\bar{v} = (-\rho_0 - h(x) + \alpha)\bar{v} \geq (\alpha - \rho_0 - \max\{\bar{f}'(-1), \bar{f}'(+1)\})\bar{v} > 0.$$

By the comparison lemma, we deduce that $|v(x,t)| \leq \bar{v}(t)$ for all $x \in \mathbb{R}$ and $t > 0$. Therefore, if $\beta = 0$, the fact that $|u^1(x)| = |v(x,t)| \leq \bar{v}(t) \to 0$ as $t \to \infty$ implies $u^1(x) \equiv 0$. The same argument applying to $u^2(x)$ yields $u^2(x) \equiv 0$. If $\beta > 0$, $|v(x,t)| = |v(x, t + 2n\pi/\beta)| \leq \bar{v}(t + 2n\pi/\beta)$. Letting $n \to \infty$, we deduce $u(x) \equiv 0$. This proves the injectivity. □

To prove the surjectivity of $\lambda - \mathcal{L}_0$, we first consider the case $D > 0$. Let $M \equiv M_m = \{u \in C(\mathbb{R}) : u(x) = 0 \text{ for } |x| \geq m\}$, where m is a positive constant, and consider the operator $\mathcal{A}_0$ on M defined by $\mathcal{A}_0 u = Du'' + c_0u' + h(x)u$ for $u \in D(\mathcal{A}_0) = M \bigcap C^2(\mathbb{R})$. We have the next lemma.

LEMMA 3.3. *If $D > 0$, then $\{\lambda : \mathrm{Re}\lambda > \max\{\bar{f}'(-1), \bar{f}'(+1)\}\} \subset \rho(\mathcal{A}_0)$.*

Proof. Similar to the proof of Lemma 3.2, we can prove that $\lambda - \mathcal{A}_0$ is injective. The lemma follows from the Fredholm theory for elliptic operators on bounded domains. □

With this preparation, we can prove the following.

LEMMA 3.4. *For $\lambda \in \mathbb{C}$ satisfying $\mathrm{Re}\lambda > \max\{\bar{f}'(-1), \bar{f}'(+1)\}$, the operator $\lambda - \mathcal{L}_0 : D(\mathcal{L}_0) \to X_0$ is surjective.*

Proof. First let us assume $D > 0$. Let θ be a smooth nonnegative function on $\mathbb{R}$ satisfying $\theta(x) = 1$ for $|x| \leq \frac{1}{2}$ and $\theta(x) = 0$ for $|x| \geq 1$. Let $\mathcal{A}_1 \equiv \mathcal{A}_{1m}$ be the operator defined by $\mathcal{A}_1 u = \mathcal{A}_0 u + \sum_{i=1}^{n} \tilde{k}_i(x)(J_i * u - u)$ for $u \in D(\mathcal{A}_1) = D(\mathcal{A}_0)$,

where $\tilde{k}_i(\cdot) = k_i(\cdot)\theta(\cdot/m)$ and m is a positive number. We will prove the surjectivity of $\mathcal{A}_1$ and let m approach the infinite to prove the surjectivety of $\lambda - \mathcal{L}_0$.

For $\lambda \in \mathbb{C}$ with $\mathrm{Re}\lambda > \max\{\bar{f}'(-1), \bar{f}'(+1)\}$, as in the proof of Lemma 3.2, one can prove that $\lambda - \mathcal{A}_1$ is injective. On the other hand, for $g \in M$, the equation

$$(\lambda - \mathcal{A}_1)u = g \tag{3.9}$$

is equivalent to $u - R(\lambda, \mathcal{A}_0)\sum_{i=1}^n \tilde{k}_i(x)(J_i * u - u) = g_1$, where $g_1 = R(\lambda, \mathcal{A}_0)g$ and $R(\lambda, \mathcal{A}_0)$ is the resolvent of $\lambda - \mathcal{A}_0$. Since $\lambda - \mathcal{A}_1$ is injective and $R(\lambda, \mathcal{A}_0)$ is compact, by the Fredholm Alternative, $u - R(\lambda, \mathcal{A}_0)\sum_{i=1}^n \tilde{k}_i(x)(J_i * u - u) = g_1$ has a solution. Therefore, for any $g \in M$, $(\lambda - \mathcal{A}_1)u = g$ has a solution.

For $g \in X_0$, let u_m be the solutions of (3.9) corresponding to $g_m = g(\cdot)\theta(\cdot/m)$. We claim that $\|u_m\|_{C(\mathbb{R})} \le C\|g\|_{C(\mathbb{R})}$ for some constant C. To prove this, suppose $\lambda = \alpha + i\beta$, $u_m(x) = u_m^1(x) + iu_m^2(x)$ and $g_m(x) = g_m^1(x) + ig_m^2(x)$. Then it is easy to verify that the problem

$$v_t - \mathcal{A}_1 v + \alpha v = g_m^1(x)\cos\beta t - g_m^2(x)\sin\beta t, \tag{3.10}$$

$$v(x,0) = u_m^1(x) \tag{3.11}$$

has the unique solution $v_m(x,t) = u_m^1(x)\cos\beta t - u_m^2(x)\sin\beta t$. Let $\bar{v}(t) = \|u_m\|_{C(\mathbb{R})}e^{-\rho_0 t} + \frac{1}{\rho_0}(1 - e^{-\rho_0 t})\|g\|_{C(\mathbb{R})}$, where $\rho_0 > 0$ is a constant satisfying $\alpha > \max\{\bar{f}'(-1), \bar{f}'(+1)\} + \rho_0$. Applying the comparison principle to $\bar{v} \pm v_m$ and using (3.10), we conclude that $|v_m(x,t)| \le \bar{v}(t) \le \|u_m\|_{C(\mathbb{R})}e^{-\rho_0 t} + \frac{1}{\rho_0}(1 - e^{-\rho_0 t})\|g\|_{C(\mathbb{R})}$. If $\beta = 0$, we have $|u_m^1(x)| = |v_m(x,t)| \le \bar{v}(t) \to \frac{1}{\rho_0}\|g\|_{C(\mathbb{R})}$ as $t \to \infty$. The same argument applies for estimate of $u^2(x)$. If $\beta \neq 0$, for each x, choose $t_n \to \infty$ such that $v_m(x,t_n) = |u_m(x)|$, implying $|u_m(x)| \le \frac{1}{\rho_0}\|g\|_{C(\mathbb{R})}$. This completes the proof of the claim.

Since $\|u_m\|_{C(\mathbb{R})}$ is uniformly bounded, there exist a function $u \in X_0$ and a subsequence, which we label the same, such that u_m converges to u in the weak $*$ topology. Note that the dual space of linear bounded functionals is represented by the space of complex Borel measures. It is easily seen that u satisfies $\lambda u - \mathcal{L}_0 u = g$ in the sense of distributions. Therefore u is a solution in the classical sense. This completes the proof of the lemma when $D > 0$.

For the case $D = 0$, we choose a positive sequence such that $\lim_{j\to\infty} D_j = 0$. Let u_j be the solutions of $-D_j u'' - c_0 u' + (\lambda - h(x)) - \sum_{i=1}^n k_i(x)(J^i * u - u) = g$. Since the proof above does not involve the size of D, we have that $\|u_j\|_{C(\mathbb{R})} \le C\|g\|_{C(\mathbb{R})}$. Again, by taking a subsequence, we can assume u_j converges to some function $u \in X_0$ in the weak $*$ topology. Similar to the above argument, we see that u is a solution to $(\lambda - \mathcal{L}_0)u = g$. This completes the proof. □

As an immediate consequence of Lemma 3.2 and 3.4, one gets $\{\lambda : \mathrm{Re}\lambda > \max\{\bar{f}'(-1), \bar{f}'(1)\}\} \subset \rho(\mathcal{L}_0)$.

Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be bounded linear operators on X_0 defined by $\mathcal{L}_1 u(\cdot) = (\bar{f}'(\phi) - h)u(\cdot) = \{F_u(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi)) + \sum_{i=1}^n k_i - h\}u(\cdot)$ and $\mathcal{L}_2 u(\cdot) = \sum_{i=1}^n \{F_{p_i}(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi))J_i * (s_u(\phi)u) - k_i J_i * u\}(\cdot)$, respectively, for $u \in X_0$. Then we have the next lemma.

LEMMA 3.5. *Let $\lambda \in \rho(\mathcal{L}_0)$. Then if $D > 0$ or $c_0 \neq 0$, $\mathcal{L}_i(\lambda - \mathcal{L}_0)^{-1}$ is a compact operator on X_0 for $i = 1, 2$.*

Proof. Let B be a bounded closed set in X_0. Since $\|(\lambda - \mathcal{L}_0)^{-1} f\|_{C^1(\mathbb{R})} \le C\|f\|_{C(\mathbb{R})}$ if $D > 0$ or $c_0 \neq 0$, $(\lambda - \mathcal{L}_0)^{-1}B$ is bounded in $C^1(\mathbb{R})$ and therefore it is compact

in $C([-n,n])$ for all $n>0$. A diagonal argument can be used to get a sequence $\{u_k\}\subset(\lambda-\mathcal{L}_0)^{-1}B$ such that u_k converges to some u in $C([-n,n])$ for each fixed n. Note that $\mathcal{L}_iu(x)$ $(i=1,2)$ converges to zero as $|x|\to\infty$ uniformly for u in a bounded set in X_0. It easily follows that $\mathcal{L}_i(\lambda-\mathcal{L}_0)^{-1}u_k$ converges to $\mathcal{L}_i(\lambda-\mathcal{L}_0)^{-1}u$ in $C(\mathbb{R})$, respectively, for $i=1,2$. □

Now we are ready to prove Theorem 2.2(iii).

LEMMA 3.6. *There exists a positive constant γ_0 such that $\sigma_{ess}(\mathcal{L})\subset\{\lambda : \mathrm{Re}\lambda\le-\gamma_0\}$.*

Proof. Note that $\mathcal{L}=\mathcal{L}_0+\mathcal{L}_1+\mathcal{L}_2$. We know that $\sigma_{ess}(\mathcal{L}_0)\subset\{\lambda : \mathrm{Re}\lambda\le\max\{\bar f'(-1),\bar f'(1)\}\}$. If $D>0$ or $c_0\neq0$, by Lemma 3.5 and applying Theorem A.1 (p. 136) in [20], we deduce that the half-plane $\{\lambda\mid\mathrm{Re}\lambda>\max\{f'(-1),f'(1)\}\}$ consists entirely of normal points of $\mathcal{L}$ (see Lemma 3.7). This establishes the lemma with $\gamma_0=-\max\{\bar f'(-1),\bar f'(1)\}$.

For the case $D=0$ and $c_0=0$, we first show that $\lambda-\mathcal{L}$ is Fredholm with index 0 for $\lambda\in S_{\gamma_1}=\{\lambda\in\mathbb{C}:\mathrm{Re}\lambda\ge-\gamma_1\}$ for some constant $\gamma_1>0$ to be chosen. Let k_i be as defined in (3.5) and (3.6) for $i=0,1,\dots,n$. By assumption (A_3), $k_0(x)\le-\bar\delta_0<0$ for all $x\in\mathbb{R}$, where $\bar\delta_0$ is a positive constant. Since $\bar f'(\pm1)=k_0(\pm\infty)+\sum_{i=1}^n k_i(\pm\infty)<0$, by assumption (A_2), we can choose $\gamma_1>0$ such that $\lambda-k_0(x)\neq0$ for all $x\in\mathbb{R}$ and $|\sum_{i=1}^n\frac{k_i(\pm\infty)}{\lambda-k_0(\pm\infty)}|<1$ for all $\lambda\in S_{\gamma_1}$. For $\lambda\in S_{\gamma_1}$ and $i=1,2,\dots,n$, we choose smooth functions $m_i(x)$ on $\mathbb{R}$ satisfying $\lim_{x\to\pm\infty}m_i(x)=\frac{k_i(\pm\infty)}{\lambda-k_0(\pm\infty)}$ and $\sum_{i=1}^n|m_i(x)|\le\delta_1<1$, where δ_1 is a positive constant. Then, we can write

$$\begin{aligned}(\lambda-\mathcal{L})\psi&=(\lambda-k_0)\psi-\sum_{i=1}^n F_{p_i}(\phi,J_1*s^1(\phi),\dots,J_n*s^n(\phi))\cdot J_i*(s_u^i(\phi)\psi)\\&=(\lambda-k_0)\psi-\sum_{i=1}^n k_iJ_i*\psi+B_0\psi\\&=(\lambda-k_0)(B_1\psi+B_2\psi)+B_0\psi,\end{aligned}\tag{3.12}$$

where

$$\begin{aligned}B_0\psi&=\sum_{i=1}^n[k_iJ_i*\psi-F_{p_i}(\phi,J_1*s^1(\phi),\dots,J_n*s^n(\phi))J_i*(s_u^i(\phi)\psi)],\\B_1\psi&=\sum_{i=1}^n\left[m_i-\frac{k_i}{\lambda-k_0}\right]J_i*\psi,\\B_2\psi&=\psi-\sum_{i=1}^n m_iJ_i*\psi.\end{aligned}$$

By the choice of m_i, we know that the bounded linear operator B_2 is invertible on X_0. Therefore, the index of $\lambda-\mathcal{L}$ is 0 since B_0 and B_1 are compact operators on X_0.

This and the simplicity of eigenvalue 0 of $\mathcal{L}$ (see Lemma 3.8) imply that $X_0=\mathrm{Range}(\mathcal{L})\oplus N$, where N is one-dimensional. On the other hand, since 0 is an algebraically simple eigenvalue, we can take $N=\mathrm{Ker}(\mathcal{L})$ and hence $X_0=\mathrm{Range}(\mathcal{L})\oplus\mathrm{Ker}(\mathcal{L})$. Now $\mathcal{L}:\mathrm{Range}(\mathcal{L})\to\mathrm{Range}(\mathcal{L})$ is one-to-one and onto, so it has a bounded inverse. Therefore, for $|\lambda|$ small, $(\lambda-\mathcal{L})|_{\mathrm{Range}(\mathcal{L})}$ is invertible and hence if $\lambda\neq0$ with $|\lambda|$ small, $\mathcal{L}\psi=\lambda\psi$ has no nonzero solution in X_0. On the other hand, by Lemma 3.7, $\mathcal{L}\psi=\lambda\psi$ has no nonzero solution in X_0 for $\lambda\neq0$ with $\mathrm{Re}\lambda\ge0$. From this and

the fact that $\lambda - \mathcal{L}$ is Fredholm with index zero for $\lambda \in S_{\gamma_1}$, we deduce that λ is in the resolvent set of $\mathcal{L}$ for $\lambda \neq 0$ with $|\lambda|$ small or $\mathrm{Re}\lambda \geq 0$. Since $\sigma(\mathcal{L})$ is closed and bounded, we deduce that λ is in the resolvent set of $\mathcal{L}$ for $\lambda \neq 0$ with $\mathrm{Re}\lambda \geq -\gamma_0$ for some constant $\gamma_0 > 0$. □

From Lemma 3.6, we deduce that the half-plane $\{\lambda \mid \mathrm{Re}\lambda > -\gamma_0\}$ consists entirely of normal points of $\mathcal{L}$. This establishes (iii) of Theorem 2.2. The rest of the statement follows from the following lemmas.

LEMMA 3.7. $\{\lambda : \mathrm{Re}\lambda \geq 0, \lambda \neq 0\} \subset \rho(\mathcal{L})$.

Proof. It is clear that every eigenfunction of $\mathcal{L}$ in X_0 with nonzero eigenvalue is at least C^1. Suppose $\lambda = \alpha + i\beta$ satisfying $\alpha \geq 0$ and $\beta \neq 0$ is an eigenvalue with eigenfunction $u = u^1(x) + iu^2(x) \neq 0$. As in the proof of Lemma 3.2, we consider the Cauchy problem

$$v_t = \mathcal{L}v - \alpha v, \tag{3.13}$$

$$v(x,0) = u^1(x). \tag{3.14}$$

It has a solution $v(x,t) = u^1(x)\cos\beta t - u^2(x)\sin\beta t$.

Note that $v(x,t) \leq |u(x)|$ for all $x \in \mathbb{R}$ and $t \geq 0$. We claim that there is a $\tau > 0$ such that $v(x,t) \leq \tau\phi'(x)$ for all $x \in \mathbb{R}$ and $t \geq 0$. To prove this claim, let θ_0 be a constant satisfying $0 < \theta_0 < -\max\{\bar{f}'(-1), \bar{f}'(1)\}$. Since $\phi(\pm\infty) = \pm 1$, we may choose M large enough such that $F_u(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi))(x) + \sum_{i=1}^n F_{p_i}(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi)) J_i * (s_u^i(\phi))(x) \leq -\theta_0$ for all $|x| \geq M$ and such that $|u|$ is positive at some point $x \in [-M, M]$. Since $\phi'(x) > 0$, there exists a constant $\tau > 0$ such that $|u(x)| \leq \tau\phi'(x)$ for $|x| \leq M$. We prove that the claim holds with this choice of τ. Since $\lim_{x\to\pm\infty} u(x) = 0$, there exists a constant $\epsilon > 0$ such that $v(x,t) \leq \tau\phi'(x) + \epsilon$ for all x and $t \geq 0$. Let $\epsilon_0 = \inf\{\epsilon : v(x,t) \leq \tau\phi'(x) + \epsilon \text{ for all } x \in \mathbb{R} \text{ and } t \geq 0\}$. We prove that $\epsilon_0 = 0$. Consider the function $w(x,t) = \tau\phi'(x) + \epsilon_0 e^{-\theta_0 t}$. We have

$$\begin{aligned} w_t - \mathcal{L}w + \alpha w = {} & \epsilon_0(-\theta_0 - F_u(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi))(x) \\ & - \sum_{i=1}^n F_{p_i}(\phi, J_1 * s^1(\phi), \ldots, J_n * s^n(\phi)) J_i * (s_u^i(\phi))(x)) e^{-\theta_0 t} + \alpha w \\ \geq {} & 0 \end{aligned} \tag{3.15}$$

for all $|x| > M$ and $t > 0$. Therefore w is a supersolution of (3.13) on $|x| > M$. Notice that $w(x,t) \geq v(x,t)$ for $|x| \leq M$ and $t > 0$ and $w(x,0) \geq v(x,0)$ for all x. The comparison principle (Lemma 3.1) yields $w(x,t) \geq v(x,t)$ for all x and $t > 0$. Therefore $v(x,t) = v(x, t + 2n\pi/\beta) \leq \tau\phi'(x) + \epsilon_0 e^{-\theta_0(t+2n\pi/\beta)}$ for all $n \in Z^+$, $x \in \mathbb{R}$, and $t > 0$. Letting $n \to \infty$, we get $v(x,t) \leq \tau\phi'(x)$ for all x. Therefore $\epsilon_0 = 0$ and the claim is proved.

Clearly τ_0 can be chosen such that $|u(x)| \leq \tau_0\phi'(x)$ for all $|x| \leq M$ and there is a point $x_0 \in [-M, M]$ such that $|u(x_0)| = \tau_0\phi'(x_0)$. By the comparison principle, we deduce that $v(x,t) < \tau_0\phi'(x)$ for all x and $t > 0$. If we choose t such that $u(x_0)/|u(x_0)| = e^{-i\beta t}$, then $v(x_0,t) = |u(x_0)| = \tau_0\phi'(x_0) > v(x_0,t)$, which is a contradiction. Therefore $u(x) = 0$ for all x and λ is not an eigenvalue.

Now assume that $\lambda > 0$ is an eigenvalue with an eigenfunction $u(x)$. Without loss of generality, we may assume u is real and there is a point where u is positive. Then $v(x,t) \equiv u(x)$ is a solution of (3.13) with $\alpha = \lambda > 0$ and the initial condition $v(x,0) = u(x)$. A similar argument to the above can be used to prove that $u(x) \equiv 0$ and so $\lambda > 0$ is not an eigenvalue. This completes the proof. □

LEMMA 3.8. *0 is an eigenvalue of $\mathcal{L}$ and it is algebraically simple.*

Proof. Since $(\phi(\cdot + s), c_0)$ satisfies (1.4) for each $s \in \mathbb{R}$ and $\lim_{z\to\infty} \phi(\pm z) = \pm 1$, $p = \phi'$ is a positive eigenfunction of $\mathcal{L}$ in X_0 with corresponding eigenvalue 0.

To show that 0 is a geometrically simple eigenvalue we use the comparison principle. Suppose that $\mathcal{L}v = 0$ with $v \in X_0$ and assume, without loss of generality, that $v(x_0) > 0$ for some x_0. For $\theta > 0$, let $w_\theta \equiv \theta\phi' - v$. Note that $\phi'(x) > 0$. We can choose θ large enough such that $w_\theta(x) \geq 0$ for $|x| \leq M$, where M is a constant given as in the proof of Lemma 3.7. Note that $v_\theta(x,t) \equiv w_\theta(x)$ as a function of (x,t) (it is independent of t) satisfies (3.13) with initial data $v(x,0) = w_\theta(x)$. Similar to the proof in Lemma 3.7, we deduce that $w_\theta(x) \geq 0$ for all $x \in \mathbb{R}$. Let $\theta_0 = \inf\{\theta : w_\theta(x) \geq 0\}$. We claim that $w_{\theta_0}(x) \equiv 0$. In fact, if w_{θ_0} is not identically zero, we apply the comparison principle to $v_{\theta_0}(x,t) \equiv w_{\theta_0}(x)$ to deduce that $w_{\theta_0} > 0$ for all $x \in \mathbb{R}$. Therefore there exists $\epsilon_0 > 0$ such that $w_{\theta_0-\epsilon_0}(x) \geq 0$ for $x \in [-M, M]$. This contradicts the choice of θ_0. Therefore, ϕ is a simple eigenfunction corresponding to eigenvalue 0.

To show algebraic simplicity of eigenvalue 0, suppose that there is v such that $\mathcal{L}v = -\phi'$ and recall that $\phi'(x) > 0$ on $\mathbb{R}$. Let $w_\tau \equiv v + \tau\phi'$ and note that for $\tau > 0$ and sufficiently large, $w_\tau > 0$ on $[-M, M]$ and, as we argued above, $w_\tau > 0$ on $\mathbb{R}$. Taking the infimum T of all such τ produces a function $w_T \geq 0$ on $\mathbb{R}$ and a point x_T where this function is zero. Then

$$\begin{aligned} 0 &\leq Dw_T''(x_T) + \sum_{i=1}^{n} F_{p_i}(\phi, J_1 * s^1(\phi), \cdots, J_n * s^n(\phi)) \cdot J_i * (s_u^i(\phi) w_T)(x_T) \\ &= \mathcal{L}w_T(x_T) = -\phi_0'(x_T) < 0. \end{aligned}$$

This provides the contradiction which establishes the result. ∎

Let $\mathcal{L}^*$ be the adjoint operator of $\mathcal{L}$ on X_0^*. Then we have the next lemma.

LEMMA 3.9. *The operator $\mathcal{L}^*$ on X_0^* has a positive eigenvector corresponding to the simple eigenvalue* 0.

Proof. Since 0 is a simple eigenvalue of $\mathcal{L}$, by the Fredholm theory, $\mathcal{L}^*v = 0$ has a nonzero solution v and it is unique modulo a constant multiple. For any solution $v \neq 0$ of $\mathcal{L}^*v = 0$ in X_0^*, v or $-v$ must be a positive Borel measure. To show this, if v or $-v$ is not positive, we can find a function $f \in X_0$ such that $f(x) < 0$ for all x and the dual pairing $< f, v >= 0$. Then, by the Fredholm Alternative, $\mathcal{L}u = f$ has a solution u. As before, we can choose τ such that $\tau\phi'(x) \geq -u(x)$ and $\tau\phi'(x_0) = -u(x_0)$ for some x_0. With $w(x,t) = \tau\phi'(x_0) + u(x_0)$, we have $w_t - \mathcal{L}w = -\mathcal{L}u = -f > 0$ and $w(x_0, t) = 0$, in contradiction to the strong comparison principle. ∎

Acknowledgment. The authors thank the referees for valuable suggestions and comments.

REFERENCES

[1] P. W. BATES AND F. CHEN, *Periodic traveling wave solutions of an integrodifferential model for phase transition*, Electron. J. Differential Equations, 26 (1999), pp. 1–19.

[2] P. W. BATES, F. CHEN, AND J. WANG, *Global existence and uniqueness of solutions to a nonlocal phase-field system*, in US-Chinese Conference on Differential Equations and Applications, P. W. Bates, S.-N. Chow, K. Lu and X. Pan, eds., International Press, Cambridge, MA, 1997, pp. 14–21.

[3] P. W. BATES, X. CHEN, AND A. CHMAJ, *Equilibria and traveling waves for bistable equations with non-local and discrete dissipation*, in Nonlinear Diffusive Systems—Dynamics and Asymptotics, E. Yanagida and Y. Morita, eds., RIMS, Kyoto, Japan, 2001, pp. 48–71.

[4] P. W. Bates and A. Chmaj, *An integrodifferential model for phase transitions: Stationary solutions in higher space dimensions*, J. Statist. Phys., 95 (1999), pp. 1119–1139.
[5] P. W. Bates, P. C. Fife, R. A. Gardner, and C. K. R. T. Jones, *The existence of traveling wave solutions of a generalized phase-field model*, SIAM J. Math. Anal., 28 (1997), pp. 60–93.
[6] P. W. Bates, P. C. Fife, X. Ren, and X. Wang, *Traveling waves in a nonlocal model of phase transitions*, Arch. Ration. Mech. Anal. 138 (1997), pp. 105–136.
[7] P. W. Bates and C. K. R. T. Jones, *Invariant Manifolds for Semilinear Partial Differential Equations*, in Dynam. Report. Ser. Dynam. Systems Appl. 2, Wiley, Chichester, UK, 1989, pp. 1–38.
[8] P. W. Bates, K. Lu, and C. Zeng, *Existence and persistence of invariant manifolds for semiflows in banach space*, Mem. Amer. Math. Soc., 135 (1998).
[9] P. W. Bates, K. Lu, and C. Zeng, *Invariant foliations for semiflows near a normally hyperbolic invariant manifold*, Trans. Amer. Math. Soc., 352 (2000), pp. 4641–4676.
[10] P. W. Bates, K. Lu, and C. Zeng, *Persistence of overflowing manifolds for semiflow*, Comm. Pure Appl. Math., 52 (1999), pp. 983–1046.
[11] F. Chen, *Almost periodic traveling waves of nonlocal evolution equations*, Nonlinear Anal., 50 (2002), pp. 807–838.
[12] X. Chen, *Existence, uniqueness and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.
[13] Z. Chen, B. Ermentrout, and B. McLeod, *Traveling fronts for a class of non-local convolution differential equations*. Appl. Anal., 64 (1997), pp. 235–253.
[14] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.
[15] A. de Masi, E. Orlandi, E. Presutti, and L. Triolo, *Stability of the interface in a model of phase separation*, Proc. Roy. Soc. Edinburgh, 124 (1994), pp. 1013–1022.
[16] B. Ermentrout and J. B. Mcleod, *Existence and uniqueness of traveling waves for a neural network*, Proc. Roy. Soc. Edinburgh, 123 (1993), pp. 461–478.
[17] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.
[18] P. C. Fife and X. Wang, *A convolution model for interfacial motion: The generation and propagation of internal layers in higher space dimensions*, Adv. Differential Equations, 3 (1998), pp. 85–110.
[19] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.
[20] D. Henry, *Geometric theory of semilinear parabolic equations*, in Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 127–142

LOCALIZATION FOR THE SCHRÖDINGER EQUATION IN A LOCALLY PERIODIC MEDIUM*

GRÉGOIRE ALLAIRE† AND MARIAPIA PALOMBARO‡

Abstract. We study the homogenization of a Schrödinger equation in a locally periodic medium. For the time and space scaling of semiclassical analysis we consider well-prepared initial data that are concentrated near a stationary point (with respect to both space and phase) of the energy, i.e., the Bloch cell eigenvalue. We show that there exists a localized solution which is asymptotically given as the product of a Bloch wave and of the solution of an homogenized Schrödinger equation with quadratic potential.

Key words. homogenization, localization, Bloch waves, Schrödinger

AMS subject classifications. 35B27, 35J10

DOI. 10.1137/050635572

1. Introduction. We study the homogenization of the following Schrödinger equation:

$$\begin{cases} \dfrac{i}{\varepsilon}\dfrac{\partial u_\varepsilon}{\partial t} - \operatorname{div}\left(A\left(x,\dfrac{x}{\varepsilon}\right)\nabla u_\varepsilon\right) + \dfrac{1}{\varepsilon^2}c\left(x,\dfrac{x}{\varepsilon}\right)u_\varepsilon = 0 & \text{in } \mathbb{R}^N\times\mathbb{R}^+, \\ u_\varepsilon(0,x) = u_\varepsilon^0(x) & \text{in } \mathbb{R}^N, \end{cases} \tag{1.1}$$

where the unknown $u_\varepsilon(t,x)$ is a complex-valued function. The coefficients $A(x,y)$ and $c(x,y)$ are real and sufficiently smooth bounded functions defined for $x\in\mathbb{R}^N$ (the macroscopic variable) and $y\in\mathbb{T}^N$ (the microscopic variable in the unit torus). The period ε is a small positive parameter which is intended to go to zero. Furthermore, the matrix A is symmetric uniformly positive definite. Of course, the usual Schrödinger equation is recovered when $A\equiv Id$, but since there is no additional difficulty, we keep the general form of (1.1) in what follows (which can be interpreted as introducing a nonflat locally periodic metric).

The scaling of (1.1) is that of semiclassical analysis (see, e.g., [5], [8], [10], [11], [12], [13], [14], [18], [19]): if the period is rescaled to 1, it amounts to looking at large time and space variables of order ε^{-1}. At least in the case when $A\equiv Id$ and $c(x,y)=c_0(x)+c_1(y)$, there is a well-known theory for the asymptotic limit of (1.1) when ε goes to zero. By using WKB asymptotic expansion or the notion of semiclassical measures (or Wigner transforms), the homogenized problem is in some sense the Liouville transport equation for a classical particle which is the limit of the wave function u_ε. In other words, for initial data living in the nth Bloch band and under some technical assumptions on the Bloch spectral cell problem (1.4), the semiclassical limit of (1.1) is given by the dynamic of the following Hamiltonian system

*Received by the editors July 8, 2005; accepted for publication November 14, 2005; published electronically April 12, 2006. This work was partly supported by the Research Training Network MULTIMAT funded by the EEC.

http://www.siam.org/journals/sima/38-1/63557.html

†Centre de Mathématiques Appliquées, École Polytechnique, 91128 Palaiseau, France (gregoire.allaire@polytechnique.fr).

‡Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany (Mariapia.Palombaro@mis.mpg.de).

in the phase space $(x,\theta)\in\mathbb{R}^N\times\mathbb{T}^N$:

$$\begin{cases}\dot{x}=\nabla_\theta\lambda_n(x,\theta),\\ \dot{\theta}=-\nabla_x\lambda_n(x,\theta),\end{cases} \tag{1.2}$$

where the Hamiltonian $\lambda_n(x,\theta)$ is precisely the nth Bloch eigenvalue of (1.4) (see [8], [10], [11], [12], [13], [14], [18], [19] for more details).

Our approach to (1.1) is different since we consider special initial data that are monochromatic and have zero group velocity and zero applied force. Namely, the initial data are concentrated at a point (x^n,θ^n) of the phase space, where $\nabla_\theta\lambda_n(x^n,\theta^n)=\nabla_x\lambda_n(x^n,\theta^n)=0$. In such a case, the previous Hamiltonian system (1.2) degenerates (its solution is constant) and is unable to describe the precise dynamic of the wave function u_ε. We exhibit another limit problem which is again a Schrödinger equation with quadratic potential. In other words, we build a sequence of approximate solutions of (1.1) which are the product of a Bloch wave and of the solution of a homogenized Schrödinger equation. Furthermore, if the full Hessian tensor of the Bloch eigenvalue $\lambda_n(x,\theta)$ is positive definite at (x^n,θ^n), we prove that all the eigenfunctions of a homogenized Schrödinger equation are exponentially decreasing at infinity. In other words, we exhibit a localization phenomenon for (1.1) since we build a sequence of approximate solutions that decay exponentially fast away from x^n. The root of this localization phenomenon is the macroscopic modulation (i.e., with respect to x) of the periodic coefficients, which is similar in spirit to the randomness that causes Anderson's localization (see [9] and references therein).

Let us describe more precisely the type of well-prepared initial data that we consider. For a given point $(x^n,\theta^n)\in\mathbb{R}^N\times\mathbb{T}^N$ and a given function $v^0\in H^1(\mathbb{R}^N)$ we take

$$u_\varepsilon^0(x)=\psi_n\Big(x^n,\frac{x}{\varepsilon},\theta^n\Big)e^{2i\pi\frac{\theta^n\cdot x}{\varepsilon}}v^0\Big(\frac{x-x^n}{\sqrt{\varepsilon}}\Big), \tag{1.3}$$

where $\psi_n(x,y,\theta)$ is a so-called Bloch eigenfunction, solution of the Bloch spectral cell equation

$$-(\mathrm{div}_y+2i\pi\theta)(A(x,y)(\nabla_y+2i\pi\theta)\psi_n)+c(x,y)=\lambda_n(x,\theta)\psi_n \qquad \text{in } \mathbb{T}^N, \tag{1.4}$$

corresponding to the nth eigenvalue or energy level λ_n. The Bloch wave ψ_n is periodic with respect to y, but v^0 is not periodic, so $v^0\big(\frac{x-x^n}{\sqrt{\varepsilon}}\big)$ means that the initial data is concentrated around x^n with a support of asymptotic size $\sqrt{\varepsilon}$. The Bloch frequency $\theta^n\in\mathbb{T}^N$, the localization point $x^n\in\mathbb{R}^N$, and the energy level n are chosen such that $\lambda_n(x^n,\theta^n)$ is simple and $\nabla_x\lambda_n(x^n,\theta^n)=\nabla_\theta\lambda_n(x^n,\theta^n)=0$.

Our main result (Theorem 3.2) shows that the solution of (1.1) is approximately given by

$$u_\varepsilon(t,x)\approx\psi_n\Big(x^n,\frac{x}{\varepsilon},\theta^n\Big)e^{i\frac{\lambda_n(x^n,\theta^n)t}{\varepsilon}}e^{2i\pi\frac{\theta^n\cdot x}{\varepsilon}}v\Big(t,\frac{x-x^n}{\sqrt{\varepsilon}}\Big), \tag{1.5}$$

where v is the unique solution of the homogenized Schrödinger equation

$$\begin{cases} i\dfrac{\partial v}{\partial t}-\mathrm{div}\,(A^*\nabla v)+\mathrm{div}(vB^*z)+c^*v+vD^*z\cdot z=0 & \text{in } \mathbb{R}^N\times\mathbb{R}^+,\\ v(0,z)=v^0(z) & \text{in } \mathbb{R}^N,\end{cases} \tag{1.6}$$

where c^* is a constant coefficient and A^*, B^*, D^* are constant matrices defined by

$$A^* = \frac{1}{8\pi^2}\nabla_\theta\nabla_\theta\lambda_n(x^n,\theta^n), \quad B^* = \frac{1}{2i\pi}\nabla_\theta\nabla_x\lambda_n(x^n,\theta^n), \quad D^* = \frac{1}{2}\nabla_x\nabla_x\lambda_n(x^n,\theta^n).$$

In Proposition 3.4 we show that the homogenized problem (1.6) is well-posed since the underlying operator is self-adjoint. Furthermore, under the additional assumption that the Hessian tensor $\nabla\nabla\lambda_n(x^n,\theta^n)$ (with respect to both variables x and θ) is positive definite, we prove that (1.6) admits a countable number of eigenvalues and eigenfunctions which all decay exponentially at infinity (see Proposition 3.5). In such a case, formula (1.5) defines a family of approximate (exponentially) localized solutions of (1.1).

Let us indicate that the case of the first eigenvalue (ground state) $n = 1$ with $\theta^1 = 0$ was already studied in [3] (for the spectral problem rather than the evolution equation). The case of purely periodic coefficients (i.e., that depend only on y and not on x) is completely different and was studied in [4]. Indeed, in this latter case there is no localization effect, and one proves that, for a longer time scale (of order ε^{-1} with respect to (1.1)), the homogenized limit is again a Schrödinger equation without the drift and quadratic potential in (1.6).

2. Preliminaries. In the present section we give our main assumptions and set some notation and a few preliminary results needed in the proof of the main results in section 3.

We first assume that the coefficients $A_{ij}(x,y)$ and $c(x,y)$ are real, bounded, and Carathéodory functions (measurable with respect to y and continuous in x), which are periodic with respect to y. In other words, they belong to $C_b(\mathbb{R}^N; L^\infty(\mathbb{T}^N))$. Furthermore, the tensor $A(x,y)$ is symmetric uniformly coercive. Under these assumptions, it is well known that, for any values of the parameters $\theta \in \mathbb{T}^N$ and $x \in \mathbb{R}^N$, the cell problem (1.4) defines a compact self-adjoint operator on $L^2(\mathbb{T}^N)$, which admits a countable sequence of real increasing eigenvalues $\{\lambda_n(x,\theta)\}_{n\geq 1}$ (repeated with their multiplicity) with corresponding eigenfunctions $\{\psi_n(x,\theta,y)\}_{n\geq 1}$ normalized by

$$\|\psi_n(x,\theta,\cdot)\|_{L^2(\mathbb{T}^N)} = 1.$$

Our main assumptions are the following.

Hypothesis H1. There exist $x^n \in \mathbb{R}^N$ and $\theta^n \in \mathbb{T}^N$ such that

$$\text{(2.1)} \qquad \begin{cases} \text{(i) } \lambda_n(x^n,\theta^n) \text{ is a simple eigenvalue,} \\ \text{(ii) } (x^n,\theta^n) \text{ is a critical point of } \lambda_n(x,\theta), \\ \qquad \text{i.e., } \nabla_x\lambda_n(x^n,\theta^n) = \nabla_\theta\lambda_n(x^n,\theta^n) = 0. \end{cases}$$

Hypothesis H2. The coefficients $A(x,y)$ and $c(x,y)$ are of class C^2 with respect to the variable x in a neighborhood of $x = x^n$.

Then we set

$$A_{1,h}(y) := \frac{\partial A}{\partial x_h}(x^n,y), \quad A_{2,lh}(y) := \frac{\partial^2 A}{\partial x_l \partial x_h}(x^n,y), \quad \text{for } l,h = 1,\dots,N.$$

Similar notation is used to denote the derivatives of the function c with respect to the x-variable. With an abuse of notation we further set

$$A(y) := A(x^n,y), \quad \lambda_n := \lambda_n(x^n,\theta^n), \quad \psi_n(y) := \psi_n(x^n,y,\theta^n),$$

and analogous notation holds for all derivatives of ψ_n and λ_n with respect to the x-variable and the θ-variable evaluated at $x = x^n$ and $\theta = \theta^n$. Without loss of generality we will assume in what follows that $x^n = 0$.

Notation. For any function $\rho(y)$ defined on $\mathbb{T}^N$ we set

$$\rho^\varepsilon(z) := \rho(z/\sqrt{\varepsilon}),$$

where $z := \sqrt{\varepsilon}y \equiv x/\sqrt{\varepsilon}$. In what follows the symbols div_y and ∇_y will stand for the divergence and gradient operators which act with respect to the y-variable, while div and ∇ will indicate the divergence and gradient operators which act with respect to the z-variable. Finally, throughout this paper the Einstein summation convention is used.

Under Hypothesis H1-(i) it is a classical matter to prove that the nth eigencouple of (1.4) is smooth with respect to the variable θ in a neighborhood of $\theta = \theta^n$ (see [16]) and has the same differentiability property as the coefficients with respect to the variable x. Introducing the unbounded operator $\mathbb{A}_n(x,\theta)$ defined on $L^2(\mathbb{T}^N)$ by

$$\mathbb{A}_n(x,\theta)\psi = -(\mathrm{div}_y + 2i\pi\theta)\big(A(x,y)(\nabla_y + 2i\pi\theta)\psi\big) + c(x,y)\psi - \lambda_n(x,\theta)\psi,$$

it is easy to differentiate (1.4). Denoting by $(e_k)_{1\le k\le N}$ the canonical basis of $\mathbb{R}^N$, the first derivatives satisfy

$$\begin{aligned}\mathbb{A}_n(x,\theta)\frac{\partial\psi_n}{\partial\theta_k} &= 2i\pi e_k A(x,y)(\nabla_y + 2i\pi\theta)\psi_n \\ &\quad + (\mathrm{div}_y + 2i\pi\theta)\,(A(x,y)2i\pi e_k\psi_n) + \frac{\partial\lambda_n}{\partial\theta_k}(x,\theta)\psi_n,\end{aligned} \tag{2.2}$$

$$\begin{aligned}\mathbb{A}_n(x,\theta)\frac{\partial\psi_n}{\partial x_l} &= (\mathrm{div}_y + 2i\pi\theta)\left(\frac{\partial A}{\partial x_l}(x,\theta)(\nabla_y + 2i\pi\theta)\psi_n\right) \\ &\quad - \frac{\partial c}{\partial x_l}(x,y)\psi_n + \frac{\partial\lambda_n}{\partial x_l}(x,\theta)\psi_n.\end{aligned} \tag{2.3}$$

Similar formulas hold for second order derivatives. By integrating the cell equations for the second order derivatives against ψ_n, we obtain the following formulas that will be useful in what follows (their proofs are safely left to the reader).

LEMMA 2.1. *Assume that assumptions* H1 *and* H2 *hold true. Then the following equalities hold:*

$$\begin{aligned}&\int_{\mathbb{T}^N}\frac{1}{2\pi i}\left[A_{1,h}(\nabla_y + 2i\pi\theta^n)\frac{\partial\psi_n}{\partial\theta_k}\cdot(\nabla_y - 2i\pi\theta^n)\bar\psi_n + c_{1,h}\,\frac{\partial\psi_n}{\partial\theta_k}\bar\psi_n\right]dy \\ &\quad + \int_{\mathbb{T}^N}\left[A_{1,h}e_k\psi_n\cdot(\nabla_y - 2i\pi\theta^n)\bar\psi_n + Ae_k\frac{\partial\psi_n}{\partial x_h}\cdot(\nabla_y - 2i\pi\theta^n)\bar\psi_n\right]dy \\ &\quad - \int_{\mathbb{T}^N}\left[e_k\bar\psi_n A_{1,h}\cdot(\nabla_y + 2i\pi\theta^n)\psi_n + e_k\bar\psi_n A\cdot(\nabla_y + 2i\pi\theta^n)\frac{\partial\psi_n}{\partial x_h}\right]dy \\ &\quad - \frac{1}{2i\pi}\frac{\partial^2\lambda_n}{\partial x_h\partial\theta_k} = 0,\end{aligned} \tag{2.4}$$

$$
\begin{aligned}
(2.5)\qquad &\int_{\mathbb{T}^N}\left[A_{2,lh}(\nabla_y+2i\pi\theta^n)\psi_n\cdot(\nabla_y-2i\pi\theta^n)\bar\psi_n+\left(c_{2,lh}-\frac{\partial^2\lambda_n}{\partial x_l\partial x_h}\right)|\psi_n|^2\right]dy\\
&+\int_{\mathbb{T}^N}\left[A_{1,h}(\nabla_y+2i\pi\theta^n)\frac{\partial\psi_n}{\partial x_l}\cdot(\nabla_y-2i\pi\theta^n)\bar\psi_n+c_{1,h}\,\frac{\partial\psi_n}{\partial x_l}\bar\psi_n\right]dy\\
&+\int_{\mathbb{T}^N}\left[A_{1,l}(\nabla_y+2i\pi\theta^n)\frac{\partial\psi_n}{\partial x_h}\cdot(\nabla_y-2i\pi\theta^n)\bar\psi_n+c_{1,l}\,\frac{\partial\psi_n}{\partial x_h}\bar\psi_n\right]dy=0,
\end{aligned}
$$

(2.6)

$$
\begin{aligned}
&\int_{\mathbb{T}^N}\left[2i\pi e_kA(y)(\nabla_y+2i\pi\theta^n)\frac{\partial\psi_n}{\partial\theta_l}\bar\psi_n-\left(A(y)2i\pi e_k\frac{\partial\psi_n}{\partial\theta_l}\right)(\nabla_y-2i\pi\theta^n)\bar\psi_n\right]dy\\
&+\int_{\mathbb{T}^N}\left[2i\pi e_lA(y)(\nabla_y+2i\pi\theta^n)\frac{\partial\psi_n}{\partial\theta_k}\bar\psi_n-\left(A(y)2i\pi e_l\frac{\partial\psi_n}{\partial\theta_k}\right)(\nabla_y-2i\pi\theta^n)\bar\psi_n\right]dy\\
&-\int_{\mathbb{T}^N}\left[4\pi^2e_kA(y)e_l|\psi_n|^2+4\pi^2e_lA(y)e_k|\psi_n|^2\right]dy+\frac{\partial^2\lambda_n}{\partial\theta_l\partial\theta_k}(\theta^n)=0.
\end{aligned}
$$

We now give the variational formulations of the above cell problems, rescaled at size ε.

LEMMA 2.2. *Assume that assumptions* H1 *and* H2 *hold true, and let $\varphi(z)$ be a smooth compactly supported function defined from $\mathbb{R}^N$ into $\mathbb{C}$. Then the following equalities hold:*

$$
(2.7)\qquad \int_{\mathbb{R}^N}\left[A^\varepsilon(\nabla_y+2i\pi\theta^n)\psi_n^\varepsilon\cdot(\sqrt{\varepsilon}\nabla-2i\pi\theta^n)\bar\varphi(z)+(c^\varepsilon-\lambda_n^\varepsilon)\psi_n^\varepsilon\bar\varphi\right]dz=0,
$$

(2.8)

$$
\begin{aligned}
&\int_{\mathbb{R}^N}\left[A^\varepsilon(\nabla_y+2i\pi\theta^n)\frac{\partial\psi_n^\varepsilon}{\partial\theta_k^n}\cdot(\sqrt{\varepsilon}\nabla-2i\pi\theta^n)\bar\varphi+(c^\varepsilon-\lambda_n^\varepsilon)\frac{\partial\psi_n^\varepsilon}{\partial\theta_k^n}\bar\varphi\right]dz\\
&+\int_{\mathbb{R}^N}\left[-2\pi ie_k\cdot A^\varepsilon(\nabla_y+2i\pi\theta^n)\psi_n^\varepsilon\bar\varphi+A^\varepsilon\,2\pi ie_k\psi_n^\varepsilon\cdot(\sqrt{\varepsilon}\nabla-2i\pi\theta^n)\bar\varphi\right]dz=0,
\end{aligned}
$$

$$
\begin{aligned}
(2.9)\qquad &\int_{\mathbb{R}^N}\left[A^\varepsilon(\nabla_y+2i\pi\theta^n)\frac{\partial\psi_n^\varepsilon}{\partial x_h}\cdot(\sqrt{\varepsilon}\nabla-2i\pi\theta^n)\bar\varphi+(c^\varepsilon-\lambda_n^\varepsilon)\frac{\partial\psi_n^\varepsilon}{\partial x_h}\bar\varphi\right]dz\\
&+\int_{\mathbb{R}^N}\left[A_{1,h}^\varepsilon(\nabla_y+2i\pi\theta^n)\psi_n^\varepsilon\cdot(\sqrt{\varepsilon}\nabla-2i\pi\theta^n)\bar\varphi+c_{1,h}^\varepsilon\,\psi_n^\varepsilon\bar\varphi\right]dz=0.
\end{aligned}
$$

Proof. Formula (2.7) follows straightforwardly from (1.4), while (2.8)–(2.9) are consequences of (2.2)–(2.3). □

Finally, we recall the notion of two-scale convergence introduced in [1], [17] (which will be used with $\delta=\sqrt{\varepsilon}$).

PROPOSITION 2.3. *Let f_δ be a sequence uniformly bounded in $L^2(\mathbb{R}^N)$.*

(1) *There exists a subsequence, still denoted by f_δ, and a limit $f_0(x,y)\in L^2(\mathbb{R}^N\times\mathbb{T}^N)$ such that f_δ* two-scale converges *weakly to f_0 in the sense that*

$$
\lim_{\delta\to0}\int_{\mathbb{R}^N}f_\delta(x)\phi(x,x/\delta)\,dx=\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}f_0(x,y)\phi(x,y)\,dx\,dy
$$

for all functions $\phi(x,y)\in L^2(\mathbb{R}^N;C(\mathbb{T}^N))$.

(2) *Assume further that f_δ two-scale converges weakly to f_0 and that*

$$\lim_{\delta\to 0}\|f_\delta\|_{L^2(\mathbb{R}^N)}=\|f_0\|_{L^2(\mathbb{R}^N\times\mathbb{T}^N)}.$$

Then f_δ is said to two-scale converge strongly *to its limit f_0 in the sense that, if f_0 is smooth enough, e.g., $f_0\in L^2(\mathbb{R}^N;C(\mathbb{T}^N))$, we have*

$$\lim_{\delta\to 0}\int_{\mathbb{R}^N}|f_\delta(x)-f_0(\,x,x/\delta\,)|^2\,dx=0.$$

(3) *Assume that $\delta\nabla f_\delta$ is also uniformly bounded in $L^2(\mathbb{R}^N)^N$. Then there exists a subsequence, still denoted by f_δ, and a limit $f_0(x,y)\in L^2(\mathbb{R}^N;H^1(\mathbb{T}^N))$ such that f_δ two-scale converges to $f_0(x,y)$ and $\delta\nabla f_\delta$ two-scale converges to $\nabla_y f_0(x,y)$.*

3. Main results. We begin by recalling the usual a priori estimates for the solution of the Schrödinger equation (1.1), which hold true since the coefficients are real. They are obtained by multiplying the equation successively by $\overline{u}_\varepsilon$ and $\frac{\partial\overline{u}_\varepsilon}{\partial t}$ and integrating by parts.

LEMMA 3.1. *There exists $C>0$ independent of ε such that the solution of* (1.1) *satisfies*

$$\begin{aligned}\|u_\varepsilon\|_{L^\infty(\mathbb{R}^+;L^2(\mathbb{R}^N))}&=\|u_\varepsilon^0\|_{L^2(\mathbb{R}^N)},\\ \varepsilon\|\nabla u_\varepsilon\|_{L^\infty(\mathbb{R}^+;L^2(\mathbb{R}^N))}&\le C\big(\|u_\varepsilon^0\|_{L^2(\mathbb{R}^N)}+\varepsilon\|\nabla u_\varepsilon^0\|_{L^2(\mathbb{R}^N)}\big).\end{aligned}$$

THEOREM 3.2. *Assume that assumptions* H1 *and* H2 *hold true and that the initial datum u_ε^0 is of the form* (1.3). *Then the solution of* (1.1) *can be written as*

$$u_\varepsilon(t,x)=e^{i\frac{\lambda_n t}{\varepsilon}}e^{2i\pi\frac{\theta^n\cdot x}{\varepsilon}}v_\varepsilon\Big(t,\frac{x-x^n}{\sqrt{\varepsilon}}\Big),\tag{3.1}$$

where $v_\varepsilon(t,z)$ two-scale converges strongly to $\psi_n(y)v(t,z)$; i.e.,

$$\lim_{\varepsilon\to 0}\int_{\mathbb{R}^N}\left|v_\varepsilon(t,z)-\psi_n\left(\frac{z}{\sqrt{\varepsilon}}\right)v(t,z)\right|^2dz=0,\tag{3.2}$$

uniformly on compact time intervals in $\mathbb{R}^+$, and v is the unique solution of the homogenized Schrödinger equation

$$\begin{cases}i\dfrac{\partial v}{\partial t}-\operatorname{div}(A^*\nabla v)+\operatorname{div}(vB^*z)+c^*v+vD^*z\cdot z=0 & \text{in } \mathbb{R}^N\times\mathbb{R}^+,\\ v(0,z)=v^0(z) & \text{in } \mathbb{R}^N,\end{cases}\tag{3.3}$$

where

$$A^*=\frac{1}{8\pi^2}\nabla_\theta\nabla_\theta\lambda_n(x^n,\theta^n),\quad B^*=\frac{1}{2i\pi}\nabla_\theta\nabla_x\lambda_n(x^n,\theta^n),\quad D^*=\frac{1}{2}\nabla_x\nabla_x\lambda_n(x^n,\theta^n),$$

and c^* is given by

$$\begin{aligned}c^*=\int_{\mathbb{T}^N}\Big[&A\,(\nabla_y+2i\pi\theta^n)\psi_n\cdot\frac{\partial\bar\psi_n}{\partial x_k}e_k\\ &-A(\nabla_y-2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial x_k}\cdot\psi_n\,e_k-A_{1,k}(\nabla_y-2i\pi\theta^n)\bar\psi_n\cdot\psi_n e_k\Big]dy.\end{aligned}$$

Remark 3.3. Notice that even if the tensor A^* might be noncoercive, the homogenized problem (3.3) is well posed. Indeed, the operator $\mathbb{A}^* : L^2(\mathbb{R}^N) \to L^2(\mathbb{R}^N)$ defined by

$$\mathbb{A}^*\varphi = -\mathrm{div}\,(A^*\nabla\varphi) + \mathrm{div}(\varphi B^* z) + c^*\varphi + \varphi D^* z \cdot z \tag{3.4}$$

is self-adjoint (see Proposition 3.4), and therefore by using semigroup theory (see, e.g., [6] or Chapter X in [20]), one can show that there exists a unique solution in $C(\mathbb{R}^+; L^2(\mathbb{R}^N))$, although it may not belong to $L^2(\mathbb{R}^+; H^1(\mathbb{R}^N))$.

The next result establishes the conservation of the L^2-norm for the solution v of the homogenized equation (3.3) and the self-adjointness of the operator $\mathbb{A}^*$.

PROPOSITION 3.4. *Let $v \in C(\mathbb{R}^+; L^2(\mathbb{R}^N))$ be solution to* (3.3). *Then*

$$\|v(t,\cdot)\|_{L^2(\mathbb{R}^N)} = \|v^0\|_{L^2(\mathbb{R}^N)} \quad \forall\, t \in \mathbb{R}^+. \tag{3.5}$$

Moreover, the operator $\mathbb{A}^*$ defined in (3.4) *is self-adjoint.*

Proof. We multiply (3.3) by $\bar{v}$ and take the imaginary part to obtain

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^N} |v|^2\,dz = \mathrm{Im}\left(\int_{\mathbb{R}^N} vB^* z\cdot\nabla\bar{v} - c^*|v|^2\,dz\right). \tag{3.6}$$

After integrating by parts one finds that the right-hand side of (3.6) equals

$$-\left(\frac{1}{2i}\mathrm{tr}\,B^* + \mathrm{Im}c^*\right)\int_{\mathbb{R}^N}|v|^2\,dz,$$

and therefore (3.5) is proved as soon as we show that

$$\frac{1}{2i}\mathrm{tr}\,B^* + \mathrm{Im}c^* = 0. \tag{3.7}$$

In order to do this we first rewrite the coefficients c^* and B^* in a suitable form. Denoting by $\langle\cdot,\cdot\rangle$ the Hermitian inner product in $L^2(\mathbb{T}^N)$ and using (2.2), we write

$$c^* = \frac{1}{2i\pi}\left\langle \mathbb{A}_n\frac{\partial\psi_n}{\partial\theta_k}, \frac{\partial\psi_n}{\partial x_k}\right\rangle - \int_{\mathbb{T}^N} A_{1,k}(\nabla_y - 2i\pi\theta^n)\bar{\psi}_n\cdot\psi_n e_k\,dy, \tag{3.8}$$

while by (2.2)–(2.4) it follows that

$$\begin{aligned}\frac{1}{2i\pi}\frac{\partial^2\lambda_n}{\partial x_h\partial\theta_k} =& -\frac{1}{2i\pi}\left\langle\overline{\mathbb{A}_n\frac{\partial\psi_n}{\partial\theta_k}, \frac{\partial\psi_n}{\partial x_h}}\right\rangle - \frac{1}{2i\pi}\left\langle\overline{\mathbb{A}_n\frac{\partial\psi_n}{\partial x_h}, \frac{\partial\psi_n}{\partial\theta_k}}\right\rangle \\ &+ 2i\mathrm{Im}\int_{\mathbb{T}^N} A_{1,h}(\nabla_y - 2i\pi\theta^n)\bar{\psi}_n\cdot\psi_n e_k\,dy.\end{aligned} \tag{3.9}$$

By formulas (3.8)–(3.9) it is readily seen that equality (3.7) holds true.

In order to prove the self-adjointness of the operator $\mathbb{A}^*$, one first checks that $\mathbb{A}^*$ is symmetric, which easily follows by (3.7) and the fact that $\overline{B}^* = -B^*$, and then observes that up to addition of a multiple of the identity the operator $\mathbb{A}^*$ is monotone (see, e.g., [7, Chapter VII]). ∎

In the next proposition we will denote by $\nabla\nabla\lambda_n$ the Hessian matrix of the function $\lambda_n(x,\theta)$ evaluated at the point (x^n,θ^n), namely,

$$\nabla\nabla\lambda_n = \begin{pmatrix}\nabla_x\nabla_x\lambda_n & \nabla_\theta\nabla_x\lambda_n\\ \nabla_\theta\nabla_x\lambda_n & \nabla_\theta\nabla_\theta\lambda_n\end{pmatrix}(x^n,\theta^n).$$

PROPOSITION 3.5. *Assume that the matrix $\nabla\nabla\lambda_n$ is positive definite. Then there exists an orthonormal basis $\{\varphi_n\}_{n\geq 1}$ of eigenfunctions of $\mathbb{A}^*$; moreover, for each n there exists a real constant $\gamma_n > 0$ such that*

$$e^{\gamma_n|z|}\varphi_n,\ e^{\gamma_n|z|}\nabla\varphi_n \in L^2(\mathbb{R}^N). \tag{3.10}$$

Proof. Up to shifting the spectrum of the operator $\mathbb{A}^*$, we may assume that $\mathrm{Re}(c^*) = 0$. In order to prove the existence of an orthonormal basis of eigenfunctions we introduce the inverse operator of $\mathbb{A}^*$, denoted by G^*,

$$\begin{aligned} G^* : L^2(\mathbb{R}^N) &\to L^2(\mathbb{R}^N), \\ f &\to \varphi \text{ unique solution in } H^1(\mathbb{R}^N) \text{ of} \\ \mathbb{A}^*\varphi &= f \quad \text{in } \mathbb{R}^N, \end{aligned} \tag{3.11}$$

and we show that G^* is compact. Indeed multiplication of (3.11) by $\bar\varphi$ yields

$$\int_{\mathbb{R}^N} [A^*\nabla\varphi\cdot\nabla\bar\varphi - iB^*\mathrm{Im}(\varphi z\cdot\nabla\bar\varphi) + D^*z\cdot z|\varphi|^2]\,dz = \int_{\mathbb{R}^N} f\bar\varphi\,dz. \tag{3.12}$$

Upon defining the $2N$-dimensional vector-valued function Φ,

$$\Phi := \begin{pmatrix} 2i\pi z\varphi \\ \nabla\varphi \end{pmatrix},$$

we rewrite (3.12) in agreement with this block notation:

$$\int_{\mathbb{R}^N} \frac{1}{8\pi^2}\nabla\nabla\lambda_n\Phi\cdot\overline{\Phi}\,dz = \int_{\mathbb{R}^N} f\bar\varphi\,dz.$$

By the positivity assumption on the matrix $\nabla\nabla\lambda_n$ it follows that there exists a positive constant c_0 such that

$$c_0\big(\|\nabla\varphi\|^2_{L^2(\mathbb{R}^N)} + \|z\varphi\|^2_{L^2(\mathbb{R}^N)}\big) \leq \|f\|_{L^2(\mathbb{R}^N)}\|\varphi\|_{L^2(\mathbb{R}^N)},$$

which implies by a standard argument

$$\|\varphi\|^2_{L^2(\mathbb{R}^N)} + \|\nabla\varphi\|^2_{L^2(\mathbb{R}^N)} + \|z\varphi\|^2_{L^2(\mathbb{R}^N)} \leq C\|f\|^2_{L^2(\mathbb{R}^N)},$$

from which we deduce the compactness of G^* in $L^2(\mathbb{R}^N)$-strong. Thus there exists an infinite countable number of eigenvalues for $\mathbb{A}^*$.

We are left to prove the exponential decay of the eigenfunctions (this fact is quite standard; see, e.g., [2]). Let φ_n be an eigenfunction, and let σ_n be the associated eigenvalue

$$\mathbb{A}^*\varphi_n = \sigma_n\varphi_n. \tag{3.13}$$

Let $R_0 > 0$ and $\rho \in C^\infty(\mathbb{R})$ be a real function such that $0 \leq \rho \leq 1$, $\rho(s) = 0$ for $s \leq R_0$, and $\rho(s) = 1$ for $s \geq R_0 + 1$, and for every positive integer k define $\rho_k \in C^\infty(\mathbb{R}^N)$ in the following way:

$$\rho_k(z) := \rho(|z| - k).$$

We now multiply (3.13) by $\bar{\varphi}_n \rho_k^2$ to get

$$\int_{\mathbb{R}^N} \rho_k^2 \left(A^* \nabla \varphi_n \cdot \nabla \bar{\varphi}_n - iB^* \mathrm{Im}(\varphi_n z \cdot \nabla \bar{\varphi}_n) + D^* z \cdot z |\varphi_n|^2 - \sigma_n |\varphi_n|^2 \right) dz = \int_{\mathbb{R}^N} \left(\rho_k |\varphi_n|^2 B^* z \cdot \nabla \rho_k - 2\rho_k \, \bar{\varphi}_n A^* \nabla \varphi_n \cdot \nabla \rho_k \right) dz. \tag{3.14}$$

Next remark that since the left-hand side of (3.14) is real, the right-hand side must also be real, and therefore it is equal to

$$\int_{\mathbb{R}^N} -2\rho_k \, \mathrm{Re}(\bar{\varphi}_n A^* \nabla \varphi_n) \cdot \nabla \rho_k \, dz. \tag{3.15}$$

Let B_k denote the ball of radius $R_0 + k$ and center $z = 0$, and observe that the support of $\nabla \rho_k$ is contained in $B_{k+1} \setminus B_k$. Then putting up together (3.14) and (3.15) and again using the positive definiteness of the matrix $\nabla \nabla \lambda_n$, we obtain for R_0 sufficiently large ($\sqrt{R_0} > \sigma_n$ does the job)

$$\|\varphi_n\|^2_{H^1(\mathbb{R}^N \setminus B_{k+1})} \leq c_1 \big(\|\varphi_n\|^2_{H^1(\mathbb{R}^N \setminus B_k)} - \|\varphi_n\|^2_{H^1(\mathbb{R}^N \setminus B_{k+1})} \big),$$

where c_1 is a positive constant independent of k. Thus we deduce that

$$\|\varphi_n\|^2_{H^1(\mathbb{R}^N \setminus B_{k+1})} \leq \left(\frac{c_1}{1 + c_1} \right)^k \|\varphi_n\|^2_{H^1(\mathbb{R}^N \setminus B_0)}. \tag{3.16}$$

Upon defining a positive constant $\gamma_0 > 0$ by

$$\left(\frac{c_1}{1 + c_1} \right)^k = e^{-2\gamma_0 (k + R_0)},$$

it is finally seen that (3.16) implies the estimate (3.10) for any exponent $0 < \gamma_n < \gamma_0$. □

Proof of Theorem 3.2. We rescale the space variable by introducing

$$z = \frac{x}{\sqrt{\varepsilon}},$$

and define the sequence v_ε by

$$v_\varepsilon(t, z) := e^{-i \frac{\lambda_n t}{\varepsilon}} e^{-2i\pi \frac{\theta^n \cdot x}{\varepsilon}} u_\varepsilon(t, x). \tag{3.17}$$

By the a priori estimates of Lemma 3.1 it follows that $v_\varepsilon(t, z)$ satisfies

$$\|v_\varepsilon\|_{L^\infty(\mathbb{R}^+; L^2(\mathbb{R}^N))} + \sqrt{\varepsilon} \|\nabla v_\varepsilon\|_{L^\infty(\mathbb{R}^+; L^2(\mathbb{R}^N))} \leq C,$$

and applying the compactness of two-scale convergence (see Proposition 2.3), up to a subsequence, there exists a limit $v^*(t, z, y) \in L^2(\mathbb{R}^+ \times \mathbb{R}^N; H^1(\mathbb{T}^N))$ such that v_ε and $\sqrt{\varepsilon} \nabla v_\varepsilon$ two-scale converge to v^* and $\nabla_y v^*$, respectively. Similarly, by definition of the initial data, $v_\varepsilon(0, z)$ two-scale converges to $\psi_n(y) v^0(z)$.

Although v_ε is the unknown which will pass to the limit in what follows, it is simpler to write an equation for another function, namely

$$w_\varepsilon(t, z) := e^{2i\pi \frac{\theta^n \cdot z}{\sqrt{\varepsilon}}} v_\varepsilon(t, z) = e^{-i \frac{\lambda_n t}{\varepsilon}} u_\varepsilon(t, x). \tag{3.18}$$

By (3.18) it follows that

$$\nabla w_\varepsilon = e^{2i\pi\frac{\theta^n\cdot z}{\sqrt{\varepsilon}}}\left(\nabla + 2i\pi\frac{\theta^n}{\sqrt{\varepsilon}}\right)v_\varepsilon, \tag{3.19}$$

and it can be checked that the new unknown w_ε solves the following equation:

(3.20)
$$\begin{cases} i\dfrac{\partial w_\varepsilon}{\partial t} - \operatorname{div}[A\left(\sqrt{\varepsilon}z, z/\sqrt{\varepsilon}\right)\nabla w_\varepsilon] + \dfrac{1}{\varepsilon}[c(\sqrt{\varepsilon}z, z/\sqrt{\varepsilon}) - \lambda_n]w_\varepsilon = 0 & \text{in } \mathbb{R}^N\times\mathbb{R}^+, \\ w_\varepsilon(0,z) = u_\varepsilon^0(\sqrt{\varepsilon}z) & \text{in } \mathbb{R}^N, \end{cases}$$

where the differential operators div and ∇ act with respect to the new variable z.

First step. We multiply (3.20) by the complex conjugate of

$$\varepsilon\phi\Big(t, z, \frac{z}{\sqrt{\varepsilon}}\Big)e^{2i\pi\frac{\theta^n\cdot z}{\sqrt{\varepsilon}}},$$

where $\phi(s,z,y)$ is a smooth test function defined on $\mathbb{R}^+\times\mathbb{R}^N\times\mathbb{T}^N$, with compact support in $\mathbb{R}^+\times\mathbb{R}^N$. Since this test function has compact support (fixed with respect to ε), the effect of the nonperiodic variable in the coefficients is negligible for sufficiently small ε. Therefore we can replace the value of each coefficient at $(\sqrt{\varepsilon}z, z/\sqrt{\varepsilon})$ by its Taylor expansion of order two about the point $(0, z/\sqrt{\varepsilon})$. Integrating by parts and using (3.18) and (3.19) yields

$$\begin{aligned} &-i\varepsilon\int_0^{+\infty}\!\!\int_{\mathbb{R}^N} v_\varepsilon\frac{\partial\bar\phi^\varepsilon}{\partial t}\,dt\,dz - i\varepsilon\int_{\mathbb{R}^N} v_\varepsilon(0,z)\bar\phi\Big(0,z,\frac{z}{\sqrt{\varepsilon}}\Big)\,dz \\ &\quad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[A^\varepsilon + A^\varepsilon_{1,h}\sqrt{\varepsilon}z_h + \frac{1}{2}A^\varepsilon_{2,lh}\,\varepsilon z_l z_h + o(\varepsilon)\right](\sqrt{\varepsilon}\nabla + 2i\pi\theta^n)v_\varepsilon \\ &\qquad\cdot(\sqrt{\varepsilon}\nabla - 2i\pi\theta^n)\bar\phi^\varepsilon\,dz\,dt \\ &\quad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[c^\varepsilon + c^\varepsilon_{1,h}\sqrt{\varepsilon}z_h + \frac{1}{2}c^\varepsilon_{2,lh}\,\varepsilon z_l z_h + o(\varepsilon) - \lambda_n\right]v_\varepsilon\bar\phi^\varepsilon\,dz\,dt = 0. \end{aligned}$$

Passing to the two-scale limit, we get the variational formulation of

$$-(\operatorname{div}_y + 2i\pi\theta^n)(A(y)(\nabla_y + 2i\pi\theta^n)v^*) + c(y)v^* = \lambda_n v^* \quad \text{in } \mathbb{T}^N.$$

The simplicity of λ_n implies that there exists a scalar function $v(t,z)\in L^2(\mathbb{R}^+\times\mathbb{R}^N)$ such that

$$v^*(t,z,y) = v(t,z)\psi_n(y). \tag{3.21}$$

Second step. We multiply (3.20) by the complex conjugate of

$$\Psi_\varepsilon(t,z) = e^{2i\pi\theta^n\cdot\frac{z}{\sqrt{\varepsilon}}}\left[\psi_n^\varepsilon\phi(t,z) + \sqrt{\varepsilon}\sum_{k=1}^N\left(\frac{1}{2i\pi}\frac{\partial\psi_n^\varepsilon}{\partial\theta_k}\frac{\partial\phi}{\partial z_k}(t,z) + z_k\frac{\partial\psi_n^\varepsilon}{\partial x_k}\phi(t,z)\right)\right],$$

where $\phi(t,z)$ is a smooth test function with compact support in $\mathbb{R}^+ \times \mathbb{R}^N$. We first look at those terms of the equation involving time derivatives:

(3.22)
$$\begin{aligned}
&\int_0^{+\infty}\!\!\int_{\mathbb{R}^N} i\frac{\partial w_\varepsilon}{\partial t}\bar{\Psi}_\varepsilon\,dt\,dz \\
&\quad= \int_0^{+\infty}\!\!\int_{\mathbb{R}^N} -iv_\varepsilon\left[\bar{\psi}_n^\varepsilon\frac{\partial\bar{\phi}}{\partial t}+\sqrt{\varepsilon}\sum_{k=1}^N\left(-\frac{1}{2i\pi}\frac{\partial\bar{\psi}_n^\varepsilon}{\partial\theta_k}\frac{\partial^2\bar{\phi}}{\partial t\partial z_k}+z_k\frac{\partial\bar{\psi}_n^\varepsilon}{\partial x_k}\frac{\partial\bar{\phi}}{\partial t}\right)\right]dt\,dz \\
&\qquad- i\int_{\mathbb{R}^N} v_\varepsilon(0,z)\left[\bar{\psi}_n^\varepsilon\bar{\phi}(0,z)+\sqrt{\varepsilon}\sum_{k=1}^N\left(-\frac{1}{2i\pi}\frac{\partial\bar{\psi}_n^\varepsilon}{\partial\theta_k}\frac{\partial\bar{\phi}}{\partial z_k}(0,z)+z_k\frac{\partial\bar{\psi}_n^\varepsilon}{\partial x_k}\bar{\phi}(0,z)\right)\right]dz.
\end{aligned}$$

Recalling the normalization $\int_{\mathbb{T}^N}|\psi_n|^2\,dy=1$, we find that the two-scale limit of the term on the left-hand side of (3.22) is given by the expression

$$-i\int_0^{+\infty}\!\!\int_{\mathbb{R}^N} v\frac{\partial\bar{\phi}}{\partial t}\,dz\,dt - i\int_{\mathbb{R}^N} v^0\bar{\phi}(0,z)\,dz. \tag{3.23}$$

We further decompose Ψ_ε as follows:

$$\Psi_\varepsilon=\Psi_\varepsilon^1+\Psi_\varepsilon^2\cdot z \quad\text{ with }\quad \Psi_\varepsilon^2=\sqrt{\varepsilon}e^{2i\pi\theta^n\cdot\frac{z}{\sqrt{\varepsilon}}}\sum_{k=1}^N\frac{\partial\psi_n^\varepsilon}{\partial x_k}\phi(t,z)e_k.$$

Getting rid of all terms multiplied by $o(\varepsilon)$ and taking into account (3.18) and (3.19), we next pass to the limit in the remaining terms of (3.20) multiplied by $\bar{\Psi}_\varepsilon$. The computation is similar to these in [4], but it involves new terms since ψ_n and its derivatives also depend on x. We first look at those terms which are of zero order with respect to z, namely,

(3.24)
$$\begin{aligned}
&\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[A^\varepsilon\nabla w_\varepsilon\cdot(\nabla\bar{\Psi}_\varepsilon^1+\bar{\Psi}_\varepsilon^2)+\frac{1}{\varepsilon}(c^\varepsilon-\lambda_n)w_\varepsilon\bar{\Psi}_\varepsilon^1\right]dz\,dt \\
&\quad=\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[\frac{1}{\varepsilon}A^\varepsilon\big(\sqrt{\varepsilon}\nabla+2i\pi\theta^n\big)v_\varepsilon\cdot(\nabla_y-2i\pi\theta^n)\bar{\psi}_n^\varepsilon\bar{\phi}+\frac{1}{\varepsilon}(c^\varepsilon-\lambda_n)\bar{\psi}_n^\varepsilon v_\varepsilon\bar{\phi}\right]dz\,dt \\
&\qquad-\frac{1}{2i\pi}\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[\frac{1}{\sqrt{\varepsilon}}A^\varepsilon\big(\sqrt{\varepsilon}\nabla+2i\pi\theta^n\big)v_\varepsilon\cdot(\nabla_y-2i\pi\theta^n)\frac{\partial\bar{\psi}_n^\varepsilon}{\partial\theta_k}\frac{\partial\bar{\phi}}{\partial z_k}\right. \\
&\qquad\qquad\left.+\frac{1}{\sqrt{\varepsilon}}(c^\varepsilon-\lambda_n)v_\varepsilon\frac{\partial\bar{\psi}_n^\varepsilon}{\partial\theta_k}\frac{\partial\bar{\phi}}{\partial z_k}\right]dz\,dt \\
&\qquad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\frac{1}{\sqrt{\varepsilon}}A^\varepsilon\big(\sqrt{\varepsilon}\nabla+2i\pi\theta^n\big)v_\varepsilon\cdot\bar{\psi}_n^\varepsilon\nabla\bar{\phi}\,dz\,dt \\
&\qquad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}-\frac{1}{2\pi i}A^\varepsilon\big(\sqrt{\varepsilon}\nabla+2i\pi\theta^n\big)v_\varepsilon\cdot\frac{\partial\bar{\psi}_n^\varepsilon}{\partial\theta_k}\nabla\frac{\partial\bar{\phi}}{\partial z_k}\,dz\,dt \\
&\qquad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}A^\varepsilon\big(\sqrt{\varepsilon}\nabla+2i\pi\theta^n\big)v_\varepsilon\cdot\frac{\partial\bar{\psi}_n^\varepsilon}{\partial x_k}\bar{\phi}\,e_k\,dz\,dt.
\end{aligned}$$

Using (2.7) with $\varphi=v_\varepsilon\bar{\phi}$ and (2.8) with $\varphi=v_\varepsilon\frac{\partial\bar{\phi}}{\partial z_k}$, we rewrite the first two integrals

in the right-hand side of (3.24) as follows:

$$\int_0^{+\infty}\!\!\int_{\mathbb{R}^N} -\frac{1}{\sqrt{\varepsilon}}A^\varepsilon(\nabla_y - 2i\pi\theta^n)\bar\psi_n^\varepsilon\cdot v_\varepsilon\nabla\bar\phi\,dz\,dt$$
$$+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[\frac{1}{2i\pi}A^\varepsilon(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n^\varepsilon}{\partial\theta_k}\cdot v_\varepsilon\nabla\frac{\partial\bar\phi}{\partial z_k}+\frac{1}{\sqrt{\varepsilon}}A^\varepsilon e_k\cdot v_\varepsilon\frac{\partial\bar\phi}{\partial z_k}(\nabla_y - 2i\pi\theta^n)\bar\psi_n^\varepsilon\right.$$
$$\left.-\frac{1}{\sqrt{\varepsilon}}A^\varepsilon\bar\psi_n^\varepsilon e_k\cdot\big(\sqrt{\varepsilon}\nabla + 2i\pi\theta^n\big)\left(v_\varepsilon\frac{\partial\bar\phi}{\partial z_k}\right)\right]dz\,dt.$$

Combining the above terms with the other terms in (3.24) and passing to the two-scale limit in (3.24) yields

$$\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\left[\frac{1}{2i\pi}A\psi_n(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial\theta_k}\right.$$
$$\left.-\frac{1}{2i\pi}A\frac{\partial\bar\psi_n}{\partial\theta_k}(\nabla_y + 2i\pi\theta^n)\psi_n - A|\psi_n|^2e_k\right]\cdot v\nabla\frac{\partial\bar\phi}{\partial z_k}\,dy\,dz\,dt$$
$$+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}A(\nabla_y + 2i\pi\theta^n)\psi_n\cdot\frac{\partial\bar\psi_n}{\partial x_k}v\bar\phi\,e_k\,dy\,dz\,dt. \tag{3.25}$$

By (2.6) it can be seen that the first integral of (3.25) equals

$$\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}A^*\nabla v\nabla\bar\phi\,dz\,dt. \tag{3.26}$$

We now focus on those terms which are linear in z:

(3.27)

$$\int_0^{+\infty}\int_{\mathbb{R}^N}\left[A^\varepsilon\nabla w_\varepsilon\cdot(\nabla\bar\Psi_\varepsilon^2 z)+\frac{1}{\varepsilon}(c^\varepsilon-\lambda_n)w_\varepsilon\bar\Psi_\varepsilon^2 z+A_{1,k}^\varepsilon\sqrt{\varepsilon}z_k\nabla w_\varepsilon\cdot(\nabla\bar\Psi_\varepsilon^1+\bar\Psi_\varepsilon^2)\right.$$
$$\left.+\frac{1}{\sqrt{\varepsilon}}c_{1,k}^\varepsilon z_k w_\varepsilon\bar\Psi_e^1\right]dz\,dt$$
$$=\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[\frac{1}{\sqrt{\varepsilon}}A^\varepsilon\big(\sqrt{\varepsilon}\nabla+2i\pi\theta^n\big)v_\varepsilon\cdot(\nabla_y-2i\pi\theta^n)\frac{\partial\bar\psi_n^\varepsilon}{\partial x_k}\,\bar\phi z_k\right.$$
$$\left.+\frac{1}{\sqrt{\varepsilon}}(c^\varepsilon-\lambda_n)v_\varepsilon\frac{\partial\bar\psi_n^\varepsilon}{\partial x_k}\,\bar\phi z_k\right]dz\,dt$$
$$+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[\frac{1}{\sqrt{\varepsilon}}A_{1,k}^\varepsilon(\sqrt{\varepsilon}\nabla+2i\pi\theta^n)v_\varepsilon\cdot(\nabla_y-2i\pi\theta^n)\bar\psi_n^\varepsilon\,\bar\phi z_k+\frac{1}{\sqrt{\varepsilon}}c_{1,k}^\varepsilon v_\varepsilon\bar\psi_n^\varepsilon\,\bar\phi\,z_k\right]dz\,dt$$
$$+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[A^\varepsilon(\sqrt{\varepsilon}\nabla+2i\pi\theta^n)v_\varepsilon\cdot\frac{\partial\bar\psi_n^\varepsilon}{\partial x_k}\nabla\bar\phi\,z_k\right.$$
$$\left.+A_{1,k}^\varepsilon(\sqrt{\varepsilon}\nabla+2i\pi\theta^n)v_\varepsilon\cdot\bar\psi_n^\varepsilon\nabla\bar\phi\,z_k\right]dz\,dt$$
$$-\frac{1}{2i\pi}\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[A_{1,h}^\varepsilon(\sqrt{\varepsilon}\nabla+2i\pi\theta^n)v_\varepsilon\cdot(\nabla_y-2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial\theta_k}\frac{\partial\bar\phi}{\partial z_k}\,z_h\right.$$
$$\left.+c_{1,h}^\varepsilon v_\varepsilon\frac{\partial\bar\psi_n}{\partial\theta_k}\frac{\partial\bar\phi}{\partial z_k}\,z_h\right]dz\,dt$$
$$+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[\sqrt{\varepsilon}A_{1,h}^\varepsilon(\sqrt{\varepsilon}\nabla+2i\pi\theta^n)v_\varepsilon\cdot\left(-\frac{1}{2i\pi}\frac{\partial\bar\psi_n^\varepsilon}{\partial\theta_k}\nabla\frac{\partial\bar\phi}{\partial z_k}+\frac{\partial\bar\psi_n^\varepsilon}{\partial x_k}\bar\phi\,e_k\right)z_h\right]dz\,dt.$$

By (2.9) with $\varphi = v_\varepsilon \bar\phi z_k$, it can be seen that the sum of the first two integrals in the right-hand side of (3.27) gives

(3.28)
$$-\int_0^{+\infty}\int_{\mathbb{R}^N}\left(A^\varepsilon(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n^\varepsilon}{\partial x_k}\cdot v_\varepsilon\nabla(\bar\phi z_k) + A^\varepsilon_{1,k}(\nabla_y - 2i\pi\theta^n)\bar\psi_n^\varepsilon\cdot v_\varepsilon\nabla(\bar\phi z_k)\right)dz\,dt.$$

Therefore passing to the two-scale limit in (3.27), we find

$$\begin{aligned}
(3.29)\qquad &-\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\Big[A(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial x_k}\cdot v\psi_n\bar\phi\,e_k\\
&\qquad\qquad + A_{1,k}(\nabla_y - 2i\pi\theta^n)\bar\psi_n\cdot v\psi_n\bar\phi\,e_k\Big]\,dy\,dz\,dt\\
&-\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\Big[A(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial x_k}\cdot v\psi_n z_k\nabla\bar\phi\\
&\qquad\qquad + A_{1,k}(\nabla_y - 2i\pi\theta^n)\bar\psi_n\cdot v\psi_n z_k\nabla\bar\phi\Big]\,dy\,dz\,dt\\
&+\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\Big[A(\nabla_y + 2i\pi\theta^n)\psi_n\cdot v\frac{\partial\bar\psi_n}{\partial x_k}z_k\nabla\bar\phi\\
&\qquad\qquad + A_{1,k}(\nabla_y + 2i\pi\theta^n)\psi_n\cdot v\bar\psi_n z_k\nabla\bar\phi\Big]\,dy\,dz\,dt\\
&-\frac{1}{2i\pi}\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\Big[A_{1,h}(\nabla_y + 2i\pi\theta^n)\psi_n\cdot(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial\theta_k}vz_h\frac{\partial\bar\phi}{\partial z_k}\\
&\qquad\qquad + c_{1,h}\psi_n\frac{\partial\bar\psi_n}{\partial\theta_k}vz_h\frac{\partial\bar\phi}{\partial z_k}\Big]\,dy\,dz\,dt.
\end{aligned}$$

By (2.4) it follows that the last integral in (3.29) is equal to

$$\begin{aligned}
(3.30)\qquad &\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\Big[A_{1,h}\psi_n e_k\cdot(\nabla_y - 2i\pi\theta^n)\bar\psi_n\\
&\qquad\qquad + A\psi_n e_k\cdot(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial x_h}\psi_n\Big]vz_h\frac{\partial\bar\phi}{\partial z_k}\;dy\,dz\,dt\\
&-\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\Big[A_{1,h}\bar\psi_n e_k\cdot(\nabla_y + 2i\pi\theta^n)\psi_n\\
&\qquad\qquad + A\frac{\partial\bar\psi_n}{\partial x_h}e_k\cdot(\nabla_y + 2i\pi\theta^n)\psi_n\Big]vz_h\frac{\partial\bar\phi}{\partial z_k}\;dy\,dz\,dt\\
&-\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\frac{1}{2i\pi}\frac{\partial^2\lambda_n}{\partial x_h\partial\theta_k}|\psi_n|^2vz_h\frac{\partial\bar\phi}{\partial z_k}\;dy\,dz\,dt.
\end{aligned}$$

Next notice that the first and the second lines of (3.30) cancel out with the second and the third lines of (3.29), respectively, and therefore (3.29) reduces to

$$\begin{aligned}
(3.31)\qquad &-\int_0^{+\infty}\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\Big[A(\nabla_y - 2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial x_k}\cdot v\psi_n\bar\phi\,e_k\\
&\qquad\qquad + A_{1,k}(\nabla_y - 2i\pi\theta^n)\bar\psi_n\cdot v\psi_n\bar\phi\,e_k\Big]\;dy\,dz\,dt\\
&-\int_0^{+\infty}\int_{\mathbb{R}^N}\frac{1}{2i\pi}\frac{\partial^2\lambda_n}{\partial x_h\partial\theta_k}v\frac{\partial\bar\phi}{\partial z_k}\,z_h\,dz\,dt.
\end{aligned}$$

Finally we consider all quadratic in z terms:

$$
\begin{aligned}
&\frac{1}{2}\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[A^\varepsilon_{2,lh}\,\varepsilon z_l z_h \nabla w_\varepsilon\cdot(\nabla\bar\Psi^1_\varepsilon+\bar\Psi^2_\varepsilon)+c^\varepsilon_{2,lh}z_l z_h w_\varepsilon\bar\Psi^1_e\right]dz\,dt\\
&\quad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\left[A^\varepsilon_{1,k}\sqrt{\varepsilon}z_k\nabla w_\varepsilon\cdot(z\nabla\bar\Psi^2_\varepsilon)+\frac{1}{\sqrt{\varepsilon}}c^\varepsilon_{1,k}z_k w_\varepsilon z\cdot\bar\Psi^2_\varepsilon\right]dz\,dt\\
&\quad=\frac{1}{2}\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}A^\varepsilon_{2,lh}\,\sqrt{\varepsilon}z_l z_h(\sqrt{\varepsilon}\nabla+2i\pi\theta^n)v_\varepsilon\cdot\left[\frac{1}{\sqrt{\varepsilon}}(\nabla_y-2i\pi\theta^n)\bar\psi^\varepsilon_n\bar\phi+\bar\psi^\varepsilon_n\nabla\bar\phi\right]dz\,dt\\
&\quad-\frac{1}{2}\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}A^\varepsilon_{2,lh}\,\sqrt{\varepsilon}z_l z_h(\sqrt{\varepsilon}\nabla+2i\pi\theta^n)v_\varepsilon\\
&\qquad\qquad\cdot\left[\frac{1}{2\pi i}\nabla_y\frac{\partial\bar\psi^\varepsilon_n}{\partial\theta_k}\frac{\partial\bar\phi}{\partial z_k}+\sqrt{\varepsilon}\left(\frac{1}{2i\pi}\frac{\partial\bar\psi^\varepsilon_n}{\partial\theta_k}\nabla\frac{\partial\bar\phi}{\partial z_k}+e_k\frac{\partial\bar\psi^\varepsilon_n}{\partial x_k}\bar\phi\right)\right]dz\,dt\\
&\quad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}A^\varepsilon_{1,h}\,z_h\Big(\sqrt{\varepsilon}\nabla+2i\pi\theta^n\Big)v_\varepsilon\cdot\left[z_k(\nabla_y-2i\pi\theta^n)\frac{\partial\bar\psi^\varepsilon_n}{\partial x_k}\bar\phi+\sqrt{\varepsilon}z_k\frac{\partial\bar\psi^\varepsilon_n}{\partial x_k}\nabla\bar\phi\right]dz\,dt\\
&\quad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\frac{1}{2}c^\varepsilon_{2,lh}z_l z_h v_\varepsilon\left(\bar\psi^\varepsilon_n\bar\phi-\sqrt{\varepsilon}\frac{1}{2i\pi}\frac{\partial\bar\psi^\varepsilon_n}{\partial\theta_k}\frac{\partial\bar\phi}{\partial z_k}\right)dz\,dt\\
&\quad+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}c^\varepsilon_{1,h}z_h v_\varepsilon z_k\frac{\partial\bar\psi^\varepsilon_n}{\partial x_k}\bar\phi\,dz\,dt,
\end{aligned}
$$

which gives, on passing to the two-scale limit,

(3.32)

$$
\begin{aligned}
&\frac{1}{2}\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\left[A_{2,lh}(\nabla_y+2i\pi\theta^n)\psi_n\cdot(\nabla_y-2i\pi\theta^n)\bar\psi_n+c_{2,lh}\psi_n\bar\psi_n\right]v\bar\phi\,z_l z_h\,dy\,dz\,dt\\
&+\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\int_{\mathbb{T}^N}\left[A_{1,h}(\nabla_y+2i\pi\theta^n)\psi_n\cdot(\nabla_y-2i\pi\theta^n)\frac{\partial\bar\psi_n}{\partial x_k}\right.\\
&\qquad\qquad\left.+\,c_{1,h}\psi_n\frac{\partial\bar\psi_n}{\partial x_k}\right]v\bar\phi\,z_h z_k\,dy\,dz\,dt.
\end{aligned}
$$

Now, using (2.5), we find that (3.32) reduces itself to

$$
\int_0^{+\infty}\!\!\int_{\mathbb{R}^N}\frac{1}{2}\frac{\partial^2\lambda_n}{\partial x_l\partial x_h}\,v\bar\phi\,z_l z_h\,dz\,dt. \tag{3.33}
$$

Summing up together (3.23), (3.25), (3.26), (3.31), and (3.33) yields the weak formulation of (3.3). By uniqueness of the solution of the homogenized problem (3.3), we deduce that the entire sequence v_ε two-scale converges weakly to $\psi_n(y)v(t,x)$.

It remains to prove the strong two-scale convergence of v_ε. By Lemma 3.1 we have

$$
\|v_\varepsilon(t)\|_{L^2(\mathbb{R}^N)}=\|u_\varepsilon(t)\|_{L^2(\mathbb{R}^N)}=\|u^0_\varepsilon\|_{L^2(\mathbb{R}^N)}\to\|\psi_n v^0\|_{L^2(\mathbb{R}^N\times\mathbb{T}^N)}=\|v^0\|_{L^2(\mathbb{R}^N)}
$$

by the normalization condition of ψ_n. From the conservation of energy of the homogenized equation (3.3) we have

$$
\|v(t)\|_{L^2(\mathbb{R}^N)}=\|v^0\|_{L^2(\mathbb{R}^N)},
$$

and thus we deduce the strong convergence from Proposition 2.3. □

Remark 3.6. As usual in periodic homogenization [1], [5], [15], the choice of the test function Ψ_ε in the proof of Theorem 3.2 is dictated by the formal two-scale asymptotic expansion that can be obtained for the solution w_ε of (3.20), namely,

$$w_\varepsilon(t,z) \approx e^{2i\pi\theta^n\cdot\frac{z}{\sqrt{\varepsilon}}}\left[\psi_n\left(\frac{z}{\sqrt{\varepsilon}}\right)v(t,z)+\sqrt{\varepsilon}\sum_{k=1}^{N}\left(\frac{1}{2i\pi}\frac{\partial\psi_n}{\partial\theta_k}\left(\frac{z}{\sqrt{\varepsilon}}\right)\frac{\partial v}{\partial z_k}(t,z)\right.\right.$$
$$\left.\left.+\,z_k\frac{\partial\psi_n}{\partial x_k}\left(\frac{z}{\sqrt{\varepsilon}}\right)v(t,z)\right)\right],$$

where v is the homogenized solution of (3.3). Actually the homogenized equation that one gets by the asymptotic expansion method is

$$i\frac{\partial v}{\partial t}-\operatorname{div}\left(A^*\nabla v\right)+B^*\nabla v\cdot z+\bar{c}^*v+vD^*z\cdot z=0, \tag{3.34}$$

which apparently differs from (3.3) by the following zero order term:

$$\left(\operatorname{tr}\left(\nabla_\theta\nabla_x\lambda_n\right)-4\pi\mathrm{Im}(c^*)\right)v.$$

By virtue of (3.7), the above term vanishes, so that formulas (3.34) and (3.3) are equivalent.

Acknowledgments. This work was done while M. Palombaro was working as a post doc at the Centre de Mathématiques Appliquées of Ecole Polytechnique. The hospitality of people there is gratefully acknowledged.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.
[2] G. Allaire and M. Amar, *Boundary layer tails in periodic homogenization*, ESIAM Control Optim. Calc. Var., 4 (1999), pp. 209–243.
[3] G. Allaire and A. Piatnitski, *Uniform spectral asymptotics for singularly perturbed locally periodic operators,* Comm. Partial Differential Equations, 27 (2002), pp. 705–725.
[4] G. Allaire and A. Piatnitski, *Homogenization of the Schrödinger equation and effective mass theorems,* Comm. Math. Phys., 258 (2005), pp. 1–22.
[5] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures,* North–Holland, Amsterdam, 1978.
[6] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert,* North–Holland, Amsterdam, 1973.
[7] H. Brézis, *Analyse fonctionelle*, Masson, Paris, 1983.
[8] V. Buslaev, *Semiclassical approximation for equations with periodic coefficients*, Russ. Math. Surv., 42 (1987), pp. 97–125.
[9] R. Carmona and J. Lacroix, *Spectral Theory of Random Schrödinger Operators*, Birkhäuser Boston, Cambridge, MA, 1990.
[10] M. Dimassi, J.-C. Guillot, and J. Ralston, *Semiclassical asymptotics in magnetic Bloch bands*, J. Phys. A, 35 (2002), pp. 7597–7605.
[11] P. Gérard, *Mesures semi-classiques et ondes de Bloch*, Séminaire sur les équations aux Dérivées Partielles, 1990–1991, Exp. No. XVI, École Polytech., Palaiseau, 1991, pp. 1–18.
[12] P. Gérard, P. Markowich, N. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms,* Comm. Pure Appl. Math., 50 (1997), pp. 323–379.
[13] C. Gérard, A. Martinez, and J. Sjöstrand, *A mathematical approach to the effective Hamiltonian in perturbed periodic problems*, Comm. Math. Phys., 142 (1991), pp. 217–244.
[14] J.-C. Guillot, J. Ralston, and E. Trubowitz, *Semi-classical methods in solid state physics*, Comm. Math. Phys., 116 (1988), pp. 401–415.

[15] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, New York, 1994.
[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
[17] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[18] G. Panati, H. Sohn, and S. Teufel, *Effective dynamics for Bloch electrons: Peierls substitution and beyond*, Comm. Math. Phys., 242 (2003), pp. 547–578.
[19] F. Poupaud and C. Ringhofer, *Semi-classical limits in a crystal with exterior potentials and effective mass theorems*, Comm. Partial Differential Equations, 21 (1996), pp. 1897–1918.
[20] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1978.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 143–165

VISCOSITY SOLUTIONS OF INCREASING FLOWS OF SETS. APPLICATION OF THE HELE–SHAW PROBLEM FOR POWER-LAW FLUIDS*

PIERRE CARDALIAGUET† AND ELISABETH ROUY‡

Abstract. An existence and inclusion principle of viscosity solutions for increasing flows of sets is established. A typical example—the Hele–Shaw moving boundary problem for a power-law fluid—is discussed extensively.

Key words. moving boundary problems, power-law fluids, viscosity solutions

AMS subject classifications. 53C44, 35D05, 35K55

DOI. 10.1137/040613366

1. Introduction.

The aim of this work is to investigate a viscosity solutions approach for increasing flows of sets with velocity laws of the form

$$V_{t,x} = h(x, \Omega(t)) \qquad \forall x \in \partial\Omega(t), \tag{1}$$

where $V_{t,x}$ denotes the normal velocity to the expanding family of sets $(\Omega(t))$ and $h = h(x, \Omega(t))$ is some nonnegative velocity law depending nonlocally on the set $\Omega(t)$.

A typical example of such evolution is the Hele–Shaw approximation model of the injection of a power-law fluid between two closely situated plates. Since the fluid is supposed to be surrounded by another fluid with small viscosity, this is a one-phase moving boundary problem.

Let us denote by S the source of the injection, by $\Omega(t)$ the portion of space occupied by the fluid at time t, and by $\Sigma(t)$ the moving boundary. According to [2], [3], and [22], $\Sigma(t)$ evolves with a normal velocity $V_{t,x}$ given, at each point $x \in \Sigma(t)$, by the quasi-static equation

$$V_{t,x} = h(x, \Omega(t)), \qquad \text{where} \qquad h(x, \Omega(t)) = |\nabla u(t,x)|^{p-1}\,, \tag{2}$$

and where $u(t,x)$ satisfies at any time $t > 0$ the p-Laplace equation (with $p > 1$)

$$\begin{cases} -\text{div}\left(|\nabla u(t,x)|^{p-2}\nabla u(t,x)\right) = f(x) & \text{in } \Omega(t)\backslash S, \\ u(t,x) = 0 & \text{on } \Sigma(t), \\ u(t,x) = g(x) & \text{on } S \end{cases} \tag{3}$$

for some positive functions f and g.

In this paper, we define a notion of viscosity solutions for (1) and prove existence of such solutions. Our most important statement is the inclusion principle which holds for those generalized solutions. Our results are true under some structure condition on the velocity h that we detail further.

*Received by the editors August 13, 2004; accepted for publication (in revised form) November 18, 2005; published electronically April 12, 2006. The authors are partially supported by the ACI grant JC 1041 "Mouvements d'interface avec termes non-locaux" from the French Ministry of Research.
http://www.siam.org/journals/sima/38-1/61336.html

†Laboratoire de Mathématiques UMR 6205, Université de Bretagne Occidentale, 6 Av. Le Gorgeu, BP 809, 29285 Brest, France (Pierre.Cardaliaguet@univ-brest.fr).

‡Département de Mathématiques Informatique, Ecole Centrale de Lyon, 36 Av. Guy de Collongue, 69134 Ecully Cedex, France (rouy@univ-lyon1.fr).

When $p = 2$, this example is the well-known Hele–Shaw problem. If the initial data are smooth, the evolution equation has a smooth short time solution (see [15]), but singularities generally appear in finite time. In order to define the solutions after the onset of singularities, various notions of generalized solutions have been introduced. For instance, the Hele–Shaw problem is reformulated in terms of variational inequalities via the Baiocchi transform in [14]; Hele–Shaw with surface tension is understood as the gradient flow of some functional in the space of measures in [17], while Kim proposes in [19] a definition of viscosity solutions for this problem. To the best of our knowledge, the case of $p \neq 2$ has never been studied up to now.

The main assumption we require on the velocity law $h = h(x, \Omega)$ is that it is nondecreasing with respect to Ω. Under this assumption, at least formally, the flow preserves inclusion. Namely, if $(\Omega_1(t))_t$ and $(\Omega_2(t))_t$ are two families of solutions, with $\Omega_1(0) \subset \Omega_2(0)$, then this inclusion is preserved for all time. The main result of the paper (Theorem 3.1) is that this "inclusion principle" holds true even for weak solutions when h satisfies some additional regularity conditions.

The inclusion principle is one of the key tools for constructing generalized solutions of front propagation problems: viscosity solutions in [16], [12] for the so-called level-set approach of mean curvature motion; related but more geometric viscosity solutions in [23], [4], [5]; and barrier solutions in [6], [7], and [8]. Viscosity solutions have also been introduced for the porous-medium equation [9] and for a free boundary problem motivated by combustion [20]. In [19] Kim proved the inclusion principle for the viscosity solutions of the Hele–Shaw problem when $p = 2$, $f \equiv 0$, and a particular source S. Let us underline that, although our work uses some ideas and techniques which have counterparts in [19], its point of view is completely different: the main unknown in [19] is the evolving function u (given by (3) for $p = 2$) while our unknown is the evolving family of sets. Both approaches should lead to different extensions. In that respect, we spent some effort in this paper to prove the inclusion principle for a wide class of (nonnegative) velocity laws h and not only for (2). This is completely new in this framework. Let us underline that the result could be extended to velocity laws (1) depending at each time t not only on the set $\Omega(t)$ but also on the evolving family $(\Omega(s))_{s \leq t}$.

For proving this inclusion principle we use several ideas introduced by the first author in [11] for simpler moving boundary problems. In particular, we use two basic ingredients of [11]: an equivalent definition of solutions (Proposition 2.7) and an Ilmanen interposition lemma [18]. However, proofs differ substantially. Indeed, in [11], the velocity is invariant by translation, a property which is no longer satisfied here.

We now briefly explain the organization of the paper. We first introduce the notion of viscosity solutions for the evolution equation (1) and investigate the main properties of the velocity defined by (2). Next, we state and prove the inclusion principle. Finally, we apply this inclusion principle to derive existence, uniqueness, and stability of solutions when the velocity is given by (2).

2. Definitions and preliminary results.

2.1. Definition of the solutions. Let us first fix some notations: throughout the paper $|\cdot|$ denotes the Euclidean norm (of $\mathbb{R}^N$ or $\mathbb{R}^{N+1}$, depending on the context). If K is a subset of $\mathbb{R}^N$ and $x \in \mathbb{R}^N$, then $d_K(x)$ denotes the usual distance from x to K: $d_K(x) = \inf_{y \in K} |y - x|$. Finally, we denote by $B(x, R)$ the open ball centered at x and of radius R.

We intend to study the evolution of compact hypersurfaces $\Sigma(t) = \partial\Omega(t)$ of $\mathbb{R}^N$, where $\Omega(t)$ is an open set, evolving with the following law:

$$\forall t \geq 0,\ x \in \Sigma(t),\ \ V_{t,x} = h(x, \Omega(t)), \tag{4}$$

where $V_{t,x}$ is the normal velocity of $\Omega(t)$ at the point x, and $h = h(x, \Omega)$ is defined for any $x \in \partial\Omega$ and for any open set Ω with $\mathcal{C}^{1,1}$ boundary which belongs to some class of sets $\mathcal{D}$. To fix the ideas we assume that the set $\mathcal{D}$ is of the form

$$\mathcal{D} = \big\{\Omega \subset \mathbb{R}^N \mid \Omega \text{ is open, bounded, and such that } S \subset\subset \Omega\big\}, \tag{5}$$

where S is some compact subset of $\mathbb{R}^N$. This suffices in most applications.

Our key assumption is that h is nonnegative and nondecreasing with respect to the set Ω: Namely

$$h(x, \Omega) \geq 0 \qquad \forall \Omega \in \mathcal{D} \text{ and } x \in \Omega, \tag{6}$$

and

$$\begin{gathered}\text{if } \Omega_1 \in \mathcal{D} \text{ and } \Omega_2 \in \mathcal{D},\ \Omega_1 \subset \Omega_2, \text{ and } x \in \Omega_1 \cap \partial\Omega_2,\\ \text{then } h(x, \Omega_1) \leq h(x, \Omega_2).\end{gathered} \tag{7}$$

We explain below that the assumptions above are typically satisfied for the Hele–Shaw problem described in the introduction.

From now on, we consider the graph of the evolving sets $\Omega(t)$. In order to underline the fact that it need not be smooth, we denote this graph by $\mathcal{K}$. Then $\mathcal{K}$ is a subset of $\mathbb{R}^+ \times \mathbb{R}^N$. Formally, with the notations above,

$$\mathcal{K} = \{(t, x) \text{ such that } x \in \Omega(t)\}.$$

The set $\mathcal{K}$ is our main unknown. We denote by (t, x) an element of such a set, where $t \in \mathbb{R}^+$ denotes the time and $x \in \mathbb{R}^N$ denotes the space. We set

$$\mathcal{K}(t) \ = \ \big\{x \in \mathbb{R}^N \mid (t, x) \in \mathcal{K}\big\}.$$

The closure of the set $\mathcal{K}$ in $\mathbb{R}^{N+1}$ is denoted by $\overline{\mathcal{K}}$. The closure of the complementary of $\mathcal{K}$ is denoted by $\widehat{\mathcal{K}}$:

$$\widehat{\mathcal{K}} = \overline{(\mathbb{R}^+ \times \mathbb{R}^N) \setminus \mathcal{K}}$$

and we set

$$\widehat{\mathcal{K}}(t) = \{x \in \mathbb{R}^N \mid (t, x) \in \widehat{\mathcal{K}}\}.$$

Let us go further into terminology: If $\mathcal{K}$ is a subset of $[0, +\infty) \times \mathbb{R}^N$, we say that

- $\mathcal{K}$ is a *tube* if $\forall T \geq 0$, $\overline{\mathcal{K}} \cap ([0, T] \times \mathbb{R}^N)$ is a compact subset of $\mathbb{R}^{N+1}$.
- $\mathcal{K}$ is *nondecreasing* if $\mathcal{K}(s) \subset \mathcal{K}(t)$ for any $0 \leq s \leq t$.
- $\mathcal{K}$ is *left lower semicontinuous* if

$$\forall t > 0,\ \forall x \in \mathcal{K}(t),\ \text{if } t_n \to t^-,\ \ \exists x_n \in \mathcal{K}(t_n) \text{ such that } x_n \to x.$$

- $\mathcal{K}_r$ is a *smooth tube* if $\mathcal{K}_r$ is closed in $I \times \mathbb{R}^N$ (where I is some open interval), has a nonempty interior, and $\partial\mathcal{K}_r \cap (I \times \mathbb{R}^N)$ is a $\mathcal{C}^{1,1}$ submanifold of $\mathbb{R}^{N+1}$, such that at any point $(t, x) \in \mathcal{K}_r$ the outward normal (ν_t, ν_x) to $\mathcal{K}_r$ at (t, x) satisfies $\nu_x \neq 0$. In this case the normal velocity $V^{\mathcal{K}_r}_{(t,x)}$ of $\mathcal{K}_r$ at the point $(t, x) \in \partial\mathcal{K}_r$ is given by $V^{\mathcal{K}_r}_{(t,x)} = -\nu_t/|\nu_x|$, where (ν_t, ν_x) is the outward normal to $\mathcal{K}_r$ at (t, x).

We use smooth tubes as "test sets." Namely, we say that the smooth tube $\mathcal{K}_r$ is *externally tangent* to a tube $\mathcal{K}$ at $(t,x) \in \partial\mathcal{K}$ if $\mathcal{K}_r$ is defined on some open interval I containing t, and if

$$\mathcal{K}(s) \subset \mathcal{K}_r(s) \quad \forall s \in I \quad \text{and} \quad (t,x) \in \partial\mathcal{K}_r.$$

In the same way, the smooth tube $\mathcal{K}_r$ is said to be *internally tangent* to $\mathcal{K}$ at $(t,x) \in \partial\widehat{\mathcal{K}}$ if $\mathcal{K}_r(s)$ is defined on some open interval I containing t, and if

$$\mathcal{K}_r(s) \subset \mathcal{K}(s) \quad \forall s \in I \quad \text{and} \quad (t,x) \in \partial\mathcal{K}_r.$$

We are now ready to define the viscosity solutions of (4). Recall that the set $\mathcal{D}$ is defined by (5).

DEFINITION 2.1. *Let $\mathcal{K}$ be a tube and $K_0 \subset \mathbb{R}^N$ be an initial position.*

1. *$\mathcal{K}$ is* a viscosity subsolution *to the front propagation problem* (4) *if $\mathcal{K}$ is nondecreasing, left lower semicontinuous, and $\mathcal{K}(0) \in \mathcal{D}$, and if for any smooth tube $\mathcal{K}_r$ externally tangent to $\mathcal{K}$ at some point (t,x), with $\mathcal{K}_r(t) \in \mathcal{D}$ and $t > 0$, we have*

$$V^{\mathcal{K}_r}_{(t,x)} \le h(x, \mathcal{K}_r(t)),$$

where $V^{\mathcal{K}_r}_{(t,x)}$ is the normal velocity of $\mathcal{K}_r$ at (t,x).

We say that $\mathcal{K}$ is a subsolution to the front propagation problem with initial position K_0 if $\mathcal{K}$ is a subsolution and if $\overline{\mathcal{K}}(0) \subset \overline{K_0}$.

2. *$\mathcal{K}$ is a* viscosity supersolution *to the front propagation problem if $\mathcal{K}$ is nondecreasing and $\mathcal{K}(0) \subset \mathcal{D}$, and if for any smooth tube $\mathcal{K}_r$ internally tangent to $\mathcal{K}$ at some point (t,x), with $\mathcal{K}_r(t) \in \mathcal{D}$ and $t > 0$, we have*

$$V^{\mathcal{K}_r}_{(t,x)} \ge h(x, \mathcal{K}_r(t)).$$

We say that $\mathcal{K}$ is a supersolution to the front propagation problem with initial position K_0 if $\mathcal{K}$ is a supersolution and if $\widehat{\mathcal{K}}(0) \subset \mathbb{R}^N \backslash \overline{K_0}$.

3. *Finally, we say that a tube $\mathcal{K}$ is a* viscosity solution *to the front propagation problem (with initial position K_0) if $\mathcal{K}$ is a sub- and a supersolution to the front propagation problem (with initial position K_0).*

Let us point out that, under assumptions (6) and (7), any classical solution is a viscosity solution. The previous definition has been introduced in [1] and was also used in [11].

2.2. Regularity properties of the velocity law for power-law fluids. We now investigate the main regularity properties of the velocity law for power-law fluids. This regularity shall be our guideline for the structure condition when investigating the inclusion principle. Let us recall that the velocity law h for power-law fluids is given by

$$h(x,\Omega) = |\nabla u(x)|^{p-1}, \tag{8}$$

where $u : \Omega \to \mathbb{R}$ is the solution of the following p.d.e.:

$$\text{(9)} \qquad \begin{cases} \text{(i)} & -\operatorname{div}(|\nabla u|^{p-2}\nabla u) = f & \text{in } \Omega\backslash S, \\ \text{(ii)} & u = g & \text{on } \partial S, \\ \text{(iii)} & u = 0 & \text{on } \partial\Omega, \end{cases}$$

and (i) is understood in the sense of distributions. The set S is a fixed source and we always assume above that $S \subset\subset \Omega(t)$. If $S = \emptyset$, then we omit condition (ii) in (9). We assume that

$$(10)\qquad \begin{cases} \text{(i)} & S \subset \mathbb{R}^N \text{ is bounded and equal to the closure of an open set with a } \mathcal{C}^2 \text{ boundary;} \\ \text{(ii)} & f : \mathbb{R}^N \to \mathbb{R} \text{ is continuous and bounded and either} \\ & \bullet f > 0 \text{ on } \mathbb{R}^N \text{ and } f \text{ is locally Lipschitz continuous, or} \\ & \bullet S \neq \emptyset \text{ and } f = 0 \text{ on } \mathbb{R}^N; \\ \text{(iii)} & g : S \to (0, +\infty) \text{ is } \mathcal{C}^{1,\beta} \text{ (for some } \beta \in (0,1)). \end{cases}$$

Remark 2.2. Following [21], $h(x, \Omega)$ is well defined as soon as Ω has a "smooth" boundary. More precisely, it is proved in [21] that, if Ω has a $\mathcal{C}^{1,\beta}$ boundary and if $S \subset\subset \Omega$, then the solution u is $\mathcal{C}^{1,\alpha}$ for some $\alpha \in (0, \beta)$. Moreover, the $\mathcal{C}^{1,\alpha}$ norm of u is bounded by a constant which depends only on $\|f\|_\infty$, $|g|_{1,\beta}$, p and on the $\mathcal{C}^{1,\beta}$ norm of the mapping which locally flattens the boundary of $\Omega \backslash S$.

Let us recall that $\mathcal{D}$ is defined in (5) by

$$\mathcal{D} = \{\Omega \subset \mathbb{R}^N,\ \Omega \text{ open bounded and } S \subset\subset \Omega\}.$$

The following proposition is a straightforward application of the maximum principle.

Proposition 2.3. *The function h defined by* (8) *is nonnegative and nondecreasing with respect to the inclusion, i.e., it satisfies* (6) *and* (7).

Proof of Proposition 2.3. The velocity law h is nonnegative by definition. In order to prove (7), let $\Omega_1 \in \mathcal{D}$ and $\Omega_2 \in \mathcal{D}$ be open with a $\mathcal{C}^{1,1}$ boundary, and assume that $\Omega_1 \subset \Omega_2$ and $x \in \Omega_1 \cap \partial\Omega_2$. Let u_1 and u_2 be solutions to (9) with Ω replaced by Ω_1 and Ω_2, respectively. Then u_1 and u_2 are nonnegative, and the maximum principle states that

$$u_1 \leq u_2 \qquad \text{in } \Omega_1.$$

Let ν be the outward unit normal to Ω_1 and Ω_2 at x. Since u_1 and u_2 vanish on $\partial\Omega_1$ and $\partial\Omega_2$, respectively, and are nonnegative inside Ω_1 and Ω_2, we have $\nabla u_1(x) = -|\nabla u_1(x)|\nu$ and $\nabla u_2(x) = -|\nabla u_2(x)|\nu$. Since moreover $u_1 \leq u_2$ in Ω_1, we get, for $h > 0$ small,

$$u_1(x - h\nu) = h|\nabla u_1(x)| + h\epsilon_1(h) \leq u_2(x - h\nu) = h|\nabla u_2(x)| + h\epsilon_2(h),$$

where $\epsilon_1(h) \to 0$ and $\epsilon_2(h) \to 0$ as $h \to 0^+$. Therefore, $|\nabla u_1(x)| \leq |\nabla u_2(x)|$, which proves (7). □

In order to describe the continuity properties of h, let us first recall that, if Ω is an open set with a $\mathcal{C}^{1,1}$ boundary, then the signed distance $\mathbf{d}$ defined by

$$\mathbf{d}(x) = \begin{cases} d_\Omega(x) & \text{if } x \notin \Omega, \\ -d_{\partial\Omega}(x) & \text{otherwise} \end{cases}$$

is $\mathcal{C}^{1,1}$ in a neighborhood of $\partial\Omega$. We say that a sequence Ω_n of sets with a $\mathcal{C}^{1,1}$ boundary converges to some set Ω with a $\mathcal{C}^{1,1}$ boundary if $\overline{\Omega_n}$ converges to $\overline{\Omega}$ and $\partial\Omega_n$ converges to $\partial\Omega$ for the Hausdorff distance, and if there is an open neighborhood $\mathcal{O}$ of $\partial\Omega$ such that, if $\mathbf{d}$ (resp., $\mathbf{d}_n$) is the signed distance to Ω (resp., to Ω_n), then $(\mathbf{d}_n)$ and $(\nabla\mathbf{d}_n)$ converge uniformly to $\mathbf{d}$ and $\nabla\mathbf{d}$ on $\mathcal{O}$ and if the L^∞ norms of $(\nabla^2\mathbf{d}_n)$ are uniformly essentially bounded on $\mathcal{O}$.

PROPOSITION 2.4. *The velocity h is sequentially continuous with respect to its arguments, i.e.,*

$$\begin{array}{c} \text{if } \Omega_n \text{ and } \Omega \in \mathcal{D} \text{ are open subsets of } \mathbb{R}^N \text{ with a } \mathcal{C}^{1,1} \text{ boundary,} \\ \text{such that } \Omega_n \text{ converge to } \Omega, \text{ if } x_n \in \partial\Omega_n \text{ converge to } x \in \partial\Omega, \\ \text{then } \lim_n h(x_n, \Omega_n) = h(x, \Omega). \end{array} \tag{11}$$

Proof of Proposition 2.4. Note that, if $\Omega \in \mathcal{D}$, then, for n large enough, the sets Ω_n also belongs to $\mathcal{D}$. The rest of the proposition is a straightforward application of the regularity results of [21] recalled in Remark 2.2. □

We now state some estimates on the variations of the mapping $v \to h(x+v, \Omega+v)$ for a set Ω with a smooth boundary and $x \in \partial\Omega$. The key point is that such an estimate has to be independent of the regularity of Ω. Here and below we set

$$S_r = \{x \in \mathbb{R}^N,\ d_S(x) \le r\}.$$

PROPOSITION 2.5. *Let $R > 0$ be some large constant and $r > 0$ be sufficiently small so that S_r has a $\mathcal{C}^2$ boundary. There is a constant $\lambda > 1/r$, such that, for any bounded set Ω with a $\mathcal{C}^{1,1}$ boundary such that $S_r \subset \Omega$ and $\Omega \subset B(0, R-r)$, for any $v \in \mathbb{R}^N$ with $|v| < 1/\lambda$ and any $x \in \partial\Omega$, we have*

$$\Omega + v \in \mathcal{D} \qquad \text{and} \qquad h(x+v, \Omega+v) \ge (1 - \lambda|v|)h(x, \Omega).$$

Proof of Proposition 2.5. We give the proof in the case $S \neq \emptyset$ and $f > 0$, the proof for the other cases being similar. Since $f > 0$ and f is Lipschitz continuous, we can find a constant $C_1 > 0$ such that

$$\forall x, y \in \mathbb{R}^N, \text{ with } |x|, |y| \le R,\ f(x) \ge f(y)(1 - C_1|x-y|). \tag{12}$$

Let u^+ and u_r^- be, respectively, the solutions of

$$\begin{cases} -\mathrm{div}(|\nabla u^+|^{p-2}\nabla u^+) = f & \text{in } B(0,R)\backslash S, \\ u^+ = g & \text{on } \partial S, \\ u^+ = 0 & \text{on } \partial B(0,R) \end{cases}$$

and

$$\begin{cases} -\mathrm{div}(|\nabla u_r^-|^{p-2}\nabla u_r^-) = f & \text{in } Int(S_r)\backslash S, \\ u_r^- = g & \text{on } \partial S, \\ u_r^- = 0 & \text{on } \partial S_r. \end{cases}$$

Since g is $\mathcal{C}^{1,\beta}$, and since S, S_r, and $B(0,R)$ have a $\mathcal{C}^2$ boundary, the functions u^+ and u_r^- belong to $\mathcal{C}^{1,\alpha}(\overline{\Omega})$ (for some $\alpha \in (0,\beta)$) and there is some constant $C_2 > 0$ such that u^+ and u_r^- are C_2-Lipschitz continuous, whence

$$\forall x \in S_r\backslash S\ ,\ u_r^-(x) \ge u^+(x) - 2C_2 d_S(x), \tag{13}$$

because $u^+ = u_r^-$ on ∂S. We now choose $\lambda = \max\{6C_2(p-1)/m, C_1, 2/r\}$, where $m = \min_{S_r\backslash S} u^+$. Note that m is positive.

Let $\Omega \subset \mathbb{R}^N$ be some open bounded set with a $\mathcal{C}^{1,1}$ boundary, such that $S_r \subset \Omega$ and $\Omega \subset B(0, R-r)$, and let $v \in \mathbb{R}^N$ with $|v| < 1/\lambda$. Let u be the solution of (9) with Ω and u_v be the solution of (9) with $\Omega+v$ in place of Ω. From the maximum principle,

we have $u_r^- \le u \le u^+$ and $u_r^- \le u_v \le u^+$ on $S_r\backslash S$ because $S_r \subset\subset \Omega \subset\subset B(0,R)$ and $S_r \subset\subset \Omega + v \subset\subset B(0,R)$ from the choice of $\lambda > 1/r$ and v.

Since $S_r \subset \Omega$ and $|v| < r$, we have $S \subset\subset \Omega$. We claim that

$$\frac{1}{(1-\lambda|v|)^{1/(p-1)}} u_v(x+v) \ge u(x) \qquad \forall x \in \Omega\backslash S_{|v|}. \tag{14}$$

For proving this claim, let us set $w(x) = \frac{1}{(1-\lambda|v|)^{1/(p-1)}} u_v(x+v)$ for $x \in \Omega\backslash S_{|v|}$ and let us show that

$$w \ge u \text{ on } \partial S_{|v|} \quad \text{and} \quad -\operatorname{div}(|\nabla w|^{p-2}\nabla w) \ge f \text{ on } \Omega\backslash S_{|v|}. \tag{15}$$

We have, for any $x \in \partial S_{|v|}$,

$$(1-\lambda|v|)^{1/(p-1)} w(x) = u_v(x+v) \ge u_r^-(x+v) \ge u_r^-(x) - C_2|v| \ge u^+(x) - 3C_2|v|$$

because u_r^- is C_2-Lipschitz continuous and thanks to (13). Thus,

$$(1-\lambda|v|)^{1/(p-1)} w(x) \ge u^+(x) - 3C_2|v| \ge (1-\lambda|v|)^{1/(p-1)} u^+(x)$$

because $u^+ \ge m$ in $S_r\backslash S$, $|v| < 1/\lambda$, and $\lambda \ge 6C_2(p-1)/m$. So we have proved that $w(x) \ge u^+(x) \ge u(x)$ for any $x \in \partial S_{|v|}$.

From the definition of C_1 in (12) and from the choice of λ, we have, for any $x \in \Omega\backslash S_{|v|}$,

$$-\operatorname{div}(|\nabla u_v(x+v)|^{p-2}\nabla u_v(x+v)) = f(x+v) \ge f(x)(1-\lambda|v|).$$

Thus,

$$-\operatorname{div}(|\nabla w(x)|^{p-2}\nabla w(x)) \ge f(x) \qquad \forall x \in \Omega\backslash S_{|v|}.$$

So (15) is proved, which entails (14). In particular, at any point $x \in \partial\Omega$, we have, since $w \ge u$ in $\Omega\backslash S_{|v|}$ and $w = u$ in $\partial\Omega$,

$$\begin{aligned} h(x,\Omega) &= |\nabla u(x)|^{p-1} \le |\nabla w(x)|^{p-1} \\ &= \frac{1}{1-\lambda|v|} |\nabla u_v(x+v)|^{p-1} = \frac{1}{1-\lambda|v|} h(x+v, \Omega+v). \end{aligned}$$ □

In order to prove the global existence of the solution, we need to control the growth of h, as shown in the following proposition.

PROPOSITION 2.6. *There are constants $r_0 > 0$ and $\sigma > 0$ such that*

$$\forall r \ge r_0,\ \forall x \in \partial B(0,r), \qquad h(x, B(0,r)) \le \sigma r. \tag{16}$$

Moreover, the constants r_0 and σ depend only on p, S, $\|f\|_\infty$, and $\|g\|_\infty$.

Proof of Proposition 2.6. Let us fix $r_0 > 0$ such that $S \subset\subset B(0, r_0/2^{\frac{(p-1)}{p}})$ and $\kappa = \max\{\frac{2\|g\|_\infty}{r_0^{p/(p-1)}}, \frac{\|f\|_\infty^{1/(p-1)}(p-1)}{N^{1/(p-1)}p}\}$. Let $r \ge r_0$ and u be the solution to

$$\begin{cases} -\operatorname{div}(|\nabla u|^{p-2}\nabla u) = f & \text{in } B(0,r)\backslash S, \\ u = g & \text{on } \partial S, \\ u = 0 & \text{on } \partial B(0,r). \end{cases}$$

We claim that $u \le w$ on $B(0,r)\backslash S$, where $w(x) = -\kappa|x|^{p/(p-1)} + \kappa r^{p/(p-1)}$ and $u = w$ on $\partial B(0,r)$. Indeed, $-\mathrm{div}(|\nabla w|^{p-2}\nabla w) = \kappa^{p-1}N[p/(p-1)]^{p-1} \ge f$ in $B(0,r)\backslash S$. Since $S \subset B(0, r_0/2^{(p-1)/p})$ we also have

$$\forall x \in \partial S, \quad w(x) \ge -\kappa \frac{r_0^{p/(p-1)}}{2} + \kappa r_0^{p/(p-1)} \ge \|g\|_\infty \ge g(x).$$

Finally, $u = w = 0$ on $\partial B(0,r)$ by construction. So $u \le w$ on $B(0,r)\backslash S$ and $u = w$ on $\partial B(0,r)$. This entails that $h(x, B(0,r)) = |\nabla u|^{p-1} \le |\nabla w|^{p-1} = \kappa^{p-1}[p/(p-1)]^{p-1} r$ for any $x \in \partial B(0,r)$, whence the result with $\sigma = \kappa^{p-1}[p/(p-1)]^{p-1}$. □

2.3. A preliminary result. We state here an equivalent definition for solutions of (4). This formulation is introduced in [11].

Let us set, for any compact set K, $x \in \partial K$, and $\nu \in \mathbb{R}^N$, $\nu \neq 0$,

$$h^\sharp(x, K, \nu) = \inf\{h(x, \Omega)\}, \tag{17}$$

where the infimum is taken over the sets $\Omega \in \mathcal{D}$ with a $\mathcal{C}^{1,1}$ boundary, such that $K \subset \overline{\Omega}$, $x \in \partial\Omega$ and ν is an outward normal to Ω at x. In the same way, we set

$$h^\flat(x, K, \nu) = \sup\{h(x, \Omega)\}, \tag{18}$$

where the supremum is taken over the sets $\Omega \in \mathcal{D}$ with a $\mathcal{C}^{1,1}$ boundary, such that $\Omega \subset K$, and $x \in \partial\Omega$ and ν is an outward normal to Ω at x.

We set $h^\sharp(x, K, \nu) = +\infty$ or $h^\flat(x, K, \nu) = -\infty$ if there is no set Ω with the required properties.

If A is a subset of some finite dimensional space and x belongs to A, we say that a vector ν is *a proximal normal* to A at x if the distance of $x + \nu$ to A is equal to $|\nu|$.

PROPOSITION 2.7. *Let $h = h(x, \Omega)$ satisfy* (6), (7), *and* (11). *Let $\mathcal{K}$ be a nondecreasing tube with $\mathcal{K}(0) \in \mathcal{D}$.*

Then $\mathcal{K}$ is a subsolution of the front propagation problem for h if and only if $\overline{\mathcal{K}}$ is left lower semicontinuous and if, for any $(t,x) \in \mathcal{K}$ with $t > 0$, for any proximal normal (ν_t, ν_x) to $\mathcal{K}$ at (t,x) such that $\nu_x \neq 0$, we have

$$-\frac{\nu_t}{|\nu_x|} \le h^\sharp(x, \mathcal{K}(t), \nu_x).$$

In the same way, the tube $\mathcal{K}$ is a supersolution of the front propagation problem for h if and only if, for any $(t,x) \in \widehat{\mathcal{K}}$ with $t > 0$, for any proximal normal (ν_t, ν_x) to $\widehat{\mathcal{K}}$ at (t,x) such that $\nu_x \neq 0$, we have

$$\frac{\nu_t}{|\nu_x|} \ge h^\flat(x, \mathcal{K}(t), -\nu_x).$$

Proof of Proposition 2.7. The proof is completely similar to that of Proposition 2.2 of [11]. The only difference is that our construction yields some approximation of the tube $\mathcal{K}$. We give only the main arguments for the case of subsolutions, the case of supersolutions being symmetrical. For any $\epsilon > 0$, let us set

$$\mathcal{K}^\epsilon = \{(t,x) \in \mathbb{R}^+ \times \mathbb{R}^N \mid \exists (s,y) \in \overline{\mathcal{K}} \text{ with } (t-s)^2 + |x-y|^2 \le \epsilon^2\}.$$

In [11], since h was translation invariant, $\mathcal{K}^\epsilon$ was also a subsolution of the front propagation problem. Instead here we have the following lemma.

LEMMA 2.8. *If $\mathcal{K}$ is a subsolution of the front propagation problem for h, then for any $\epsilon > 0$, $\mathcal{K}^\epsilon$ is a subsolution for h^ϵ on the time interval $[\epsilon, +\infty)$, where*

$$h^\epsilon(x, \Omega) = \sup_{|v| \le \epsilon} h(x + v, \Omega + v)$$

is defined for any bounded open set Ω with a $\mathcal{C}^{1,1}$ boundary such that $S_\epsilon \subset\subset \Omega$ and $x \in \partial K$ (where S is the compact set in the definition of $\mathcal{D}$; see (5)).

Once Lemma 2.8 is established, we can complete the proof of the proposition as in [11] by noticing that $h^\epsilon \to h$ as $\epsilon \to 0^+$, thanks to (11). □

Proof of Lemma 2.8. A straightforward application of the definition of $\mathcal{K}^\epsilon$ shows that $\mathcal{K}^\epsilon$ is nondecreasing and left lower semicontinuous. Moreover, since $S \subset\subset Int(\mathcal{K}(0)) \subset Int(\mathcal{K}(\epsilon))$, we have $S_\epsilon \subset\subset \mathcal{K}^\epsilon(\epsilon)$.

Let $\mathcal{K}_r$ be a smooth tube which is externally tangent to $\mathcal{K}^\epsilon$ at some point (t, x), with $t > \epsilon$. Since (t, x) belongs to the boundary of $\mathcal{K}^\epsilon$, there is some $(s, y) \in \overline{\mathcal{K}}$ such that

$$(t - s)^2 + |x - y|^2 = \epsilon^2.$$

Let us notice that, since $t > \epsilon$, we have $s > 0$. Now it is easy to check that the tube $\mathcal{K}_r - ((t, x) - (s, y))$ is externally tangent to $\mathcal{K}$ at (s, y). Since $\mathcal{K}$ is a subsolution, we have

$$V_{(s,y)}^{\mathcal{K}_r - ((t,x)-(s,y))} \le h(y, \mathcal{K}_r(t) - (x - y)),$$

where $V_{(s,y)}^{\mathcal{K}_r - ((t,x)-(s,y))}$ is the outward normal velocity of the smooth tube $\mathcal{K}_r - ((t, x) - (s, y))$ at (s, y). Using the definition of h^ϵ, this leads to

$$V_{(t,x)}^{\mathcal{K}_r} = V_{(s,y)}^{\mathcal{K}_r - ((t,x)-(s,y))} \le h(y, \mathcal{K}_r(t) - (x - y)) \le h^\epsilon(x, \mathcal{K}_r(t))$$

because $|x - y| \le \epsilon$. □

3. The inclusion principle. The aim of this part is to state and prove the inclusion principle for the generalized solutions of our problem.

3.1. Statement of the result. We need the following structure conditions on h:

1. $h = h(x, \Omega)$ is nonnegative and nondecreasing with respect to Ω (i.e., (6) and (7) hold),
2. h is sequentially continuous (i.e., satisfies (11)),
3. h has the following regularity property: For any $R > 0$ sufficiently large and $r > 0$ sufficiently small so that S_r has a $\mathcal{C}^2$ boundary, there are constants $\lambda > 1/r$ and $C > 0$, such that, for any compact set K with a $\mathcal{C}^{1,1}$ boundary such that $S_r \subset Int(K)$ and $K \subset B(0, R - r)$, for any $v \in \mathbb{R}^N$ with $|v| < 1/\lambda$ and any $x \in \partial K$, we have

$$(19)\quad K + v \in \mathcal{D} \qquad \text{and} \qquad h(x + v, K + v) \ge (1 - \lambda|v|)h(x, K) - C|v|.$$

Let us recall that, if $\mathcal{K}$ is a subset of $[0, +\infty) \times \mathbb{R}^N$, we set

$$\widehat{\mathcal{K}} = \overline{[0, +\infty) \times \mathbb{R}^N \backslash \mathcal{K}}.$$

THEOREM 3.1 (inclusion principle). *Under the above assumptions on h, let $\mathcal{K}_1$ be a subsolution of the front propagation problem on the interval $[0, T)$ for some $T > 0$*

and $\mathcal{K}_2$ be a supersolution on $[0,T)$. If $\mathcal{K}_1(t)$ and $\mathcal{K}_2(t)$ are nonempty for $t \in [0,T)$ and if

$$\overline{\mathcal{K}_1}(0) \cap \widehat{\mathcal{K}_2}(0) = \emptyset,$$

then

$$\forall t \in [0,T), \qquad \overline{\mathcal{K}_1}(t) \cap \widehat{\mathcal{K}_2}(t) = \emptyset.$$

Remarks.

1. Let K_1 and K_2 be bounded subsets of $\mathbb{R}^N$ such that $\overline{K_1} \subset Int(K_2)$. If $\mathcal{K}_1$ is a subsolution with initial condition K_1 and $\mathcal{K}_2$ is a supersolution with initial position K_2, then the assumption of the theorem holds: $\overline{\mathcal{K}_1}(0) \cap \widehat{\mathcal{K}_2}(0) = \emptyset$.
2. The statement $\overline{\mathcal{K}_1}(t) \cap \widehat{\mathcal{K}_2}(t) = \emptyset$ implies that $\overline{\mathcal{K}_1}(t) \subset Int(\mathcal{K}_2(t))$.
3. *Remarks on the structure conditions.* Thanks to Propositions 2.3, 2.4, and 2.5, our structure condition holds for the Hele–Shaw problem, where h is defined by (8). It is also satisfied for nonlocal evolution laws of the form

$$h(x,\Omega) = \int_\Omega \rho(x,y)dy,$$

where ρ is some nonnegative smooth and compactly supported function. This evolution equation is treated in particular in [10] and in [13] (the latter in the more general case of Neumann-type boundary conditions).

3.2. Proof of Theorem 3.1. From the definition of subsolutions, we can assume that $\mathcal{K}_1$ has a closed graph: $\overline{\mathcal{K}_1} = \mathcal{K}_1$. The main step of the proof amounts to showing that, for any $\gamma > 1$, such that $S_{\gamma-1} \subset Int(\mathcal{K}_1(0))$,

$$\forall t \in [0,T), \qquad \mathcal{K}_1(t) \cap \widehat{\mathcal{K}_2}(\gamma t) = \emptyset. \tag{20}$$

We explain how to obtain Theorem 3.1 from (20) at the very end of the proof.

For showing (20), we argue by contradiction, by assuming that there is some $\gamma > 1$ with $S_{\gamma-1} \subset Int(\mathcal{K}_1(0))$ and some $T^* \in [0,T)$ such that

$$\mathcal{K}_1(T^*) \cap \widehat{\mathcal{K}_2}(\gamma T^*) \neq \emptyset. \tag{21}$$

Since $\mathcal{K}_1(0) \cap \widehat{\mathcal{K}_2}(0) = \emptyset$, we have $T^* > 0$. We now introduce several notations: Let $R > 0$ be sufficiently large so that $\mathcal{K}_1(T) \subset B(0, R-(\gamma-1))$ and $\mathcal{K}_2(T) \subset B(0, R-(\gamma-1))$. We denote by λ and C the constants defined in (19) for R and $r := \gamma - 1 > 0$. Let us recall that $\lambda > 1/r$ and that, for any compact set K with a $\mathcal{C}^{1,1}$ boundary such that $S_r \subset Int(K)$ and $K \subset B(0, R-r)$, for any $v \in \mathbb{R}^N$ with $|v| < 1/\lambda$ and any $x \in \partial K$, we have

$$h(x+v, K+v) \geq (1-\lambda|v|)h(x,K) - C|v|. \tag{22}$$

We also set

$$\kappa = \gamma C + 1. \tag{23}$$

For any $\epsilon \in (0, \tau_0)$ and any $\sigma \in (0,1]$, we set

$$\mathcal{K}_1^{\epsilon,\sigma} = \left\{ (t,x) \in \mathbb{R}^+ \times \mathbb{R}^N \mid \exists (s,y) \in \mathcal{K}_1 \text{ with } \frac{1}{\sigma^2}(t-s)^2 + |x-y|^2 \leq \epsilon^2 e^{-2\kappa s} \right\} \tag{24}$$

and

$$\widehat{\mathcal{K}}_2^{\epsilon,\sigma} = \left\{ (t,x) \in \mathbb{R}^+ \times \mathbb{R}^N \mid \exists (s,y) \in \widehat{\mathcal{K}_2} \text{ with } \frac{1}{\sigma^2}(t-s)^2 + |x-y|^2 \le \epsilon^2 e^{-2\kappa s} \right\} \tag{25}$$

and

$$T^{\epsilon,\sigma,\gamma} = \min\{t \ge \epsilon \mid \mathcal{K}_1^{\epsilon,\sigma}(t) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma t) \ne \emptyset\}.$$

Let us point out that

$$T^{\epsilon,\sigma,\gamma} \le T^* \tag{26}$$

because assumption (21) implies that $\mathcal{K}_1(T^*) \cap \widehat{\mathcal{K}_2}(\gamma T^*) \ne \emptyset$ and $\mathcal{K}_1(t) \subset \mathcal{K}_1^{\epsilon,\sigma}(t)$ and $\widehat{\mathcal{K}_2}(t) \subset \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(t)$.

Let us define Π_1^σ and Π_2^σ, the projections on the sets $\mathcal{K}_1$ and $\widehat{\mathcal{K}_2}$, as follows: $\forall \sigma \in (0,1]$,

$$\Pi_1^\sigma(t,x) = \left\{ (s_1,y_1) \in \mathcal{K}_1 \;\middle|\; \begin{array}{l} \frac{1}{\sigma^2}(t-s_1)^2 + |x-y_1|^2 \\ = \inf_{(s,y)\in\mathcal{K}_1} \frac{1}{\sigma^2}(t-s)^2 + |x-y|^2 \end{array} \right\}$$

and

$$\Pi_2^\sigma(t,x) = \left\{ (s_2,y_2) \in \widehat{\mathcal{K}_2} \;\middle|\; \begin{array}{l} \frac{1}{\sigma^2}(t-s_2)^2 + |x-y_2|^2 \\ = \inf_{(s,y)\in\widehat{\mathcal{K}_2}} \frac{1}{\sigma^2}(t-s)^2 + |x-y|^2 \end{array} \right\}.$$

PROPOSITION 3.2. *One can choose ϵ and σ sufficiently small so that, for any $x \in \mathcal{K}_1^{\epsilon,\sigma}(T^{\epsilon,\sigma,\gamma}) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma T^{\epsilon,\sigma,\gamma})$, for any $(s_1,y_1) \in \Pi_1^\sigma(T^{\epsilon,\sigma,\gamma},x)$, and any $(s_2,y_2) \in \Pi_2^\sigma(\gamma T^{\epsilon,\sigma,\gamma},x)$, we have $y_1 \ne x$, $y_2 \ne x$, $s_1 > 0$, and $s_2 > 0$.*

Proof of Proposition 3.2. Let us first prove that there is some positive ϵ_0 such that

$$\text{for any } \epsilon \in (0,\epsilon_0) \text{ and any } \sigma \in (0,1], \text{ we have } T^{\epsilon,\sigma,\gamma} > \epsilon. \tag{27}$$

Since $\mathcal{K}_1(0) \cap \widehat{\mathcal{K}_2}(0) = \emptyset$ and the sets $\mathcal{K}_1$ and $\widehat{\mathcal{K}_2}$ are closed in $\mathbb{R}^+ \times \mathbb{R}^N$, there is some $\tau > 0$ such that

$$\forall 0 \le s,t \le \tau,\ \mathcal{K}_1(s) \cap \widehat{\mathcal{K}_2}(t) = \emptyset.$$

Then set

$$\theta = \min\{|y_1 - y_2| \mid y_1 \in \mathcal{K}_1(s),\ y_2 \in \widehat{\mathcal{K}_2}(t),\ 0 \le s,t \le \tau\}; \tag{28}$$

then $\theta > 0$. Set $\epsilon_0 = \min\{\frac{\theta}{2}, \frac{\tau}{1+\gamma}\}$. We claim that, for any $\epsilon \in (0,\epsilon_0)$ and for any $\sigma \in (0,1]$, we have

$$\mathcal{K}_1^{\epsilon,\sigma}(\epsilon) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma\epsilon) = \emptyset. \tag{29}$$

It is clearly enough to prove the result for $\sigma = 1$. We argue by contradiction. Suppose that, contrary to our claim, there is some $x \in \mathcal{K}_1^{\epsilon,1}(\epsilon) \cap \widehat{\mathcal{K}}_2^{\epsilon,1}(\gamma\epsilon)$. Then there is some $(s_1,y_1) \in \mathcal{K}_1$ and some $(s_2,y_2) \in \widehat{\mathcal{K}_2}$ such that

$$|(\epsilon,x) - (s_1,y_1)| \le \epsilon e^{-\kappa s_1} \text{ and } |(\gamma\epsilon,x) - (s_2,y_2)| \le \epsilon e^{-\kappa s_2}.$$

This implies, on the one hand, that $|y_1 - y_2| \le 2\epsilon < \theta$ and, on the other hand, that

$$0 \le s_1 \le 2\epsilon < \tau \text{ and } 0 \le s_2 \le (1+\gamma)\epsilon < \tau,$$

which is in contradiction with the definition of θ in (28). Thus, (29) is proved, which obviously implies that $T^{\epsilon,\sigma,\gamma} > \epsilon$, i.e., (27) holds.

From now on we fix $\epsilon \in (0, \epsilon_0)$. Let us first notice that $T^{\epsilon,\sigma,\gamma}$ is nondecreasing with respect to σ and is bounded by T^* thanks to (26). Let us set $\bar{t} = \lim_{\sigma \to 0^+} T^{\epsilon,\sigma,\gamma}$. Let us also define

$$\mathcal{K}_1^{\epsilon,0} = \{(t,x) \in \mathbb{R}^+ \times \mathbb{R}^N \mid \exists y \in \mathcal{K}_1(t) \text{ with } |x-y| \le \epsilon e^{-\kappa t}\}$$

and

$$\widehat{\mathcal{K}}_2^{\epsilon,0} = \{(t,x) \in \mathbb{R}^+ \times \mathbb{R}^N \mid \exists y \in \widehat{\mathcal{K}_2}(t) \text{ with } |x-y| \le \epsilon e^{-\kappa s}\}.$$

It is easily checked that

$$\bigcap_{\sigma \in (0,1]} \mathcal{K}_1^{\epsilon,\sigma} = \mathcal{K}_1^{\epsilon,0} \qquad \text{and} \qquad \bigcap_{\sigma \in (0,1]} \widehat{\mathcal{K}}_2^{\epsilon,\sigma} = \widehat{\mathcal{K}}_2^{\epsilon,0}. \tag{30}$$

Moreover, $\mathcal{K}_1^{\epsilon,0}$ and $\widehat{\mathcal{K}}_2^{\epsilon,0}$ are closed since $\mathcal{K}_1$ and $\widehat{\mathcal{K}_2}$ are. Hence, from the definition of $T^{\epsilon,\sigma,\gamma}$ and of $\bar{t}$, we have

$$\mathcal{K}_1^{\epsilon,0}(\bar{t}) \cap \widehat{\mathcal{K}}_2^{\epsilon,0}(\gamma\bar{t}) \neq \emptyset\ .$$

Let us also point out that, for any $t \in (0, \bar{t})$,

$$\mathcal{K}_1^{\epsilon,0}(t) \cap \widehat{\mathcal{K}}_2^{\epsilon,0}(\gamma t) = \emptyset \tag{31}$$

because, $\mathcal{K}_1^{\epsilon,0}(t) \subset \mathcal{K}_1^{\epsilon,\sigma}(t)$, $\widehat{\mathcal{K}}_2^{\epsilon,0}(\gamma t) \subset \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma t)$, and $\mathcal{K}_1^{\epsilon,\sigma}(t) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma t) = \emptyset$ as soon as $T^{\epsilon,\sigma,\gamma} > t$.

The next step of the proof amounts to showing that

$$\forall x \in \mathcal{K}_1^{\epsilon,0}(\bar{t}) \cap \widehat{\mathcal{K}}_2^{\epsilon,0}(\gamma\bar{t}),\ d_{\mathcal{K}_1(\bar{t})}(x) = \epsilon e^{-\kappa\bar{t}} \text{ and } d_{\widehat{\mathcal{K}_2}(\gamma\bar{t})}(x) = \epsilon e^{-\gamma\kappa\bar{t}}. \tag{32}$$

For proving this, we argue by contradiction, by assuming (for instance) that $d_{\mathcal{K}_1(\bar{t})}(x) < \epsilon e^{-\kappa\bar{t}}$. Then there is some $y_1 \in \mathcal{K}_1(\bar{t})$ such that $|y_1 - x| < \epsilon e^{-\kappa\bar{t}}$. Let also $y_2 \in \widehat{\mathcal{K}_2}(\gamma\bar{t})$ be such that $|y_2 - x| \le \epsilon e^{-\gamma\kappa\bar{t}}$. Since $\mathcal{K}_1$ is a subsolution, it is left lower semicontinuous. Thus, for any sequence $t_k \to \bar{t}^-$, there is some $y_1^k \to y_1$ with $y_1^k \in \mathcal{K}_1(t_k)$. In the same way, since $\mathcal{K}_2$ is a supersolution, $\widehat{\mathcal{K}_2}$ is left lower semicontinuous, and there is a sequence $y_2^k \in \widehat{\mathcal{K}_2}(\gamma t_k)$ which converges to y_2. Since $|y_1 - y_2| < \epsilon e^{-\kappa\bar{t}} + e^{-\gamma\kappa\bar{t}}$, for k large enough we still have $|y_1^k - y_2^k| \le \epsilon(e^{-\kappa t_k} + e^{-\gamma\kappa\bar{t}})$. Then it is easy to find some point $x_k \in [y_1^k, y_2^k]$ such that $|y_1^k - x_k| \le \epsilon e^{-\kappa t_k}$ and $|y_2^k - x_k| \le \epsilon e^{-\gamma\kappa\bar{t}}$, i.e., $x_k \in \mathcal{K}_1^{\epsilon,0}(t_k) \cap \widehat{\mathcal{K}}_2^{\epsilon,0}(\gamma t_k)$. This is in contradiction with (31). Hence, claim (32) is proved.

From this claim we deduce that, for $\epsilon' = \epsilon/4 < \epsilon$, we have

$$\mathcal{K}_1^{\epsilon',0}(\bar{t}) \cap \widehat{\mathcal{K}}_2^{\epsilon,0}(\gamma\bar{t}) = \emptyset.$$

Hence, there is some $\sigma_0 \in (0,1)$ such that

$$\mathcal{K}_1^{\epsilon',\sigma_0}(\bar{t}) \cap \widehat{\mathcal{K}_2}^{\epsilon,\sigma_0}(\gamma\bar{t}) = \emptyset$$

because of (30). Since $\mathcal{K}_1^{\epsilon',\sigma_0}$ and $\widehat{\mathcal{K}_2}^{\epsilon,\sigma_0}$ are closed and since $T^{\epsilon,\sigma,\gamma} \to \bar{t}$ as $\sigma \to 0^+$, we have, for any $\sigma > 0$ sufficiently small, that

$$\mathcal{K}_1^{\epsilon',\sigma_0}(T^{\epsilon,\sigma,\gamma}) \cap \widehat{\mathcal{K}_2}^{\epsilon,\sigma_0}(\gamma T^{\epsilon,\sigma,\gamma}) = \emptyset. \tag{33}$$

Let $x \in \mathcal{K}_1^{\epsilon,\sigma}(T^{\epsilon,\sigma,\gamma}) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(T^{\epsilon,\sigma,\gamma})$. Then, from (33), $x \notin \mathcal{K}_1^{\epsilon',\sigma_0}(T^{\epsilon,\sigma,\gamma})$. Therefore, if $(s_1, y_1) \in \Pi_1^{\sigma}(T^{\epsilon,\sigma,\gamma}, x)$, we have

$$\frac{1}{\sigma^2}(T^{\epsilon,\sigma,\gamma} - s_1)^2 + |x - y_1|^2 \le \epsilon^2 e^{-2\kappa s_1}$$

and

$$\frac{1}{\sigma_0^2}(T^{\epsilon,\sigma,\gamma} - s_1)^2 + |x - y_1|^2 > (\epsilon')^2 e^{-2\kappa s_1}.$$

This implies that $x \neq y_1$ as soon as $\sigma < \sigma_0/2$ (recall that $\epsilon' = \epsilon/4$). So we have proved that, for any σ sufficiently small, for any $x \in \mathcal{K}_1^{\epsilon,\sigma}(T^{\epsilon,\sigma,\gamma}) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(T^{\epsilon,\sigma,\gamma})$, and for any $(s_1, y_1) \in \Pi_1^{\sigma}(T^{\epsilon,\sigma,\gamma}, x)$, we have $x \neq y_1$. We can prove in the same way that, for any $(s_2, y_2) \in \Pi_2^{\sigma}(T^{\epsilon,\sigma,\gamma}, x)$, we have $y_2 \neq x$.

Finally, s_1 is positive because the inequality $\frac{1}{\sigma^2}(T^{\epsilon,\sigma,\gamma} - s_1)^2 + |x - y_1|^2 \le \epsilon e^{-2\kappa s_1}$ implies that

$$s_1 \ge T^{\epsilon,\sigma,\gamma} - \sigma\epsilon e^{-\kappa s_1},$$

where the right-hand side is positive thanks to (27). We can prove in the same way that $s_2 > 0$. □

From now on we fix $\epsilon > 0$ and $\sigma > 0$ as in Proposition 3.2 and also sufficiently small so that

$$\epsilon < 1/(2\lambda) \quad \text{and} \quad \frac{1}{(1 - 2\lambda\epsilon)} \le \gamma. \tag{34}$$

Recall that λ is defined at the beginning of the proof. Let x, (s_1, y_1), and (s_2, y_2) be as in Proposition 3.2. For simplicity, we set $t^* = T^{\epsilon,\sigma,\gamma}$.

Let us define, for any two sets U and V, the minimal distance $d(U, V)$ between U and V by

$$d(U, V) = \inf_{x \in U, y \in V} |x - y|.$$

PROPOSITION 3.3. *The point (t^*, x) belongs to the boundary of $\mathcal{K}_1^{\epsilon,\sigma}$, while the point $(\gamma t^*, x)$ belongs to the boundary of $\widehat{\mathcal{K}}_2^{\epsilon,\sigma}$. Moreover,*

$$d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) = |y_1 - y_2|.$$

In particular, $y_1 \in \partial\mathcal{K}_1(s_1)$, $y_2 \in \partial\widehat{\mathcal{K}_2}(s_2)$, and

$$\frac{1}{\sigma^2}(t^* - s_1)^2 + |x - y_1|^2 = \epsilon^2 e^{-2\kappa s_1} \quad \textit{and} \quad \frac{1}{\sigma^2}(\gamma t^* - s_2)^2 + |x - y_2|^2 = \epsilon^2 e^{-\kappa s_2}.$$

Proof of Proposition 3.3. For proving that the point (t^*, x) belongs to the boundary of $\mathcal{K}_1^{\epsilon,\sigma}$, we argue by contradiction by assuming that (t^*, x) belongs to the interior of $\mathcal{K}_1^{\epsilon,\sigma}$. Then, since $\widehat{\mathcal{K}_2}$ is left lower semicontinuous, so is $\widehat{\mathcal{K}}_2^{\epsilon,\sigma}$. Thus, for any

$t_n \to (t^*)^-$, there is some $x_n \to x$ such that $(\gamma t_n, x_n) \in \widehat{\mathcal{K}}_2^{\epsilon,\sigma}$. But, since (t^*, x) belongs to the interior of $\mathcal{K}_1^{\epsilon,\sigma}$, (t_n, x_n) also belongs to $\mathcal{K}_1^{\epsilon,\sigma}$ for n large enough. This is in contradiction with the definition of t^*.

Symmetric arguments show that the point $(\gamma t^*, x)$ belongs to the boundary of $\widehat{\mathcal{K}}_2^{\epsilon,\sigma}$. Let us now prove that $d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) = |y_1 - y_2|$. Since $y_1 \in \mathcal{K}_1(s_1)$ and $y_2 \in \widehat{\mathcal{K}_2}(s_2)$, we have $d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) \le |y_1 - y_2|$. Assume for a while that $d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) < |y_1 - y_2|$. Let $z_1 \in \mathcal{K}_1(s_1)$ and $z_2 \in \widehat{\mathcal{K}_2}(s_2)$ be such that $|z_1 - z_2| < |y_1 - y_2|$. One can choose $\rho \in (0,1)$ such that, if $x_\rho = \rho z_1 + (1-\rho) z_2$, then

$$|z_1 - x_\rho| < |y_1 - x| \text{ and } |z_2 - x_\rho| < |y_2 - x|$$

because $|z_1 - z_2| < |y_1 - y_2| \le |y_1 - x| + |y_2 - x|$. Therefore,

$$\frac{1}{\sigma^2}(t^* - s_1)^2 + |z_1 - x_\rho|^2 < \frac{1}{\sigma^2}(t^* - s_1)^2 + |y_1 - x|^2 \ \le \epsilon^2 e^{-2\kappa s_1}$$

and

$$\frac{1}{\sigma^2}(\gamma t^* - s_2)^2 + |z_2 - x_\rho|^2 < \frac{1}{\sigma^2}(\gamma t^* - s_2)^2 + |y_2 - x|^2 \ \le \epsilon^2.$$

So one can find some $t < t^*$ such that

$$\frac{1}{\sigma^2}(t - s_1)^2 + |z_1 - x_\rho|^2 \le \epsilon^2 e^{-2\kappa s_1} \text{ and } \frac{1}{\sigma^2}(\gamma t - s_2)^2 + |z_2 - x_\rho|^2 \le \epsilon^2,$$

which means that $x_\rho \in \mathcal{K}_1^{\epsilon,\sigma}(t) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma t)$ and $t < t^*$. This is in contradiction with the definition of t^*. Therefore, we have proved that $d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) = |y_1 - y_2|$. □

Let us introduce two new notations:

$$(\nu_t^1, \nu_x^1) = (t^* - s_1 - \epsilon^2 \sigma^2 \kappa e^{-2\kappa s_1}, \sigma^2 (x - y_1)) \tag{35}$$

and

$$(\nu_t^2, \nu_x^2) = (\gamma t^* - s_2 - \epsilon^2 \sigma^2 \kappa e^{-2\kappa s_2}, \sigma^2 (x - y_2)). \tag{36}$$

PROPOSITION 3.4. *There is some $\rho > 0$ such that the vectors $\rho(\nu_t^1, \nu_x^1)$ and $\rho(\nu_t^2, \nu_x^2)$ are proximal normals to $\mathcal{K}_1$ at (s_1, y_1) and to $\widehat{\mathcal{K}_2}$ at (s_2, y_2), respectively.*

Proof of Proposition 3.4. We do the proof only for (ν_t^1, ν_x^1), with the proof for (ν_t^2, ν_x^2) being similar. From Proposition 3.3, (t^*, x) belongs to the boundary of $\mathcal{K}_1^{\epsilon,\sigma}$. Therefore, the set $\mathcal{K}_1$ is contained in the set

$$E = \left\{ (s, y) \in \mathbb{R}^+ \times \mathbb{R}^N, \ \frac{1}{\sigma^2}(t^* - s)^2 + |y - x|^2 \ge \epsilon^2 e^{-2\kappa s} \right\}.$$

From Proposition 3.3 again, the point (s_1, y_1) belongs to the boundary of E. Moreover, E has a smooth boundary in a neighborhood of (s_1, y_1) because the gradient of the map $(s, y) \to \epsilon^2 e^{-2\kappa s} - \frac{1}{\sigma^2}(t^* - s)^2 - |y - x|^2$ at (s_1, y_1) is

$$\left(-2\epsilon^2 \kappa e^{-2\kappa s_1} - \frac{2(s_1 - t^*)}{\sigma^2}, -2(y_1 - x) \right) = \frac{2}{\sigma^2}(\nu_t^1, \nu_x^1),$$

which does not vanish since $y_2 \neq x$ from Proposition 3.2. Therefore, this gradient is, up to some small positive multiplicative constant, a proximal normal to E at the

point (s_1, y_1). Since $\mathcal{K}_1 \subset E$ with $(s_1, y_1) \in \mathcal{K}_1$, it is also a proximal normal to $\mathcal{K}_1$ at (s_1, y_1). Since (ν_t^1, ν_x^1) is proportional to this gradient, the proof is complete. □

Since $\mathcal{K}_1$ is a subsolution and $\rho(\nu_t^1, \nu_x^1)$ is a proximal normal to $\mathcal{K}_1$ at (s_1, y_1), with $\nu_x^1 \neq 0$ and $s_1 > 0$ thanks to Proposition 3.2, Proposition 2.7 states that

$$-\frac{\nu_t^1}{|\nu_x^1|} = -\frac{t^* - s_1 - \epsilon^2\sigma^2\kappa e^{-2\kappa s_1}}{\sigma^2|x - y_1|} \leq h^\sharp(y_1, \mathcal{K}_1(s_1), \nu_x^1). \tag{37}$$

Similarly, since $\mathcal{K}_2$ is a supersolution and $\rho(\nu_t^2, \nu_x^2)$ is a proximal normal to $\widehat{\mathcal{K}_1}$ at (s_2, y_2), with $\nu_x^2 \neq 0$ and $s_2 > 0$, Proposition 2.7 also states that

$$\frac{\nu_t^2}{|\nu_x^2|} = \frac{\gamma t^* - s_2 - \epsilon^2\sigma^2\kappa e^{-2\kappa s_2}}{\sigma^2|x - y_2|} \geq h^\flat(y_2, \mathcal{K}_2(s_2), -\nu_x^2). \tag{38}$$

To proceed, we need some relations between (ν_t^1, ν_x^1) and (ν_t^2, ν_x^2).

PROPOSITION 3.5. *There is some $\theta > 0$ such that*

$$\nu_x^2 = -\theta\nu_x^1 \quad \textit{and} \quad \nu_t^2 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_2} \leq -\theta(\nu_t^1 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_1})/\gamma.$$

Proof of Proposition 3.5. From the definition of t^*, we know that

$$\forall \epsilon < s < t^*,\ \mathcal{K}_1^{\epsilon,\sigma}(s) \cap \widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma s) = \emptyset.$$

Let us now notice that the sets B_1 and B_2 defined by

$$B_1 = \left\{(s, y) \mid \frac{1}{\sigma^2}(s - s_1)^2 + |y - y_1|^2 \leq \epsilon^2 e^{-2\kappa s_1}\right\}$$

and

$$B_2 = \left\{(s, y) \mid \frac{1}{\sigma^2}(\gamma s - s_2)^2 + |y - y_2|^2 \leq \epsilon^2 e^{-2\kappa s_2}\right\}$$

are, respectively, subsets of $\mathcal{K}_1^{\epsilon,\sigma}$ and of the graph of $\widehat{\mathcal{K}}_2^{\epsilon,\sigma}(\gamma\cdot)$. Therefore,

$$\forall (s, y), \text{ if } (s, y) \in B_1 \cap B_2, \text{ then } s \geq t^*. \tag{39}$$

From Proposition 3.3, the point (t^*, x) belongs to $\partial\mathcal{K}_1^{\epsilon,\sigma}$. Hence, $(t^*, x) \in \partial B_1$. In the same way, $(\gamma t^*, x)$ belongs to $\partial\widehat{\mathcal{K}}_2^{\epsilon,\sigma}$, and so $(t^*, x) \in \partial B_2$. Therefore, (39) states that (t^*, x) is a minimum in the following problem: Minimize s over the points $(s, y) \in B_1 \cap B_2$.

The necessary conditions for this problem (for the extended Lagrangian) state that there is some $(\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}_+^3$ with $(\lambda_1, \lambda_2, \lambda_3) \neq 0$, such that

$$\lambda_1\left(\frac{1}{\sigma^2}(t^* - s_1), x - y_1\right) + \lambda_2\left(\frac{\gamma}{\sigma^2}(\gamma t^* - s_2), x - y_2\right) + \lambda_3(1, 0) = 0.$$

Using the notations (35) and (36), this is equivalent to

$$\lambda_1(\nu_t^1 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_1}, \sigma^2\nu_x^1) + \lambda_2(\gamma\nu_t^2 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_2}, \sigma^2\nu_x^2) + \lambda_3(\sigma^2, 0) = 0.$$

Since, from Proposition 3.2, $\nu_x^1 \neq 0$ and $\nu_x^2 \neq 0$, we have $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$. Setting

$\theta = \lambda_1/\lambda_2 > 0$, we get $\nu_x^2 = -\theta\nu_x^1$ and

$$\begin{aligned}\nu_t^2 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_2} &= -\frac{1}{\gamma}\left(\theta(\nu_t^1 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_1}) + \frac{\lambda_3\sigma^2}{\lambda_2}\right)\\ &\le -\frac{\theta}{\gamma}(\nu_t^1 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_1}). \qquad \square\end{aligned}$$

Let us now recall the Ilmanen interposition lemma [18], which plays a crucial role in our study.

LEMMA 3.6 (Ilmanen). *Let A and B be two disjoint subsets of $\mathbb{R}^N$, A being compact and B closed. Then there exists some closed set K_r with a $\mathcal{C}^{1,1}$ boundary, such that*

$$A \subset K_r \quad \textit{and} \quad K_r \cap B = \emptyset$$

and

$$d(A,B) = d(A,\partial K_r) + d(\partial K_r, B).$$

Let us apply Lemma 3.6 to $A = \mathcal{K}_1(s_1)$ and $B = \widehat{\mathcal{K}_2}(s_2)$: There exists some set K_r with a $\mathcal{C}^{1,1}$ boundary such that

$$\mathcal{K}_1(s_1) \subset K_r \quad \text{and} \quad \widehat{\mathcal{K}_2}(s_2) \cap K_r = \emptyset$$

and

$$d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) = d(\partial K_r, \mathcal{K}_1(s_1)) + d(\partial K_r, \widehat{\mathcal{K}_2}(s_2)).$$

Let us set $\rho_1 = d(\partial K_r, \mathcal{K}_1(s_1))$, $\rho_2 = d(\partial K_r, \widehat{\mathcal{K}_2}(s_2))$, and $w = \frac{y_2 - y_1}{|y_2 - y_1|}$. Let us notice that $\nu_x^1 = |\nu_x^1| w$, while $\nu_x^2 = -|\nu_x^2| w$.

PROPOSITION 3.7. *The smooth set $K_r - \rho_1 w$ is externally tangent to $\mathcal{K}_1(s_1)$ at the point y_1 and w is a normal to $K_r - \rho_1 w$ at y_1; namely,*

$$\mathcal{K}_1(s_1) \subset (K_r - \rho_1 w) \quad \textit{and} \quad y_1 \in \partial\mathcal{K}_1(s_1) \cap \partial(K_r - \rho_1 w). \tag{40}$$

In the same way, the smooth set $K_r + \rho_2 w$ is internally tangent to $\mathcal{K}_2(s_2)$ at the point y_2 and w is a normal to $K_r + \rho_2 w$ at y_2:

$$K_r + \rho_2 w \subset \widehat{\mathcal{K}_2}(s_2) \quad \textit{and} \quad y_2 \in \partial(K_r + \rho_2 w) \cap \partial\widehat{\mathcal{K}_2}(s_2). \tag{41}$$

Finally, $S \subset Int(K_r - \rho_1 w)$ and $S \subset Int(K_r + \rho_2 w)$.

Remark. The proposition states that we can estimate the quantity $h^\sharp(y_1, \mathcal{K}_1(s_1), \nu_x^1)$ by using the set $K_r - \rho_1 w$, while the estimate of $h^\flat(y_2, \mathcal{K}_2(s_2), -\nu_x^2)$ can be done by using $K_r + \rho_2 w$.

Proof of Proposition 3.7. For proving (40) and (41), let us first notice that the fact that $\mathcal{K}_1(s_1) \subset K_r$ and $d(\partial K_r, \mathcal{K}_1(s_1)) = \rho_1$ implies the inclusion $\mathcal{K}_1(s_1) \subset (K_r - \rho_1 w)$. In the same way, since $\widehat{\mathcal{K}_2}(s_2) \cap K_r = \emptyset$ and $d(\partial K_r, \widehat{\mathcal{K}_2}(s_2)) = \rho_2$, we have $K_r + \rho_2 w \subset \widehat{\mathcal{K}_2}(s_2)$.

Let $z = y_2 - \rho_2 w = y_1 + \rho_1 w$. We have $d_{\mathcal{K}_1(s_1)}(z) \le |y_1 - z| = \rho_1$ and $d_{\widehat{\mathcal{K}_2}(s_2)}(z) \le |y_2 - z| = \rho_2$. Since, from Propositions 3.3, $d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) = |y_2 - y_1|$, this leads to

$$\rho_1 + \rho_2 \ge d_{\mathcal{K}_1(s_1)}(z) + d_{\widehat{\mathcal{K}_2}(s_2)}(z) \ge d(\mathcal{K}_1(s_1), \widehat{\mathcal{K}_2}(s_2)) = |y_1 - y_2| = \rho_1 + \rho_2.$$

So

$$d_{\mathcal{K}_1(s_1)}(z) = |y_1 - z| = \rho_1 \quad \text{and} \quad d_{\widehat{\mathcal{K}_2}(s_2)}(z)|y_2 - z| = \rho_2. \tag{42}$$

This implies that $z \notin \mathbb{R}^N \backslash K_r$, since $d(\partial K_r, \mathcal{K}_1(s_1)) = \rho_1$, and that $z \notin Int(K_r)$, since $d(\partial K_r, \widehat{\mathcal{K}_2}(s_2)) = \rho_2$. Thus, $z \in \partial K_r$, and $y_1 = z - \rho_1 w \in \partial \mathcal{K}_1(s_1) \cap \partial(K_r - \rho_1 w)$, while $y_2 = z + \rho_2 w \in \partial(K_r + \rho_2 w) \cap \partial\widehat{\mathcal{K}_2}(s_2)$. Moreover, using (42) again shows that $d(\mathcal{K}_1(s_1), \partial K_r) = |y_1 - z|$. Since K_r is smooth, this implies that w (which is proportional to $z - y_1$) is a normal to K_r at z; so (40) and (41) hold.

We now prove that $S \subset Int(K_r - \rho_1 w)$ and $S \subset Int(K_r + \rho_2 w)$. Indeed, since $S \subset Int(\mathcal{K}_1(0))$ and since $\mathcal{K}_1(0) \subset (K_r - \rho_1 w)$, we have that $S \subset Int(K_r - \rho_1 w)$. Moreover, since, from the choice of ϵ in (34),

$$S_{2\epsilon} \subset Int(S_{\gamma-1}) \subset Int(\mathcal{K}_1(0)) \subset Int(\mathcal{K}_1(s_1)) \subset Int(K_r - \rho_1 w)$$

and since $\rho_1 + \rho_2 = |y_2 - y_1| \leq 2\epsilon$, we have that $S \subset Int(K_r + \rho_2 w)$. □

We are now ready to prove the main step.

Proof of (20). Considering now the definition of $h^\sharp$ (introduced before Proposition 2.7), (40) and the facts that $\nu_x^1 = |\nu_x^1| w$ is a normal to K_r at y_1 and that $S \subset Int(K_r - \rho_1 w)$ yield

$$h^\sharp(y_1, \mathcal{K}_1(s_1), \nu_x^1) \leq h(y_1, K_r - \rho_1 w). \tag{43}$$

In the same way, (41) together with the facts that $\nu_x^2 = -|\nu_x^2| w$ is a normal to K_r at y_2 and that $S \subset Int(K_r + \rho_2 w)$ implies that

$$h^\flat(y_2, \mathcal{K}_2(s_2), -\nu_x^2) \geq h(y_2, K_r + \rho_2 w). \tag{44}$$

Using inequality (22), we can estimate the difference between the right-hand sides of the two previous inequalities:

$$h(y_2, K_r + \rho_2 w) \geq (1 - 2\lambda\epsilon) h(y_1, K_r - \rho_1 w) - C\epsilon(e^{-\kappa s_1} + e^{-\kappa s_2}) \tag{45}$$

because $y_2 - y_1 = (\rho_1 + \rho_2) w$ and $|y_1 - y_2| \leq \epsilon(e^{-\kappa s_1} + e^{-\kappa s_2}) \leq 2\epsilon$ with $\epsilon < 1/(2\lambda)$. Let us also note that

$$|\nu_x^j| \leq \sigma^2 |y_j - x| \leq \sigma^2 \epsilon e^{-\kappa s_j} \qquad \text{for } j = 1, 2. \tag{46}$$

Using Proposition 3.5 and putting together (37), (38), and the three previous inequalities finally gives

$$\begin{aligned}
h(y_2, K_r + \rho_2 w) \;&\leq\; h^\flat(y_2, \mathcal{K}_2(s_2), -\nu_x^2) \quad \text{(from (44))} \\
&\leq \frac{\nu_t^2}{|\nu_x^2|} \quad \text{(from (38))} \\
&\leq -\frac{\nu_t^1 + \epsilon^2\sigma^2\kappa e^{-2\kappa s_1}}{\gamma|\nu_x^1|} - \frac{\epsilon^2\sigma^2\kappa e^{-2\kappa s_2}}{|\nu_x^2|} \quad \text{(from Proposition 3.5)} \\
&\leq \frac{1}{\gamma} h^\sharp(y_1, \mathcal{K}_1(s_1), \nu_x^1) - \frac{\epsilon^2\sigma^2\kappa e^{-2\kappa s_1}}{\gamma|\nu_x^1|} - \epsilon\kappa e^{-\kappa s_2} \quad \text{(from (37) and (46))} \\
&\leq \frac{1}{\gamma} h(y_1, K_r - \rho_1 w) - \frac{\epsilon\kappa(e^{-\kappa s_1} + e^{-\kappa s_2})}{\gamma} \quad \text{(from (43) and (46) again)} \\
&\leq \frac{1}{\gamma(1 - 2\lambda\epsilon)} h(y_2, K_r + \rho_2 w) + \frac{C\epsilon(e^{-\kappa s_1} + e^{-\kappa s_2})}{\gamma(1 - 2\lambda\epsilon)} - \frac{\epsilon\kappa(e^{-\kappa s_1} + e^{-\kappa s_2})}{\gamma}
\end{aligned}$$

using (45) in the last inequality. This is impossible since $h(y_2, K_r + \rho_2 w) \geq 0$ and we have chosen $\gamma \geq 1/(1-2\lambda\epsilon)$ in (34) and $\kappa = C\gamma + 1$ in (23). So we have found a contradiction, and (20) is proved. □

Proof of Theorem 3.1. Since from our assumption $\overline{\mathcal{K}_1}(0) \cap \widehat{\mathcal{K}_2}(0) = \emptyset$, and since $\overline{\mathcal{K}_1}$ and $\widehat{\mathcal{K}_2}$ have a closed graph, one can find $\tau > 0$ such that

$$\forall t \in [0, \tau], \qquad \overline{\mathcal{K}_1}(t) \cap \widehat{\mathcal{K}_2}(0) = \emptyset.$$

Let us now apply (20) to the subsolution $\mathcal{K}_{1,\delta}$ and the supersolution $\mathcal{K}_2$, where

$$\delta \in (0, \tau) \qquad \forall t \geq 0, \qquad \mathcal{K}_{1,\delta}(t) = \overline{\mathcal{K}_1}(t + \delta).$$

Since $\mathcal{K}_{1,\delta}(0) \cap \widehat{\mathcal{K}_2}(0) = \emptyset$, we have, for any $\gamma > 1$,

$$\forall t \geq 0,\ \mathcal{K}_{1,\delta}(t) \cap \widehat{\mathcal{K}_2}(\gamma t) = \emptyset.$$

Applying this with $t - \delta$ (for $t > \delta$) and to $\gamma = t/(t-\delta) > 1$ gives

$$\emptyset = \mathcal{K}_{1,\delta}(t-\delta) \cap \widehat{\mathcal{K}_2}(\gamma(t-\delta)) = \mathcal{K}_1(t) \cap \widehat{\mathcal{K}_2}(t)$$

because $\gamma(t-\delta) = t$. Since we can choose $\delta > 0$ arbitrary small, the proof of Theorem 3.1 is complete. □

4. Existence, uniqueness, and stability of solutions. In this section, we prove the existence of viscosity solutions for the Hele–Shaw problem for power-law fluids. We also state some uniqueness and stability results.

Throughout this section we assume that h is given by (8). However, most results can be easily extended to the general velocity law h.

4.1. Some preliminary estimates. We give here some technical estimates which are necessary in what follows. We first establish some estimates of the growth of the solutions. For this, we recall that, according to Proposition 2.6, there are constants $r_0 > 0$ and $\sigma > 0$ such that

$$\forall r \geq r_0,\ \forall x \in \partial B(0, r), \qquad h(x, B(0, r)) \leq \sigma r.$$

Moreover, the constants r_0 and σ depend only on S, p, $\|f\|_\infty$, and $\|g\|_\infty$.

LEMMA 4.1. *If $\mathcal{K}$ is a subsolution of the front propagation problem, then*

$$\forall t \geq 0, \quad \mathcal{K}(t) \subset B\left(0, \max\{|\mathcal{K}(0)|, r_0\} e^{\sigma t}\right),$$

where $|\mathcal{K}(0)| = \sup_{y \in \mathcal{K}(0)} |y|$.

Proof of Lemma 4.1. From Proposition 2.6, for any $\epsilon > 0$, the tube

$$\mathcal{K}_2^\epsilon(t) = B\left(0, (\max\{|\mathcal{K}(0)|, r_0\} + \epsilon) e^{\sigma t}\right)$$

is a supersolution of the front propagation problem with $\overline{\mathcal{K}}(0) \cap \widehat{\mathcal{K}_2^\epsilon}(0) = \emptyset$. Hence, $\overline{\mathcal{K}}(t) \cap \widehat{\mathcal{K}_2^\epsilon}(t) = \emptyset$ for any $t \geq 0$, which entails the desired result when letting $\epsilon \to 0^+$. □

The following results state that the left lower semicontinuity of a subsolution is somehow "uniform."

LEMMA 4.2. *Let us fix $\epsilon > 0$ and $\rho > 0$ such that $\rho > \epsilon$. Then there is some constant $\eta > 0$ such that, for any subsolution $\mathcal{K}$ of the front propagation problem,*

with $\overline{\mathcal{K}}(0) \subset B(0,\rho)$ *and for any* $x_0 \in \mathbb{R}^N$ *with* $d_{\overline{\mathcal{K}}(0)}(x_0) \geq \epsilon$ *and* $|x_0| \leq \rho$, *we have* $d_{\mathcal{K}}(0,x_0) \geq \eta$.

Moreover, the constant $\eta > 0$ *depends only on* ϵ, ρ *and on the maximum of the velocity* $h(y,K_r)$ *for* $y \in \partial K_r$ *and for* K_r *belonging to the compact family of smooth sets:*

$$\left\{ B(z,R)\backslash B(z,r) \mid \frac{\epsilon}{4} \leq r \leq \frac{\epsilon}{2},\ 2\rho \leq R \leq 3\rho,\ |z| \leq \rho,\ d_S(z) \geq \epsilon \right\}.$$

Remarks.

1. Since the front propagation problem we are considering is invariant with respect to time translations, the above estimate also shows that, for any subsolution $\mathcal{K}$ of the front propagation problem, with $\overline{\mathcal{K}}(t) \subset B(0,\rho)$ and for any $x_0 \in \mathbb{R}^N$ with $d_{\overline{\mathcal{K}}(t)}(x_0) > \epsilon$ and $|x_0| \leq \rho$, we have $d_{\mathcal{K}}(t,x_0) \geq \eta$.
2. The symmetric estimates for a supersolution (i.e., if $\mathcal{K}$ is a supersolution, $d_{\widehat{\mathcal{K}}(0)}(x_0) > \epsilon$ and $|x_0| \leq \rho$ implies that $d_{\widehat{\mathcal{K}}}(0,x_0) \geq \eta$, for some η) clearly hold with $\eta = \epsilon$ because, since $\mathcal{K}$ is nondecreasing, $\widehat{\mathcal{K}}$ is nonincreasing.

Proof of Lemma 4.2. Let us denote by κ the maximum of the velocity $h(y,K_r)$ for $y \in \partial K_r$ and for K_r belonging to the compact family of smooth sets:

$$\left\{ B(z,R)\backslash B(z,r) \mid \frac{\epsilon}{4} \leq r \leq \frac{\epsilon}{2},\ 2\rho \leq R \leq 3\rho,\ |z| \leq \rho,\ d_S(z) \geq \epsilon \right\}.$$

Note that $\kappa < +\infty$ since the above family is compact and the velocity h is continuous (Proposition 2.4). Let us introduce the tube $\mathcal{K}_2$ defined by

$$\mathcal{K}_2(t) = B(x_0, 2\rho + \kappa t)\backslash B\left(x_0, \frac{\epsilon}{2} - \kappa t\right).$$

Then, from the definition of κ, $\mathcal{K}_2$ is a smooth supersolution of the Hele–Shaw problem on the time interval $[0,\tau]$, where $\tau = \min\{\frac{\epsilon}{4\kappa}, \frac{\rho}{\kappa}\}$, because, on this time interval, $2\rho \leq 2\rho + \kappa t \leq 3\rho$ and $\epsilon/4 \leq \epsilon - \kappa t \leq \epsilon/2$.

Let $\mathcal{K}$ be some subsolution of the Hele–Shaw problem, with $\overline{\mathcal{K}}(0) \subset B(0,\rho)$, and let $x_0 \in \mathbb{R}^N$ with $d_{\overline{\mathcal{K}}(0)}(x_0) > \epsilon$ and $|x_0| \leq \rho$. Then we have $\overline{\mathcal{K}}(0) \subset B(x_0,2\rho)\backslash B(x_0,\epsilon)$. Hence, $\overline{\mathcal{K}}(0) \cap \widehat{\mathcal{K}_2}(0) = \emptyset$. Then the inclusion principle (Theorem 3.1) states that

$$\forall t \in [0,\tau],\ \overline{\mathcal{K}}(t) \cap \widehat{\mathcal{K}_2}(t) = \emptyset.$$

Therefore, $d_{\mathcal{K}}(0,x_0) \geq \eta$, where $\eta = \min\{\epsilon/2, \tau\}$, because $B((0,x_0),\eta) \subset \widehat{\mathcal{K}_2}$. ∎

4.2. Existence and uniqueness of solutions. Let us first give an existence result.

THEOREM 4.3. *For any initial position* K_0, *with* $S \subset Int(K_0)$ *and* K_0 *bounded, there is (at least) one solution to the front propagation problem.*

Moreover, there is a largest solution, denoted by $S(K_0)$, *and a smallest solution, denoted by* $s(K_0)$, *to this problem. The largest solution has a closed graph while the smallest solution has an open graph in* $\mathbb{R}^+ \times \mathbb{R}^N$. *The largest solution contains all the subsolutions of the front propagation problem with initial condition* K_0, *while the smallest solution is contained in any supersolution.*

Remarks.

1. From the maximality property of the largest solution and the time invariance of the evolution law, the semigroup property holds for this solution:

$$\forall s \geq 0,\ t \geq 0, \quad S(\,S(K_0)(s)\,)(t) = S(K_0)(s+t).$$

2. For general front propagation problems, one cannot expect the uniqueness of the solutions. Soner pointed out in [23] the existence of a maximal and a minimal solution for geometric flows of mean curvature type. This result has been generalized in [10] for some class of geometric flows with nonlocal terms.

Proof of Theorem 4.3. The proof is based on Perron's method. Since it is exactly the same as the proof of Theorem 4.1 and of Corollary 4.2 of [11], we omit it. □

We say that the solution of our Hele–Shaw problem with initial position K_0 is *unique* if $\widehat{s(K_0)} = S(K_0)$ or if, equivalently, $\widehat{S(K_0)} = \widehat{s(K_0)}$ (note that $\widehat{s(K_0)} = (\mathbb{R}^+ \times \mathbb{R}^N) \backslash s(K_0)$ since $s(K_0)$ has an open graph in $\mathbb{R}^+ \times \mathbb{R}^N$).

We have the following uniqueness result.

THEOREM 4.4. *Assume that K_0 is the closure of an open, connected, and bounded subset of $\mathbb{R}^N$ with a $\mathcal{C}^2$ boundary and such that $S \subset Int(K_0)$. Then there is a unique viscosity solution to the Hele–Shaw problem.*

Remark. Some uniqueness criteria for geometric flows can be found in [23] and [4]. Our proof uses several arguments from these papers.

Proof of Theorem 4.4. Since K_0 is the closure of an open, connected, and bounded subset of $\mathbb{R}^N$ with a $\mathcal{C}^2$ boundary and such that $S \subset Int(K_0)$, the Hopf maximum principle implies that there is a constant $\delta > 0$ such that $h(x, K_0) \geq 2\delta$ for any $x \in \partial K_0$. Let us set, for any $\sigma \in \mathbb{R}$, $K_\sigma = \{x \in K_0 |\ \mathbf{d}(x) \leq \sigma\}$, where $\mathbf{d}$ is the signed distance to the boundary of K_0 (negative in $Int(K_0)$). From the continuity of h (see (11)), there is some $\epsilon > 0$ such that $h(x, K_\sigma) \geq \delta$ for any $x \in \partial K_\sigma$ and for any σ such that $|\sigma| \leq \epsilon$. Hence, the tube $\mathcal{K}(t) = K_{\delta t - \epsilon}$ is a subsolution of the Hele–Shaw problem starting from $K_{-\epsilon}$ on the time interval $[0, 2\epsilon/\delta]$, because it is smooth and has a normal velocity δ on this interval of time. In particular, $\mathcal{K}(t) \subset S(K_{-\epsilon})(t)$ on $[0, 2\epsilon/\delta]$, which proves that $\mathcal{K}(2\epsilon/\delta) = K_\epsilon \subset S(K_{-\epsilon})(2\epsilon/\delta)$.

Since $K_0 \subset Int(K_\epsilon) \subset Int(S(K_{-\epsilon})(2\epsilon/\delta))$, the inclusion principle (Theorem 3.1) combined with the semigroup property gives

$$S(K_0)(t) \subset S(S(K_{-\epsilon})(2\epsilon/\delta))(t) = S(K_{-\epsilon})(2\epsilon/\delta + t) \quad \forall t \geq 0.$$

Moreover, since $K_{-\epsilon} \subset Int(K_0)$, the inclusion principle also states that $S(K_{-\epsilon})(t) \subset s(K_0)(t)$. Accordingly, we have, for all $t \geq 0$, $\underline{S(K_0)}(t) \subset s(K_0)(2\epsilon/\delta + t)$. Letting $\epsilon \to 0^+$ gives the desired inclusion $S(K_0)(t) \subset \overline{s(K_0)}(t)$. □

4.3. Stability of the solutions. We are now investigating the stability of the flow under variations of the initial position and of the data f and g.

For this we first generalize the well-known stability result of viscosity solutions to our framework.

Let us assume that we are given a family of maps $h_n = h_n(x, K)$ which are defined for any set K with a $\mathcal{C}^{1,1}$ boundary and for any $x \in \partial K$, and continuous in the sense of (11). We also assume that h_n converges to a continuous map h; i.e., if a sequence of closed set K_n with a $\mathcal{C}^{1,1}$ boundary converges to a closed set K with a $\mathcal{C}^{1,1}$ boundary for the $\mathcal{C}^{1,1}$ convergence, and if a sequence of points (x_n), with $x_n \in \partial K_n$, converges to some $x \in \partial K$, then $h_n(x_n, K_n)$ converges to $h(x, K)$.

Let us recall that the upper limit of a sequence of sets A_n is the set of all limits of converging subsequences of sequences (x_n) with $x_n \in A_n$.

PROPOSITION 4.5. *If $\mathcal{K}_n$ is a sequence of subsolutions for h_n, locally uniformly bounded w.r.t. t, then $\mathcal{K}^*$, the upper limit of the $\mathcal{K}_n$, is also a subsolution for h.*

In a similar way, if $\mathcal{K}_n$ is a sequence of supersolutions for h_n, locally uniformly bounded w.r.t. t, then $\mathcal{K}_*$, the complementary of the upper limit of $\widehat{\mathcal{K}_n}$, is also a supersolution for h.

Proof of Proposition 4.5. We prove only the statement for the subsolutions, the proof for the supersolutions being similar. Let us first prove that $\mathcal{K}^*$ is a left lower semicontinuous tube. Indeed $\mathcal{K}^*$ is a tube because the $\mathcal{K}_n$ are locally uniformly bounded. In order to show that $\mathcal{K}^*$ is left lower semicontinuous, it is enough to establish that, for any $T \geq 0$ and for any $\epsilon > 0$, there is some $\eta > 0$ such that

$$\forall t \in [0,T],\ \forall x \in \mathbb{R}^N, \quad d_{\mathcal{K}^*(t)}(x) > \epsilon \ \Rightarrow\ d_{\mathcal{K}^*}(t,x) \geq \eta.$$

For this, let us fix $T \geq 0$ and $\epsilon > 0$. Since the h_n converge to h, Lemma 4.2 states that there is some $\eta > 0$ (independent of n and of $t \in [0,T]$) such that

$$\forall n \in \mathbb{N},\ \forall t \in [0,T],\ \forall x \in \mathbb{R}^N, \quad d_{\mathcal{K}_n(t)}(x) > \epsilon/2 \ \Rightarrow\ d_{\mathcal{K}_n}(t,x) \geq \eta.$$

Let us now assume that $d_{\mathcal{K}^*(t)}(x) > \epsilon$ for some $x \in \mathbb{R}^N$ and for some $t \in [0,T]$. Since $\mathcal{K}^*$ is equal to the upper limit of the $\mathcal{K}_n$, we have $d_{\mathcal{K}_n(t)}(x) > \epsilon/2$ for any n large enough, whence $d_{\mathcal{K}_n}(t,x) \geq \eta$. This implies that $d_{\mathcal{K}^*}(t,x) \geq \eta$. So $\mathcal{K}^*$ is a left lower semicontinuous tube.

Let us now check that $\mathcal{K}^*$ is a subsolution. For this, let us fix some smooth tube $\mathcal{K}_r$ which is externally tangent to $\mathcal{K}^*$ at some point (t,x). We denote by $\mathbf{d}$ the signed distance function to $\partial\mathcal{K}_r$. This function is $\mathcal{C}^{1,1}$ in a neighborhood $\mathcal{V} = \{|\mathbf{d}| < \eta\}$ of $\partial\mathcal{K}_r$.

Let us now consider the function $\mathbf{d}_\epsilon(s,y) = \mathbf{d}(s,y) - \epsilon|(s,y)-(t,x)|^2$, for $\epsilon > 0$. Let us underline that $\mathbf{d}_\epsilon$ has a unique maximum on $\mathcal{K}^*$ at the point (t,x). We can choose $\epsilon > 0$ sufficiently small in such a way that the set $\mathcal{K}_r^\epsilon = \{\mathbf{d}_\epsilon \leq 0\}$ has a boundary which is contained in $\mathcal{V}$ and $\nabla\mathbf{d}_\epsilon \neq 0$ on $\mathcal{V}$. Let us now consider a point (t_n,x_n) of maximum of $\mathbf{d}_\epsilon$ onto $\mathcal{K}_n$. By using a standard argument, we can prove that a subsequence of (t_n,x_n) (again denoted (t_n,x_n)) converges to (t,x), because $\mathbf{d}_\epsilon$ has a unique maximum on $\mathcal{K}^*$ at the point (t,x). Hence, the set $\mathcal{K}_r^n = \{\mathbf{d}_\epsilon \leq \mathbf{d}_\epsilon(t_n,x_n)\}$ is a smooth tube for n large enough, since, for n large enough, the boundary of $\mathcal{K}_r^n$ is in $\mathcal{V}$. Since $\mathcal{K}_n$ is a subsolution and since $\mathcal{K}_r^n$ is externally tangent to $\mathcal{K}_n$ at (t_n,x_n), we have

$$V^{\mathcal{K}_r^n}_{(t_n,x_n)} \leq h_n(x_n, \mathcal{K}_r^n(t_n)).$$

The sequence of sets $\mathcal{K}_r^n$ converges to $\mathcal{K}_r^\epsilon$ for the $\mathcal{C}^{1,1}$ topology. So we get at the limit

$$V^{\mathcal{K}_r^\epsilon}_{(t,x)} \leq h(x, \mathcal{K}_r^\epsilon(t))$$

because h_n converges to h. Letting $\epsilon \to 0$ gives the desired result, since $\mathcal{K}_r^\epsilon$ converges to $\mathcal{K}_r$ for the $\mathcal{C}^{1,1}$ topology and h is continuous. □

We finally investigate the stability of solutions with respect to the initial position K_0 and to the functions f and g.

THEOREM 4.6. *Let (f_n,g_n) and (f,g) satisfy* (10) *for any n, with $f>0$ and locally Lipschitz continuous. Let $K_n \in \mathcal{D}$ be a sequence of initial positions and $K_0 \in \mathcal{D}$.*

Let us assume that the f_n are globally bounded and converge to f locally uniformly, that the g_n converge to some $g > 0$ in $\mathcal{C}^{1,\beta}(\partial S)$ for some $\beta \in (0,1)$, and that the K_n converge to K_0, in the sense that the upper limit of the K_n is contained in K_0 and the upper limit of the $\mathbb{R}^N\backslash K_n$ is contained in $\overline{\mathbb{R}^N\backslash K_0}$. Let us also suppose that there is a unique solution, denoted by $S(K_0)$, of the Hele–Shaw problem starting from K_0 with data f and g.

If $\mathcal{K}_n$ is a solution of the Hele–Shaw problem, with data f_n and g_n, starting from K_n, then the $\mathcal{K}_n$ converge to $S(K_0)$ in the following sense: The upper limit of the $\mathcal{K}_n$ is equal to $S(K_0)$, while the upper limit of the $\widehat{\mathcal{K}_n}$ is equal to $\widehat{S(K_0)}$.

Proof of Theorem 4.6. Let $\mathcal{K}^*$ be the upper limit of the $\mathcal{K}_n$ and $\mathcal{K}_*$ be the complementary of the upper limit of the $\widehat{\mathcal{K}_n}$.

Let us first prove that $\mathcal{K}^*$ is a subsolution to the front propagation problem with initial position K_0. According to Proposition 4.5, it is enough to prove that the $\mathcal{K}_n$ are locally uniformly bounded w.r.t. t and the maps h_n, defined for any smooth set $K_r \in \mathcal{D}$ and any $x \in \partial K_r$ by $h_n(x, K_r) = |\nabla u(x)|^{p-1}$, where u is the solution to

$$\begin{cases} -\mathrm{div}(|\nabla u|^{p-1}\nabla u) = f_n & \text{in } K_r, \\ u = g_n & \text{on } \partial S, \\ u = 0 & \text{on } \partial K_r, \end{cases}$$

converge to h defined by (8). The $\mathcal{K}_n$ are locally uniformly bounded thanks to Lemma 4.1, because the f_n and the g_n are uniformly bounded. Moreover, the local uniform convergence of the h_n to h is a straightforward application of the estimates in [21]. Using Lemma 4.2, we can also show that $\mathcal{K}^*(0) \subset K_0$ (the arguments are similar to those developed for proving that $\mathcal{K}^*$ is left lower semicontinuous in Proposition 4.5). Hence, $\mathcal{K}^*$ is a subsolution to the Hele–Shaw problem with initial position K_0. In particular, this implies that $\mathcal{K}^* \subset S(K_0)$, because $S(K_0)$ contains any subsolution.

In the same way, $\mathcal{K}_*$ is a supersolution for h, with $\widehat{\mathcal{K}_*}(0) \subset \overline{\mathbb{R}^N \backslash K_0}$. Hence, $s(K_0) \subset \mathcal{K}_*$. So we have proved that

$$s(K_0) \subset \mathcal{K}_* \subset \mathcal{K}^* \subset S(K_0). \tag{47}$$

From our assumption, the Hele–Shaw problem with initial position K_0 has a unique solution, i.e., $\overline{s(K_0)} = S(K_0)$. Combining this equality with (47) gives $\overline{s(K_0)} = \overline{\mathcal{K}_*} = \mathcal{K}^* = S(K_0)$, since $\mathcal{K}^*$ and $S(K_0)$ have a closed graph.

Taking the complementary in (47) also gives $\widehat{S(K_0)} \subset \widehat{\mathcal{K}^*} \subset \widehat{\mathcal{K}_*} \subset \widehat{s(K_0)}$. Since $\widehat{S(K_0)} = \widehat{s(K_0)}$ from the uniqueness of the solution, we finally have the equality $\widehat{\mathcal{K}_*} = \widehat{S(K_0)}$, which is the desired result since $\widehat{\mathcal{K}_*} = (\mathbb{R}^+ \times \mathbb{R}^N)\backslash\mathcal{K}_*$ is the upper limit of the $\widehat{\mathcal{K}_n}$. □

REFERENCES

[1] B. Andrews and M. Feldman, *Nonlocal geometric expansion of convex planar curves*, J. Differential Equations, 182 (2002), pp. 298–343.

[2] G. Aronsson, *On p-harmonic functions, convex duality and an asymptotic formula for injection mould filling*, European J. Appl. Math., 7 (1996), pp. 417–437.

[3] G. Aronsson and U. Janfalk, *On Hele-Shaw flow of power-law fluids*, European J. Appl. Math., 3 (1992), pp. 343–366.

[4] G. Barles, H. M. Soner, and P. E. Souganidis, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[5] G. Barles and P. E. Souganidis, *A new approach to front propagation problems: Theory and applications*, Arch. Rational Mech. Anal., 141 (1998), pp. 237–296.

[6] G. Bellettini and M. Novaga, *Comparison results between minimal barriers and viscosity solutions for geometric evolutions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 97–131.

[7] G. Bellettini and M. Novaga, *Minimal barriers for geometric evolutions*, J. Differential Equations, 139 (1997), pp. 76–103.

[8] G. Bellettini and M. Paolini, *Some results on minimal barriers in the sense of De Giorgi applied to driven motion by mean curvature*, Rend. Accad. Naz. Sci. XL Mem. Mat. Appl. (5), 19 (1995), pp. 43–67.

[9] L. A. Caffarelli and J. L. Vazquez, *Viscosity solutions for the porous medium*, Proc. Sympos. Pure Math., 65 (1999), pp. 13–26.
[10] P. Cardaliaguet, *On front propagation problems with nonlocal terms*, Adv. Differential Equations, 5 (2000), pp. 213–268.
[11] P. Cardaliaguet, *Front propagation problems with nonlocal terms,* II, J. Math. Anal. Appl., 260 (2001), pp. 572–601.
[12] Y.-G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.
[13] F. Da Lio, C. I. Kim, and D. Slepscev, *Nonlocal front propagation problems in bounded domains with Neumann-type boundary conditions and applications*, Asymptot. Anal., 37 (2004), pp. 257–292.
[14] C. M. Elliott and V. Janovský, *A variational inequality approach to Hele-Shaw flow with a moving boundary*, Proc. Roy. Soc. Edinburgh Sect. A, 88 (1981), pp. 93–107.
[15] J. Escher and G. Simonett, *Classical solutions of multidimensional Hele-Shaw models*, SIAM J. Math. Anal., 28 (1997), pp. 1028–1047.
[16] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature,* I, J. Differential Geom., 33 (1991), pp. 635–681.
[17] L. Giacomelli and F. Otto, *Variational formulation for the lubrication approximation of the Hele-Shaw flow*, Calc. Var. Partial Differential Equations, 13 (2000), pp. 377–403.
[18] T. Ilmanen, *The level-set flow on a manifold*, Proc. Sympos. Pure Math., 54 (1993), pp. 193–204.
[19] C. I. Kim, *Uniqueness and existence results for the Hele-Shaw and the Stefan problems*, Arch. Ration. Mech. Anal., 168 (2003), pp. 299–328.
[20] C. Lederman, J. L. Vásquez, and N. Wolanski, *Uniqueness of solution to a free boundary problem from combustion*, Trans. Amer. Math. Soc., 353 (2000), pp. 655–692.
[21] G. M. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal., 12 (1988), pp. 1203–1219.
[22] S. V. Rogosin, *On classical formulation of Hele-Shaw moving boundary problem for power-law fluid*, Math. Model. Anal., 7 (2002), pp. 159–168.
[23] H. M. Soner, *Motion of a set by the mean curvature of its boundary*, J. Differential Equations, 101 (1993), pp. 313–372.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 166–187

PERIODIC OSCILLATIONS OF BLOOD CELL POPULATIONS IN CHRONIC MYELOGENOUS LEUKEMIA*

MICHAEL C. MACKEY†, CHUNHUA OU‡, LAURENT PUJO-MENJOUET§, AND JIANHONG WU¶

Abstract. Periodic chronic myelogenous leukemia and cyclical neutropenia are two hematological diseases that display oscillations in circulating cell numbers with a period far in excess of what one might expect based on the stem cell cycle duration. Motivated by this observation and a desire to understand how long period oscillations can arise, we analytically prove the existence and stability of long period oscillations in a G_0 phase cell cycle model described by a nonlinear differential delay equation. This periodic oscillation p_∞ can be analytically constructed when the proliferative control is of a "bang-bang" type (the Hill coefficient involved in the nonlinear feedback is infinite). We further obtain a contractive return map (for the semiflow generated by the functional differential equation) in a closed and convex cone containing p_∞ when the proliferative control is smooth (the Hill coefficient is large but finite). The fixed point of this contractive map gives the long period oscillation previously observed both numerically and experimentally.

Key words. cell proliferation, G_0 cell cycle model, periodic chronic myelogenous leukemia, long period oscillations, delay differential equations, Hill function, Walther's method

AMS subject classifications. 34C25, 34K18, 37G15

DOI. 10.1137/04061578X

1. Introduction.

Periodic hematological diseases have attracted a significant amount of modeling attention from mathematicians, notably the disorders periodic autoimmune hemolytic anemia [3, 17] and cyclical thrombocytopenia [27, 29]. Periodic hematological diseases of this type, in which only a single cell type is typically involved, usually display a periodicity in circulating cell numbers between two and four times the bone marrow production delay. This clinical observation has a clear explanation within a modeling context [10].

Other periodic hematological diseases such as cyclical neutropenia [4, 10, 11, 15, 16, 18] and periodic chronic myelogenous leukemia (PCML) [8] have more than one circulating blood cell type (i.e., white cells, red blood cells, and platelets) that display oscillatory levels. The oscillations in cell numbers in these two diseases have period durations ranging from weeks to months in general and are thought to originate in the pluripotential stem cell compartment [10]. In the particular case of PCML, the period can range from 40 to 80 days. Two lines of evidence indicate that the PCML oscillations originate in the stem cell population based in the bone marrow. The first suggestion that this is the case comes from the presence of the Philadelphia

*Received by the editors September 26, 2004; accepted for publication (in revised form) November 29, 2005; published electronically April 12, 2006.
http://www.siam.org/journals/sima/38-1/61578.html

†Department of Physiology, Centre for Nonlinear Dynamics, McGill University, 3655 Drummond, Montréal, Québec H3G 1Y6, Canada (mackey@cnd.mcgill.ca).

‡Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada (chqu@mathstat.yorku.ca). Current address: Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Canada, A1C 5S7.

§Institut Camille Jordan, Université Claude Bernard Lyon 1, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France (pujo@math.univ-lyon1.fr).

¶Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada (wujh@mathstat.yorku.ca).

chromosome in all hematopoietic cells in PCML [5, 7, 9, 12, 30]. Second, in PCML it is observed that white blood cells, erythrocytes, and platelets all oscillate with the same period [8].

"How do 'short' cell cycles give rise to 'long' period oscillations?" This question has arisen from the observation of circulating blood cell oscillations in PCML [8]. There is an enormous difference between the relatively short cell cycle duration, which ranges between 1 and 4 days [13, 18, 19], and the long period oscillations in PCML (between 40 and 80 days) [8]. The link between these relatively short cycle durations and the long periods of peripheral cell oscillations in PCML is unclear and has been neither biologically explained nor understood.

Using a G_0 model of the cell cycle [6, 20, 28], an attempt to answer this question has been made in [1, 25, 24], where the role of various model parameters on the period and amplitude of the cellular oscillations was examined. When cellular reentry from G_0 into the proliferative phase is subject to "bang-bang" control (technically, where the Hill coefficient in the model re-entry rate n is infinite—see below), qualitatively the cell cycle regulation parameters have a major influence on the oscillation amplitude, while the oscillation period is primarily determined by the cell death and differentiation parameters. Under this strong assumption, the cell cycle model is described by a piecewise linear scalar delay differential equation that, after nontrivial but straightforward calculations, has a periodic solution with large period and amplitude and strong stability properties.

Here, we prove analytically that similar conclusions hold in the more biologically realistic case that the re-entry rate is a smooth monotone function. We construct a convex closed cone containing the periodic solution when $n = \infty$ and a contractive return map defined on this cone such that a fixed point of the return map gives a stable periodic solution of the model equation when n is large. This method was first developed by Walther [31, 32] for a scalar delay differential equation with constant linear instantaneous friction and a negative delayed feedback, and was later extended to state-dependent delay differential equations [33, 34] and to delay differential systems [34, 36]. This method was further developed in [23] by incorporating some ideas from classical asymptotic analysis and using matching methods. Applications of this method to the present cell cycle model are nontrivial since both the instantaneous loss and the delayed production of stem cells involve the nonlinearity and there is no analytic formula for the periodic solution in the limiting case ($n = \infty$).

This paper is organized as follows. In section 2 we present the model in detail. Section 3 summarizes previous results from [24] in the case where the Hill coefficient n is infinite. Then, we introduce a more general result for the perturbed delay equation given in section 4, and we present our main results in section 5 including the full asymptotic expansion for the periodic solutions.

2. Description of the model. The G_0 model of the cell cycle (see Figure 2.1 for a depiction) is conceptually based on the work of Lajtha [14] and was first developed by Burns and Tannock [6]. It can be derived from an age structured system of two coupled partial differential equations, along with appropriate boundary and initial conditions [15, 16, 21, 26]. Integrating along characteristics [35] these equations can be transformed into a pair of coupled nonlinear first-order differential delay equations [15, 16, 18]. The resulting model depicted in Figure 2.1 consists of a proliferating phase cellular population $P(t)$ at time t and a G_0 resting phase, with a population of cells $N(t)$. The proliferative phase cells consist of cells in the G_1 phase of the cell cycle, the DNA synthesis (S) phase, G_2, and mitosis M. In this proliferative phase,

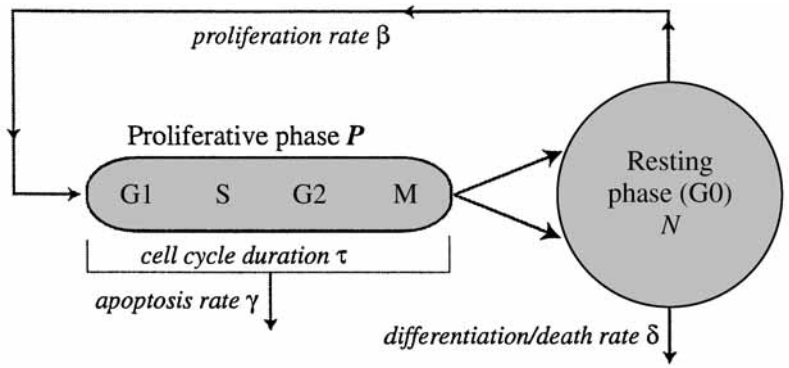


FIG. 2.1. *A schematic representation of the G_0 stem cell model. Proliferating phase cells (P) include those cells in G_1, S (DNA synthesis), G_2, and M (mitosis), while the resting phase (N) cells are in the G_0 phase. δ is the rate of differentiation into all the committed populations arising from the stem cells, and γ represents the apoptotic loss of proliferating phase cells. β is the rate of cell re-entry from G_0 into the proliferative phase, and the cell cycle time τ is the duration of the proliferative phase. See* [15, 16, 18] *for further details.*

cells are committed to undergo cell division a constant time τ after their entry into G_1. The choice of a constant cell cycle time τ simplifies the problem, though some models with a nonconstant value of τ have been examined [2, 4]. The proliferative phase death rate γ is due to apoptosis (programmed cell death). At the point of cytokinesis (cell division), a cell divides into two daughter cells, both of which are assumed to enter the resting (N) phase. In this phase, cells cannot divide but they may have one of three possible fates: differentiate at a constant rate δ, re-enter the proliferative phase at a rate β, or remain in G_0. The re-entry rate β is a nonlinear function of the cellular density and the central focus of this study.

The full model, described by a coupled nonlinear first-order delay equation, takes the form

$$\frac{dP(t)}{dt} = -\gamma P(t) + \beta(N)N - e^{-\gamma\tau}\beta(N_\tau)N_\tau \tag{2.1}$$

and

$$\frac{dN(t)}{dt} = -[\beta(N) + \delta]N + 2e^{-\gamma\tau}\beta(N_\tau)N_\tau, \tag{2.2}$$

where $N_\tau = N(t-\tau)$. The resting (G_0) to proliferative phase feedback rate β is taken to be a monotone Hill function of the form

$$\beta(N) = \frac{\beta_0\theta^n}{\theta^n + N^n}.$$

In (2.2), the first term represents the loss of nonproliferating cells to the proliferative phase (flux $\beta(N)N$) and to differentiation (flux δN). The second term represents the production of G_0 phase cells from the proliferating stem cells. The factor 2 accounts for the amplifying effect of cell division while $e^{-\gamma\tau}$ accounts for the attenuation in the proliferative phase due to apoptosis. Note that we need to study only the dynamics of the G_0 phase resting population (governed by (2.2)) since the proliferating phase dynamics (governed by (2.1)) are driven by the dynamics of the resting cells. This is

strictly a consequence of the fact that we have assumed β to be a function of N alone [21, 22].

Introducing the dimensionless variable $x = N/\theta$, we can rewrite (2.2) as

$$\frac{dx}{dt} = -[\beta(x) + \delta]x + k\beta(x_\tau)x_\tau, \tag{2.3}$$

where

$$\beta(x) = \beta_0 \frac{1}{1 + x^n}, \tag{2.4}$$

and $k = 2e^{-\gamma\tau}$. The steady states x_* of (2.3) are given by the solution of $dx/dt \equiv 0$. Thus we have $x_* \equiv 0$, and

$$x_* = \left(\beta_0 \frac{k-1}{\delta} - 1\right)^{1/n}. \tag{2.5}$$

Here we require

$$\tau < -\frac{1}{\gamma} \ln \frac{\delta + \beta_0}{2\beta_0},$$

so $\beta_0 \frac{k-1}{\delta} > 1$ in (2.5) and the second nontrivial steady state will be positive. Note that when $n \to \infty$, $x_* \to 1$ in (2.5) and $\beta(x)$ tends to a piecewise constant function (the Heaviside step function).

A solution of (2.3) is a continuous function $x : [-\tau, +\infty) \to \mathbf{R}_+$ obeying (2.3) for all $t > 0$. The continuous function $\varphi : [-\tau, 0] \to \mathbf{R}_+, \varphi(t) = x(t)$ for all $t \in [-\tau, 0]$, is called the initial condition for x. Using the method of steps, it is easy to prove that for every $\varphi \in C([-\tau, 0])$, where $C([-\tau, 0])$ is the space of continuous functions on $[-\tau, 0]$, there is a unique solution of (2.3) subject to the initial condition φ.

3. Periodic solutions: Limiting nonlinearity. In this section we study the dynamics of (2.3) when $\beta(x)$ is the step function

$$\beta(x) = \begin{cases} 0, & x \geq 1, \\ \beta_0, & x < 1. \end{cases}$$

By a solution of (2.3) in this case, we mean a continuous function $x(t)$ on the interval $[-\tau, \infty)$ which is piecewise differentiable and satisfies (2.3) for $t \in [0, \infty)$ except at the point t where $x(t)$ or $x(t - \tau)$ is equal to 1. For any initial data $\varphi \in C[-\tau, 0]$, it is not difficult to obtain a unique solution $x(t)$ by using the method of steps. As in [24], we introduce two constants

$$\alpha = \beta_0 + \delta, \quad \Gamma = 2\beta_0 e^{-\gamma\tau} = k\beta_0.$$

Inserting the step function $\beta(x)$ into (2.3), we obtain

$$\frac{dx}{dt} = \begin{cases} -\delta x, & 1 < x, x_\tau, \\ -\alpha x, & 0 < x < 1 < x_\tau, \\ -\alpha x + \Gamma x_\tau, & 0 < x, x_\tau < 1, \\ -\delta x + \Gamma x_\tau, & 0 < x_\tau < 1 < x, \end{cases} \tag{3.1}$$

where $x_\tau = x(t - \tau)$.

For (3.1), we choose the initial function $\varphi(t) \geq 1+\eta$ for $t \in [-\tau, 0)$ and $\varphi(0) = 1+\eta$ where η is a small positive constant specified later. By the continuity of the solution x, we have from (3.1) the existence of t_1 such that $x(t)$ and $x(t-\tau)$ are greater than 1 for $t \in [0, t_1)$ and $x(t_1) = 1$. The solution $x(t)$ then satisfies

$$\frac{dx}{dt} = -\delta x \quad \text{for} \quad t \in [0, t_1]. \tag{3.2}$$

Thus solving the above equation, we have $x(t) = \varphi(0)e^{-\delta t} = (1+\eta)e^{-\delta t}$. It follows that

$$t_1 = \frac{\ln \varphi(0)}{\delta} = \frac{\ln(1+\eta)}{\delta}. \tag{3.3}$$

In the next interval of time, defined by $(t_1, t_1+\tau)$, we have $x(t-\tau) > 1$. From the first two lines in (3.1), the solution is decreasing and thus crosses the level $x = 1$. The dynamics are given by

$$\frac{dx}{dt} = -\alpha x \tag{3.4}$$

as long as $x(t) < 1$. The solution is then given by $x(t) = e^{-\alpha(t-t_1)}$ for $t \in [t_1, t_1+\tau]$ and $x(t_1+\tau) = e^{-\alpha\tau}$ independent of the initial function $\varphi(t)$. Thus, the dynamics eventually destroy all memory of the initial function.

The solution in the next interval will be such that $x, x_\tau < 1$. In order that (3.1) has periodic solutions, we impose an extra condition on Γ and α so that

$$-\alpha x + \Gamma x_\tau > 0. \tag{3.5}$$

Otherwise, if $-\alpha x + \Gamma x_\tau \leq 0$, then the solution may tend to zero as t approaches infinity and thus we cannot expect a periodic solution. In particular, if

$$-\alpha x + \Gamma x_\tau \approx 0,$$

then the solution may stay below the line $x = 1$ so long that the resulting analysis becomes very complicated. Note that for $t \in [t_1+\tau, t_1+2\tau]$, we have $x(t-\tau) = e^{-\alpha(t-t_1-\tau)}$. Then if $x(t) < 1$, from (3.1), we have $\frac{dx}{dt} = -\alpha x + \Gamma x_\tau = -\alpha x + \Gamma e^{-\alpha(t-t_1-\tau)}$ which gives

$$x(t) = e^{-\alpha(t-t_1-\tau)}(e^{-\alpha\tau} + \Gamma(t-t_1-\tau)). \tag{3.6}$$

For the sake of simplicity, we impose an extra condition on Γ:

$$\Gamma > \max\left\{\frac{1}{\tau}(e^{\alpha\tau} - e^{-\alpha\tau}), \alpha e^{\alpha\tau}\right\}. \tag{3.7}$$

Note that condition (3.7) clearly holds if β_0 is large.

Equation (3.6) is only valid if the value of $x(t)$ is less than or equal to 1. However when we directly replace t in (3.6) by $t_1+2\tau$, we have $x(t_1+2\tau) = e^{-\alpha\tau}(e^{-\alpha\tau}+\Gamma\tau) > 1$. Thus we need to use (3.6) to find a point $t_2 \in (t_1+\tau, t_1+2\tau)$ such that $x(t_2) = 1$ and (3.6) is valid for $t \in [t_1+\tau, t_2]$. Assume $t_2 = t_1+\tau+u$, $u \in (0, \tau)$. Then from (3.6) we have

$$e^{\alpha u} = e^{-\alpha\tau} + \Gamma u. \tag{3.8}$$

Equation (3.8) is a transcendental equation and cannot be solved explicitly. However, the existence of a positive solution $u \in (0, \tau)$ is obvious given (3.7). Therefore (3.5) holds for $t \in [t_1 + \tau, t_2]$ (due to the fact that $x(t-\tau) \geq e^{-\alpha\tau}$).

Next for $t \in (t_2, t_2 + \tau)$, we claim that

$$x(t) > 1. \tag{3.9}$$

Indeed, from the above analysis, we know that $e^{-\alpha\tau} < x(t-\tau) < 1$ and at the particular point t_2, $x(t_2 + 0) = \lim_{t \to t_2 + 0} x(t) = 1$, so $x(t_2 - \tau) \geq e^{-\alpha\tau}$. By (2.3) and (3.7) we have

$$x'(t_2 + 0) = -[\beta(x) + \delta]x + k\beta(x_\tau)x_\tau > -\alpha + \Gamma x_\tau > 0.$$

The solution $x(t)$ is differentiable with respect to t as long as $x(t)$ and $x(t-\tau)$ are not equal to 1. To see our claim suppose, by contradiction, that there exists a point $h \in (t_2, t_2 + \tau)$ such that $x(h) = 1$, $x'(h - 0) \leq 0$, and $x(t) > 1$ for $t \in (t_2, h)$. Then using (3.1), we have by (3.7) that

$$x'(h - 0) = -\delta + \Gamma x(h - \tau) \geq -\delta + \Gamma e^{-\alpha\tau} > 0.$$

This is a contradiction, and our claim is true.

Splitting $[t_2, t_2 + \tau]$ into two subintervals $[t_2, t_1 + 2\tau]$ and $[t_1 + 2\tau, t_2 + \tau]$, we can give explicit formulae for the solution $x(t)$ as follows.

For $t \in [t_2, t_1 + 2\tau]$, we know that $x(t-\tau) = e^{-\alpha(t - t_1 - \tau)} < 1$. The dynamics are thus given by

$$\frac{dx}{dt} = -\delta x + \Gamma x_\tau = -\delta x + \Gamma e^{-\alpha(t - t_1 - \tau)},$$

which has the solution

$$x(t) = e^{-\delta\tau(t - t_2)} \left\{ 1 - \frac{\Gamma}{\beta_0} e^{\alpha(t_1 + \tau) - \delta t_2} \left(e^{-\beta_0 t} - e^{-\beta_0 t_2} \right) \right\}. \tag{3.10}$$

Moreover, since the solutions are differentiable provided that $x(t)$ and $x(t-\tau)$ are not equal to 1, and the solutions are continuous everywhere, for $t \in [t_1 + 2\tau, t_2 + \tau]$ we have

$$\begin{aligned} \frac{dx}{dt} &= -\delta x + \Gamma x_\tau \\ &= -\delta x + \Gamma e^{-\alpha(t - t_1 - 2\tau)}(e^{-\alpha\tau} + \Gamma(t - t_1 - 2\tau)), \end{aligned}$$

so

$$x(t) = e^{-\delta(t - t_1 - 2\tau)} \left[x(t_1 + 2\tau) + \Gamma \left(j(t) - j(t_1 + 2\tau) \right) \right],$$

where

$$j(t) = \frac{1}{(\delta - \alpha)} \left(e^{-\alpha\tau} + \Gamma(t - t_1 - 2\tau) - \frac{\Gamma}{\delta - \alpha} \right) e^{(\delta - \alpha)(t - t_1 - 2\tau)}.$$

After the time $t_2 + \tau$, both x_τ and x are greater than 1, and the solution satisfies

$$x' = -\delta x \tag{3.11}$$

as long as $x(t) > 1$ and thus is decreasing. Therefore, there exists a point, say, $t = d$, so that $x(d) = 1$. Note that in the interval $[t_2, d]$, the graph of the solution $x(t)$ is

independent of the initial function $\varphi(t)$. Now we can use (3.9) and (3.11) to choose a small positive constant $\eta < 1$ such that the following hold:

1. We have

$$t_1 = \frac{\log(1+\eta)}{\delta} < \tau; \tag{3.12}$$

2. we have

$$\Gamma > \max\left\{\frac{1}{\tau}(e^{\alpha\tau} - e^{-\alpha\tau}), \alpha(1+\eta)e^{\alpha\tau}\right\}, \tag{3.13}$$

and $x(t)$ reaches $1+\eta$ at a point $t_3 \in (t_2, t_2+\tau)$; and

3. there is a point T_x, $t_3 + \tau < T_x < d$ so that

$$x(T_x) = 1+\eta,\ x(T_x+s) > 1+\eta,\ s \in [-\tau, 0). \tag{3.14}$$

With this choice of η, we have $x(t) > 1+\eta$ for $t \in (t_3, T_x)$ and the solution is strictly increasing for $t \in [t_2, t_3]$ (due to (3.13)). Finally, when we continue to solve (3.1) step by step, we have $x(t) = x(t+T_x)$ for $t \geq 0$. Summarizing the above analysis, we have the following result.

THEOREM 3.1. *Suppose that* Γ *satisfies* (3.7). *Assume that* x *is the solution of* (3.1) *subject to the initial condition* $\phi \geq 1+\eta$ *where* η *is chosen to satisfy* (3.12), (3.13), *and* (3.14). *Then the solution* x *satisfies* $x(t) = x(t+T_x)$ *for* $t \geq 0$.

4. Periodic solutions: General nonlinearity.

4.1. Perturbed delay equation. With the preceding analysis of the G_0 phase cell cycle model when the feedback function β is a Heaviside step function, we turn to a consideration of the general continuous nonlinearity. More precisely, we consider

$$\frac{dy}{dt} = -[\beta(y) + \delta]y + k\beta(y_\tau)y_\tau, \tag{4.1}$$

returning to the original problem with $\beta = \beta_0 \frac{1}{1+y^n}$. Let $\varepsilon = 1/n$. Then we can rewrite the Hill function as

$$\beta_\varepsilon(y) = \beta_0 \frac{1}{1+y^{1/\varepsilon}}.$$

Let the initial function φ be chosen from the closed convex set

$$A_\eta = \{\varphi \in C([-\tau, 0]) : 1+\eta \leq \varphi(t) \quad \text{for} \quad t \in [-\tau, 0], \quad \text{and} \quad \varphi(0) = 1+\eta\},$$

where $\eta < 1$ is a small positive constant as chosen in the previous section. For given φ in A_η, we have a unique solution to (4.1). The relations

$$F_\varepsilon(t, \varphi) = y_t,\ y_t = y(t+s),\ -\tau \leq s \leq 0,\ t \geq 0,$$

define a continuous semiflow $F = F_\varepsilon$ on $C([-\tau, 0])$.

As a technical preparation, we now describe some elementary properties of the Hill function we employ here.

LEMMA 4.1. *Assume* $\varepsilon = \frac{1}{n} < 1$. *The following inequalities hold:*
(a) *If* $y > \left(\frac{1}{\varepsilon}\right)^{\varepsilon/(1-\varepsilon)}$, *then*

$$\beta_\varepsilon(y) < \beta_0\varepsilon, \quad y\beta_\varepsilon(y) < \beta_0\varepsilon$$

and if $0 < y < \varepsilon^\varepsilon$, *then*

$$\beta_0 > \beta_\varepsilon(y) > \beta_0(1-\varepsilon) \quad \textit{and} \quad |y\beta_\varepsilon(y) - \beta_0 y| < \beta_0\epsilon. \tag{4.2}$$

(b) *Also,*

$$\left|\frac{d(y\beta_\varepsilon(y))}{dy}\right| < \beta_0\varepsilon \quad \textit{for } y > \left(\frac{1}{\varepsilon}\right)^{2\varepsilon},$$

and

$$\left|\frac{d(y\beta_\varepsilon(y) - \beta_0 y)}{dy}\right| < \beta_0\varepsilon \quad \textit{for } 0 < y < \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^{\varepsilon}.$$

Proof. (a) If $y > \left(\frac{1}{\varepsilon}\right)^{\varepsilon/(1-\varepsilon)}$, then

$$\beta_\varepsilon(y) = \frac{\beta_0}{1+y^{1/\varepsilon}} < \frac{\beta_0}{y^{1/\varepsilon}} < \frac{\beta_0}{\left(\frac{1}{\varepsilon}\right)^{1/(1-\varepsilon)}} < \beta_0\varepsilon,$$

and

$$y\beta_\varepsilon(y) = \frac{\beta_0 y}{1+y^{1/\varepsilon}} < \frac{\beta_0}{y^{\frac{1}{\varepsilon}-1}} < \beta_0\varepsilon.$$

If $0 < y < \varepsilon^\varepsilon$, then

$$\beta_0 > \beta_\varepsilon(y) = \frac{\beta_0}{1+y^{1/\varepsilon}} > \beta_0(1-y^{1/\varepsilon}) \geq \beta_0(1-\varepsilon),$$

and

$$|y\beta_\varepsilon(y) - \beta_0 y| = \left|\beta_0 \frac{y^{1/\varepsilon+1}}{1+y^{1/\varepsilon}}\right| < \beta_0 y^{1/\varepsilon+1} < \beta_0\varepsilon.$$

(b) If $y > (1/\varepsilon)^{2\varepsilon}$, then

$$\left|\frac{d\,(y\beta_\varepsilon(y))}{dy}\right| = \beta_0 \frac{\left|\left(\frac{1}{\varepsilon}-1\right)y^{1/\varepsilon} - 1\right|}{(1+y^{1/\varepsilon})^2} \leq \beta_0\left(\frac{1}{\varepsilon}-1\right)y^{-1/\varepsilon} < \beta_0\varepsilon.$$

Since

$$f(x) = \frac{\left(1+\frac{1}{\varepsilon}\right)x + \frac{1}{\varepsilon}x^2}{1+x}$$

is strictly increasing for $x \in (0, \frac{\varepsilon^2}{1+\varepsilon})$ and $f(\frac{\varepsilon^2}{1+\varepsilon}) < \varepsilon$, we obtain

$$\left|\frac{d\,(y\beta_\varepsilon(y) - \beta_0 y)}{dy}\right| = \beta_0 \frac{(1+\frac{1}{\varepsilon})y^{1/\varepsilon} + \frac{1}{\varepsilon}y^{2/\varepsilon}}{1+y^{1/\varepsilon}} < \beta_0\varepsilon$$

for $0 < y < (\frac{\varepsilon^2}{1+\varepsilon})^\varepsilon$. ∎

We found that for (3.1), if $\varphi \in A_\eta$, then the solution will return to A_η after finite time. The following lemma shows a similar property for (4.1).

LEMMA 4.2. *Let y be the solution of* (4.1) *with an initial function $\varphi \in A_\eta$. Then there exists a point $T_y > 0$ such that $y(T_y) = 1 + \eta$ and*

$$y(t) \geq 1 + \eta \quad \text{for } t \in [T_y - \tau, T_y]. \tag{4.3}$$

Moreover, there exists a constant $\varepsilon_1, \varepsilon_1 \in (0,1)$, such that for each $\varepsilon \in (0, \varepsilon_1)$, we have

$$T_y = T_x + O(\varepsilon \log \varepsilon) \tag{4.4}$$

and

$$y(t) = x(t) + O(\varepsilon \log \varepsilon), \tag{4.5}$$

uniformly for $t \in [0, T_x]$ and $\varphi \in A_\eta$, where T_x is the period of the periodic solution x to (3.1) *obtained in Theorem* 3.1.

Proof of Lemma 4.2. We first claim that there exist three points η_1, t_1^y, η_2, $0 < \eta_1 < t_1^y < \eta_2$, which are dependent on ε and φ, such that

$$y(\eta_1) = \left(\frac{1}{\varepsilon}\right)^{2\varepsilon} > 1, \;\; y(t_1^y) = 1, \; y(\eta_2) = \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^{\varepsilon} < 1. \tag{4.6}$$

Indeed, if $y(t) > (\frac{1}{\varepsilon})^{2\varepsilon} > \left(\frac{1}{\varepsilon}\right)^{\varepsilon/(1-\varepsilon)} > 1$ and $y(t-\tau) > (\frac{1}{\varepsilon})^{2\varepsilon} > (\frac{1}{\varepsilon})^{\varepsilon/(1-\varepsilon)}$, then we have by Lemma 4.1 that

$$\beta_\varepsilon(y(t))y(t) < \beta_0 \varepsilon, \;\; \beta_\varepsilon(y(t-\tau))y(t-\tau) < \beta_0 \varepsilon$$

and

$$\begin{aligned} \frac{dy(t)}{dt} &= -(\delta + \beta_\varepsilon(y(t)))y(t) + k\beta_\varepsilon(y(t-\tau))y(t-\tau), \\ &= -\delta y(t) + O(\varepsilon) \\ &< -\frac{\delta}{2} \text{ for } \varepsilon \in (0, \sigma_1). \end{aligned} \tag{4.7}$$

Here σ_1 is chosen so that for each $\varepsilon \in (0, \sigma_1)$, we have $-\delta y(t) + O(\varepsilon) < -\frac{\delta}{2}$. This means that y is decreasing as long as $y(t) \geq (\frac{1}{\varepsilon})^{2\varepsilon} > (\frac{1}{\varepsilon})^{\varepsilon/(1-\varepsilon)}$. Therefore there is a point $\eta_1 > 0$ so that $y(\eta_1) = (1/\varepsilon)^{2\varepsilon}$ and $1 + \eta > y(t) > (1/\varepsilon)^{2\varepsilon}$ for $t \in (0, \eta_1)$. Using $\frac{dy}{dt} = -\delta y + O(\varepsilon)$ and $y(0) = 1 + \eta$, we also have

$$\eta_1 = \frac{\log(1+\eta)}{\delta} + O(-\varepsilon \log \varepsilon) = t_1 + O(-\varepsilon \log \varepsilon). \tag{4.8}$$

Here the term $O(-\varepsilon \log \varepsilon)$ holds uniformly for all the initial functions φ in A_η. Next in the interval $(\eta_1, \eta_1 + \tau)$, we have $\beta_\varepsilon(y(t-\tau))y(t-\tau) = O(\varepsilon)$ and

$$\begin{aligned} \frac{dy(t)}{dt} &= -(\delta + \beta_\varepsilon(y(t)))y(t) + k\beta_\varepsilon(y(t-\tau))y(t-\tau) \\ &< -\delta y(t) + k\beta_\varepsilon(y(t-\tau))y(t-\tau) \\ &= -\delta y(t) + O(\varepsilon) \\ &< -\frac{\delta}{2} \text{ for } \varepsilon \in (0, \sigma_2) \end{aligned}$$

as long as $y(t) \geq (\frac{\varepsilon^2}{1+\varepsilon})^{\varepsilon}$. Here σ_2 is chosen so that for each $\varepsilon \in (0, \sigma_2)$, we have

$$-\delta y(t) + O(\varepsilon) < -\delta \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^{\varepsilon} + O(\varepsilon) < -\frac{\delta}{2}.$$

This means that the solution is decreasing and there exist two points t_1^y, η_2, $\eta_1 < t_1^y < \eta_2$, such that

$$y(t_1^y) = 1, \;\; y(\eta_2) = \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^{\varepsilon}.$$

By the mean value theorem, it is easy to show that

$$|y(\eta_1) - y(\eta_2)| \geq \frac{\delta}{2} |\eta_1 - \eta_2|$$

or, equivalently,

$$\eta_2 - \eta_1 \leq \frac{2}{\delta}(y(\eta_1) - y(\eta_2)) = \frac{2}{\delta}\left[\left(\frac{1}{\varepsilon}\right)^{2\varepsilon} - \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^{\varepsilon}\right] = O(-\varepsilon \log \varepsilon).$$

Therefore,

$$0 < t_1^y - \eta_1 < \eta_2 - \eta_1 = O(-\varepsilon \log \varepsilon). \tag{4.9}$$

Now using (4.1) for $t \in [0, \eta_1]$, we have

$$y' = -\delta y + O(\varepsilon), \;\; y(0) = 1 + \eta,$$

which gives

$$y(t) = (1+\eta)e^{-\delta t} + O(\varepsilon).$$

We claim that

$$y(t) = x(t) + O(\varepsilon) \tag{4.10}$$

uniformly for $t \in [0, \xi_1]$ and $\varphi \in A_\eta$, where

$$\xi_1 = \min\{t_1, \eta_1\}.$$

Indeed, this is true, since $x(t) = (1+\eta)e^{-\delta t}$ for $t \in [0, t_1]$.

Next for $t \in [\xi_1, \eta_2]$, using an argument that the length of the interval $[t_1, \eta_1]$ is of order $O(-\varepsilon \log \varepsilon)$, and both $|x'(t)|$ and $|y'(t)|$ are bounded by a constant, say, M, which is independent of ε and η, we conclude from (4.10) that

$$y(t) = x(t) + O(-\varepsilon \log \varepsilon). \tag{4.11}$$

For $t \in [\eta_2, \tau + \xi_1]$, we can show that $y(t-\tau) > (1/\varepsilon)^{2\varepsilon}$. Note that $\eta_2 \leq \xi_1 + \tau$ since $\eta_2 - \xi_1 = O(\varepsilon \log \varepsilon)$ and τ is a constant. Here we have assumed that $\varepsilon \in (0, \sigma_3)$, where σ_3 is small enough so that for each $\varepsilon \in (0, \sigma_3)$, we have $O(\varepsilon \log \varepsilon) < \tau$. By Lemma 4.1, we have

$$y(t-\tau)\beta(y(t-\tau)) = O(\varepsilon).$$

Using (4.1) we know that

$$-\alpha y + O(\varepsilon) \leq y' \leq -\delta y + O(\varepsilon),$$

and thus the solution $y(t) \geq (\varepsilon^2/(1+\varepsilon))^\varepsilon e^{-\alpha\tau} + O(\varepsilon)$ and its derivative

$$y'(t) \leq -\delta(\varepsilon^2/(1+\varepsilon))^\varepsilon e^{-\alpha\tau} + O(\varepsilon) < 0 \text{ for } \varepsilon \in (0, \sigma_4),$$

where σ_4 is chosen so that for each $\varepsilon \in (0, \sigma_4)$, we have $-\delta(\varepsilon^2/(1+\varepsilon))^\varepsilon e^{-\alpha\tau} + O(\varepsilon) < 0$. So $y(t)$ is decreasing for $t \in [\eta_2, \tau + \xi_1]$. Note that $0 < y < y(\eta_2) \leq \varepsilon^\varepsilon$ so that (4.2) in Lemma 4.1 holds. Thus we can derive from (4.1) that

$$y'(t) = -\alpha y(t) + O(\varepsilon) \tag{4.12}$$

for $t \in [\eta_2, \tau + \xi_1]$. Coupling this equation with (3.4) and using (4.11) at the point $t = \eta_2$ give

$$y(t) = x(t) + O(-\varepsilon \log \varepsilon)$$

for $t \in [\eta_2, \tau + \xi_1]$.

For $t \in [\tau + \xi_1, \tau + \eta_2]$, again using the fact that both the derivatives of x and y are bounded and the length of this interval is of order $O(-\varepsilon \log \varepsilon)$, we have

$$y(t) = x(t) + O(-\varepsilon \log \varepsilon).$$

For $t \geq \tau + \eta_2$, the solution y begins to increase since Γ satisfies (3.7). To be precise, we have $\beta_\varepsilon(y(t)) < \beta_0$, $\beta_\varepsilon(y(t-\tau))y(t-\tau) = \beta_0 y(t-\tau) + O(-\varepsilon \log \varepsilon)$ and

$$\begin{aligned} y'(t) &= -(\delta + \beta_\varepsilon(y(t)))y(t) + k\beta_\varepsilon(y(t-\tau))y(t-\tau) \\ &\geq -\alpha y(t) + k\beta_0 y(t-\tau) + O(-\varepsilon \log \varepsilon) \\ &= -\alpha y(t) + \Gamma x(t-\tau) + O(-\varepsilon \log \varepsilon) \\ &\geq -\alpha(1+\eta) + \Gamma e^{-\alpha\tau} + O(-\varepsilon \log \varepsilon) \\ &> 0 \quad \text{for } \varepsilon \in (0, \sigma_5) \end{aligned}$$

as long as $y(t) \leq 1 + \eta$ and $t \leq 2\tau + \eta_2$. Here σ_5 is sufficiently small so that for each $\varepsilon \in (0, \sigma_5)$, we have $-\alpha(1+\eta) + \Gamma e^{-\alpha\tau} + O(-\varepsilon \log \varepsilon) > 0$. Using similar arguments as above, we conclude that there exist three points η_3, t_2^y, η_4, with $\eta_3 < t_2^y < \eta_4$ such that

$$y(\eta_3) = \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^\varepsilon, \quad y(t_2^y) = 1, \quad y(\eta_4) = \left(\frac{1}{\varepsilon}\right)^{2\varepsilon},$$

$$\eta_3 = t_2^y + O(-\varepsilon \log \varepsilon), \quad \eta_4 = t_2^y + O(-\varepsilon \log \varepsilon), \tag{4.13}$$

and

$$t_2^y = t_2 + O(-\varepsilon \log \varepsilon). \tag{4.14}$$

We can continue this process to find that y will satisfy

$$y(t) = x(t) + O(-\varepsilon \log \varepsilon)$$

for $t \in [0, \tau + \xi_2]$, where $\xi_2 = \min\{t_2, \eta_3\}$. From the expression for x, we know from the preceding equation that there exists a point $t_3^y \in (\eta_4, \tau+\xi_2)$ such that $y(t_3^y) = 1+\eta$ and $t_3^y = t_3 + O(-\varepsilon \log \varepsilon)$.

For $t \in [\tau+\xi_2, \tau+\eta_4]$, using the same argument as in the interval $[\tau+\xi_1, \tau+\eta_2]$, we again have

$$y(t) = x(t) + O(-\varepsilon \log \varepsilon). \tag{4.15}$$

Finally for $t \geq \tau+\eta_4$, the solution is decreasing and will reach the value $1+\eta$ at some point T_y. In the whole interval $[0, T_x]$, if we choose $\varepsilon_1 = \min\{\sigma_i, 1 \leq i \leq 5\}$, then we can show as before that for each $\varepsilon \in (0, \varepsilon_1)$, we have

$$y(t) = x(t) + O(-\varepsilon \log \varepsilon),\ x \in [0, T_x], \tag{4.16}$$

and

$$T_y = T_x + O(-\varepsilon \log \varepsilon). \tag{4.17}$$

Furthermore, we also have $y(T_y) = 1 + \eta$ and

$$y(t) \geq 1 + \eta \text{ for } [T_y - \tau, T_y]. \qquad \square \tag{4.18}$$

Remark 4.3. By Lemma 4.2 and (4.1) we have two positive constants M_1 and M_2 which are independent of ε and the initial data φ so that for $t \geq 0$,

$$|y(t)| \leq M_1 \tag{4.19}$$

and

$$\left| \frac{dy(t)}{dt} \right| \leq M_2. \tag{4.20}$$

Now we are ready to define a continuous return map

$$R : A_\eta \ni \varphi \to y_{q(\varphi)} = F_\varepsilon(q(\varphi), \varphi) \in A_\eta,$$

where $q(\varphi) = T_y$. To verify that there exists a unique fixed point in A_η for the map R, we need to show that the map R is contractive, i.e., derive an estimation for the Lipschitz constant and show that the Lipschitz constant is less than 1.

4.2. Lipschitz constant for the map $\mathbb{R}$. The Lipschitz constant of a given map $T : D_T \to Y$, $D_T \subset X$, where X and Y are normed linear spaces, is given by

$$L(T) = \sup_{u \in D_T, v \in D_T, u \neq v} \frac{||T(u) - T(v)||}{||u - v||}.$$

In the case where $D_T = X = Y = \mathbb{R}$, $[u_1, u_2] \subset \mathbb{R}$, and $f = T$, we set

$$L_{[u_1,u_2]}(f) = L(f|[u_1, u_2]).$$

If $f(u) = u\beta_\varepsilon(u)$, $u \in \mathbb{R}$, we define the following four Lipschitz constants:

$$\begin{aligned} L_1^\varepsilon &= L_{[1+\eta, +\infty)}(u\beta_\varepsilon(u)), \\ L_2^\varepsilon &= L_{[(\frac{1}{\varepsilon})^{2\varepsilon}, +\infty)}(u\beta_\varepsilon(u)), \\ L_3^\varepsilon &= L_{(0, +\infty)}(u\beta_\varepsilon(u)), \\ L_4^\varepsilon &= L_{(0, (\frac{\varepsilon^2}{1+\varepsilon})^\varepsilon)}(u\beta_\varepsilon(u)). \end{aligned}$$

Similarly for the function $f(u) = u\beta_\varepsilon(u) - \beta_0 u$, $u \in \mathbb{R}$, we define the following Lipschitz constant for later use:

$$L_5^\varepsilon = L_{(0,(\frac{\varepsilon^2}{1+\varepsilon})^\varepsilon)}(u\beta_\varepsilon(u) - \beta_0 u).$$

When $\varepsilon << 1$, we have

$$L_1^\varepsilon = O\left(\frac{1}{\varepsilon(1+\eta)^{1/\varepsilon}}\right), \quad L_2^\varepsilon = O(\varepsilon), \quad L_3^\varepsilon = O(1/\varepsilon), \quad L_4^\varepsilon = O(1), \quad L_5^\varepsilon = O(\varepsilon). \tag{4.21}$$

THEOREM 4.4. *There exists ε_2, $\varepsilon_2 \in (0, \varepsilon_1)$, such that for each $\varepsilon \in (0, \varepsilon_2)$ the Lipschitz constant L_R^ε of the map R is less than 1. In particular, we have*

$$\lim_{\varepsilon \to 0} L_R^\varepsilon = 0.$$

Proof. Step 1. Take $\phi, \bar{\phi}$ in A_η. Using a similar argument as in the proof of Lemma 4.2, we conclude that there exist η_1, η_2 and $\bar{\eta}_1, \bar{\eta}_2$ such that

$$y^\phi(\eta_1) = \left(\frac{1}{\varepsilon}\right)^{2\varepsilon}, \; y^\phi(\eta_2) = \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^\varepsilon, \; \eta_1 - \eta_2 = O(-\varepsilon \log \varepsilon)$$

and

$$y^{\bar{\phi}}(\bar{\eta}_1) = \left(\frac{1}{\varepsilon}\right)^{2\varepsilon}, \; y^{\bar{\phi}}(\bar{\eta}_2) = \left(\frac{\varepsilon^2}{1+\varepsilon}\right)^\varepsilon, \; \bar{\eta}_1 - \bar{\eta}_2 = O(-\varepsilon \log \varepsilon).$$

Let

$$\eta_{\min} = \min\{\eta_1, \bar{\eta}_1\}$$

and

$$\eta_{\max} = \max\{\eta_2, \bar{\eta}_2\}.$$

Then by (4.8) and (4.9) we have

$$\eta_{\min} = t_1 + O(-\varepsilon \log \varepsilon), \quad \eta_{\max} = t_1 + O(-\varepsilon \log \varepsilon), \quad \text{and} \tag{4.22}$$
$$\eta_{\max} - \eta_{\min} = O(-\varepsilon \log \varepsilon).$$

Since $t_1 = \log(1+\eta)/\delta < \tau$, from (4.22) we have that $\eta_{\min} < \tau$ and $\eta_{\max} < \tau$. Here we have chosen $\sigma_6 > 0$ sufficiently small so that for each $\varepsilon \in (0, \sigma_6)$

$$\eta_{\max} = \log(1+\eta)/\delta + O(-\varepsilon \log \varepsilon) < \tau.$$

For $t \in [0, \eta_{\min}]$, using (4.1) for $y^\phi(t)$ and $y^{\bar{\phi}}(t)$ gives

$$\frac{dy^\phi(t)}{dt} = -[\delta + \beta_\varepsilon(y^\phi(t))]y^\phi(t) + k\beta_\varepsilon(y^\phi(t-\tau))y^\phi(t-\tau) \tag{4.23}$$

and

$$\frac{dy^{\bar{\phi}}(t)}{dt} = -[\delta + \beta_\varepsilon(y^{\bar{\phi}}(t))]y^{\bar{\phi}}(t) + k\beta_\varepsilon(y^{\bar{\phi}}(t-\tau))y^{\bar{\phi}}(t-\tau). \tag{4.24}$$

Now we estimate the difference between $y^{\phi}(t)$ and $y^{\bar{\phi}}(t)$. Subtracting (4.24) from (4.23) yields

$$\begin{aligned}(4.25)\quad (y^{\phi}(t)-y^{\bar{\phi}}(t))' &= -\delta(y^{\phi}(t)-y^{\bar{\phi}}(t))\\ &\quad -[\beta_{\varepsilon}(y^{\phi}(t))y^{\phi}(t)-\beta_{\varepsilon}(y^{\bar{\phi}}(t))y^{\bar{\phi}}(t)]\\ &\quad +k[\beta_{\varepsilon}(y^{\phi}(t-\tau))y^{\phi}(t-\tau)-\beta_{\varepsilon}(y^{\bar{\phi}}(t-\tau))y^{\bar{\phi}}(t-\tau)].\end{aligned}$$

Substituting the inequalities

$$|\beta_{\varepsilon}(y^{\phi}(t))y^{\phi}(t)-\beta_{\varepsilon}(y^{\bar{\phi}}(t))y^{\bar{\phi}}(t)|\le L_2^{\varepsilon}|y^{\phi}(t)-y^{\bar{\phi}}(t)|$$

and

$$|\beta_{\varepsilon}(y^{\phi}(t-\tau))y^{\phi}(t-\tau)-\beta_{\varepsilon}(y^{\bar{\phi}}(t-\tau))y^{\bar{\phi}}(t-\tau)|\le L_1^{\varepsilon}||\phi-\bar{\phi}||$$

into (4.25), we have

$$(4.26)\qquad (y^{\phi}(t)-y^{\bar{\phi}}(t))'\le(\delta+L_2^{\varepsilon})\,|y^{\phi}(t)-y^{\bar{\phi}}(t)|+kL_1^{\varepsilon}||\phi-\bar{\phi}||.$$

Integrating (4.26) from 0 to t gives

$$(y^{\phi}(t)-y^{\bar{\phi}}(t))\le\int_0^t\Big((\delta+L_2^{\varepsilon})\,|y^{\phi}(s)-y^{\bar{\phi}}(s)|+kL_1^{\varepsilon}||\phi-\bar{\phi}||\Big)\,ds.$$

Similarly, we have

$$-(y^{\phi}(t)-y^{\bar{\phi}}(t))\le\int_0^t\Big((\delta+L_2^{\varepsilon})\,|y^{\phi}(s)-y^{\bar{\phi}}(s)|+kL_1^{\varepsilon}||\phi-\bar{\phi}||\Big)\,ds.$$

Thus, we have found that

$$(4.27)\qquad |y^{\phi}(t)-y^{\bar{\phi}}(t)|\le\int_0^t\Big((\delta+L_2^{\varepsilon})\,|y^{\phi}(s)-y^{\bar{\phi}}(s)|+kL_1^{\varepsilon}||\phi-\bar{\phi}||\Big)\,ds.$$

From Gronwall's inequality, we obtain

$$(4.28)\qquad |y^{\phi}(t)-y^{\bar{\phi}}(t)|\le C_1||\phi-\bar{\phi}||,$$

where

$$(4.29)\qquad C_1=\frac{e^{(\delta+L_2^{\varepsilon})\eta_{\min}}-1}{\delta+L_2^{\varepsilon}}kL_1^{\varepsilon}.$$

Step 2. For $t\in[\eta_{\min},\eta_{\max}]$, we have

$$|\beta_{\varepsilon}(y^{\phi}(t))y^{\phi}(t)-\beta_{\varepsilon}(y^{\bar{\phi}}(t))y^{\bar{\phi}}(t)|\le L_3^{\varepsilon}|y^{\phi}(t)-y^{\bar{\phi}}(t)|$$

and

$$|\beta_{\varepsilon}(y^{\phi}(t-\tau))y^{\phi}(t-\tau)-\beta_{\varepsilon}(y^{\bar{\phi}}(t-\tau))y^{\bar{\phi}}(t-\tau)|\le L_1^{\varepsilon}||\phi-\bar{\phi}||.$$

Thus from (4.23) and (4.24) we obtain, as before,

$$|y^{\phi}(t)-y^{\bar{\phi}}(t)|\le\int_{\eta_{\min}}^t\Big((\delta+L_3^{\varepsilon})\,|y^{\phi}(s)-y^{\bar{\phi}}(s)|+kL_1^{\varepsilon}||\phi-\bar{\phi}||\Big)\,ds+C_1||\phi-\bar{\phi}||.$$

Then by Gronwall's inequality, we have

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \leq C_2 ||\phi - \bar{\phi}||, \tag{4.30}$$

where

$$C_2 = C_1 e^{(\delta + L_3^{\varepsilon})(\eta_{\max} - \eta_{\min})} + \frac{e^{(\delta + L_3^{\varepsilon})(\eta_{\max} - \eta_{\min})} - 1}{\delta + L_3^{\varepsilon}} k L_1^{\varepsilon} > C_1. \tag{4.31}$$

Remember that $\eta_{\min} \leq \tau$ since $t_1 < \tau$ in (3.14) and $\eta_{\min} = t_1 + O(-\varepsilon \log \varepsilon)$. Moreover $\eta_{\max} \leq \tau$ since $\eta_{\max} = t_1 + O(-\varepsilon \log \varepsilon)$ from (4.22).

Step 3. For $t \in [\eta_{\max}, \tau + \eta_{\min}]$,

$$|\beta_{\varepsilon}(y^{\phi}(\tau))y^{\phi}(t) - \beta_0 y^{\phi}(t) - (\beta_{\varepsilon}(y^{\bar{\phi}}(t))y^{\bar{\phi}}(t) - \beta_0 y^{\bar{\phi}}(t))| \leq L_5^{\varepsilon} |y^{\phi}(t) - y^{\bar{\phi}}(t)|$$

and

$$|\beta_{\varepsilon}(y^{\phi}(t-\tau))y^{\phi}(t-\tau) - \beta_{\varepsilon}(y^{\bar{\phi}}(t-\tau))y^{\bar{\phi}}(t-\tau)| \leq L_2^{\varepsilon} C_2 ||\phi - \bar{\phi}||.$$

It is thus easy to derive

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \leq \int_{\eta_{\max}}^{t} \left((\alpha + L_5^{\varepsilon}) |y^{\phi}(s) - y^{\bar{\phi}}(s)| + k L_2^{\varepsilon} C_2 ||\phi - \bar{\phi}|| \right) ds + C_2 ||\phi - \bar{\phi}||$$

and to conclude that (since $\tau + \eta_{\min} - \eta_{\max} < \tau$)

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \leq C_3 ||\phi - \bar{\phi}||, \tag{4.32}$$

where

$$C_3 = C_2 e^{\alpha\tau + \tau L_5^{\varepsilon}} + \frac{e^{\alpha\tau + \tau L_5^{\varepsilon}} - 1}{\alpha + L_5^{\varepsilon}} k L_2^{\varepsilon} C_2 > C_2. \tag{4.33}$$

Step 4. When $t \geq \tau + \eta_{\min}$, we have from (4.13) and (4.14) that there exist $\eta_3 < \eta_4$ and $\bar{\eta}_3 < \bar{\eta}_4$ such that

$$y^{\phi}(\eta_3) = \left(\frac{\varepsilon^2}{1+\varepsilon} \right)^{\varepsilon}, \; y^{\phi}(\eta_4) = \left(\frac{1}{\varepsilon} \right)^{2\varepsilon}, \; \eta_4 - \eta_3 = O(-\varepsilon \log \varepsilon)$$

and

$$y^{\bar{\phi}}(\bar{\eta}_3) = \left(\frac{\varepsilon^2}{1+\varepsilon} \right)^{\varepsilon}, \; y^{\bar{\phi}}(\bar{\eta}_4) = \left(\frac{1}{\varepsilon} \right)^{2\varepsilon}, \; \bar{\eta}_4 - \bar{\eta}_3 = O(-\varepsilon \log \varepsilon).$$

Let

$$\eta_{\min}^3 = \min\{\eta_3, \bar{\eta}_3\}, \; \eta_{\max}^4 = \max\{\eta_4, \bar{\eta}_4\}.$$

Then by (4.13) and (4.14) we have

$$\eta_{\min}^3 = t_2 + O(-\varepsilon \log \varepsilon), \; \eta_{\max}^4 = t_2 + O(-\varepsilon \log \varepsilon), \; \eta_{\max}^4 - \eta_{\min}^3 = O(-\varepsilon \log \varepsilon). \tag{4.34}$$

Since $t_2 > t_1 + \tau$, we can choose $\sigma_7 > 0$ sufficiently small so that for each $\varepsilon \in (0, \sigma_7)$ the inequality

$$\tau + \eta_{\max} < \eta_{\min}^3$$

holds. For $t \in [\tau + \eta_{\min}, \eta^3_{\min}]$, we similarly have

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \le \int_{\tau+\eta_{\min}}^{t} \left((\alpha + L_5^{\varepsilon})|y^{\phi}(s) - y^{\bar{\phi}}(s)| + kL_3^{\varepsilon} C_3 ||\phi - \bar{\phi}|| \right) ds + C_3 ||\phi - \bar{\phi}||$$

and

$$|y^{\phi} - y^{\bar{\phi}}| \le C_4 ||\phi - \bar{\phi}||, \tag{4.35}$$

where

$$C_4 = C_3 e^{(\alpha + L_5^{\varepsilon})(\eta^3_{\min} - \tau - \eta_{\min})} + \frac{e^{(\alpha + L_5^{\varepsilon})(\eta^3_{\min} - \tau - \eta_{\min})} - 1}{\alpha + L_5^{\varepsilon}} kL_3^{\varepsilon} C_3 > C_3. \tag{4.36}$$

Step 5. For $t \in [\eta^3_{\min}, \eta^4_{\max}]$, from (4.22) and (4.34) it is easy to demonstrate that $\eta_{\max} \le t - \tau \le \eta^3_{\min}$. Thus we have

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \le \int_{\eta^3_{\min}}^{t} \left((\delta + L_3^{\varepsilon})|y^{\phi}(s) - y^{\bar{\phi}}(s)| + kL_4^{\varepsilon} C_4 ||\phi - \bar{\phi}|| \right) ds + C_4 ||\phi - \bar{\phi}||.$$

Then it follows that

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \le C_5 ||\phi - \bar{\phi}||, \tag{4.37}$$

where

$$C_5 = C_4 \left(e^{(\delta + L_3^{\varepsilon})(\eta^4_{\max} - \eta^3_{\min})} + \frac{e^{(\delta + L_3^{\varepsilon})(\eta^4_{\max} - \eta^3_{\min})} - 1}{\delta + L_3^{\varepsilon}} kL_4^{\varepsilon} \right). \tag{4.38}$$

Step 6. For $t \in [\eta^4_{\max}, \tau + \eta^4_{\max}]$, we claim that $y^{\phi}(t) \ge (1/\varepsilon)^{2\varepsilon}$ and $y^{\bar{\phi}}(t) \ge (1/\varepsilon)^{2\varepsilon}$. We prove this claim only for the function y^{ϕ}, because the proof for the function $y^{\bar{\phi}}$ is similar and hence omitted. Note that $t_3 > \eta^4_{\max} = t_2 + O(-\varepsilon \log \varepsilon)$ for each $\varepsilon \in (0, \sigma_8)$ where σ_8 is chosen so that $t_3 > t_2 + O(-\sigma_8 \log \sigma_8)$. Using $y^{\phi}(t) = x(t) + O(-\varepsilon \log \varepsilon)$, with $y^{\phi}(t - \tau) = x(t - \tau) + O(-\varepsilon \log \varepsilon) \ge e^{-\alpha\tau} + O(-\varepsilon \log \varepsilon)$, and (3.7) and (4.34), we have from (4.1) that $dy^{\phi}(t)/dt > 0$ for $t \in [\eta^4, t_3]$, and thus y^{ϕ} is increasing and satisfies $y^{\phi}(\eta^4_{\max}) \ge y^{\phi}(\eta_4) \ge (1/\varepsilon)^{2\varepsilon}$. For $t \in [t_3, \tau + \eta^4_{\max}]$, $x(t) \ge 1 + \eta$. Then using Lemma 4.2 again we have

$$y^{\phi}(t) = x(t) + O(-\varepsilon \log \varepsilon) > \left(\frac{1}{\varepsilon} \right)^{2\varepsilon}$$

provided $\varepsilon \in (0, \sigma_9)$, where σ_9 is sufficiently small so that the above formula holds for $\varepsilon \in (0, \sigma_9)$. Therefore, we obtain

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \le \int_{\eta^4_{\max}}^{t} \left((\delta + L_2^{\varepsilon})|y^{\phi}(s) - y^{\bar{\phi}}(s)| + kL_3^{\varepsilon} C_5 ||\phi - \bar{\phi}|| \right) ds + C_5 ||\phi - \bar{\phi}||$$

and

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \le C_6 ||\phi - \bar{\phi}||, \tag{4.39}$$

where

$$C_6 = C_5 \left(e^{(\delta + L_2^{\varepsilon})\tau} + \frac{e^{(\delta + L_2^{\varepsilon})\tau} - 1}{\delta + L_2^{\varepsilon}} kL_3^{\varepsilon} \right). \tag{4.40}$$

Step 7. When $t \geq \tau + \eta_{\max}^4$, both y and $\bar{y}$ are decreasing and will take the value $1+\eta$ after a finite time. Suppose that s and $\bar{s}$ satisfy

$$y^{\phi}(s) = 1+\eta,\ y^{\bar{\phi}}(\bar{s}) = 1+\eta.$$

For the rest of the proof, we consider only the case $s < \bar{s}$, since the case when $s \geq \bar{s}$ can be similarly dealt with and the proof is omitted. By (4.4) and (4.34), we also obtain

$$s - (\tau + \eta_{\max}^4) = T_x - (\tau + t_2) + O(-\varepsilon \log \varepsilon)$$

and

$$\bar{s} - (\tau + \eta_{\max}^4) = T_x - (\tau + t_2) + O(-\varepsilon \log \varepsilon),$$

where T_x is the period of the function x. Because the distance between $\tau + \eta_{\max}^4$ and s may be greater than τ, we need to split the interval $[\tau + \eta_{\max}^4, s]$ into subintervals $[\tau + \eta_{\max}^4, 2\tau + \eta_{\max}^4]$, $[2\tau + \eta_{\max}^4, 3\tau + \eta_{\max}^4], \ldots, [m\tau + \eta_{\max}^4, s]$, where the length of each interval is exactly τ except the last one. Here m is the largest integer less than or equal to $(s - (\tau + \eta_{\max}^4))/\tau$. We can successively estimate $|y^{\phi} - y^{\bar{\phi}}|$ on the above subintervals to obtain

$$|y^{\phi}(t) - y^{\bar{\phi}}(t)| \leq C_7 ||\phi - \bar{\phi}||,\ t \in [\tau + \eta_{\max}^4, s], \tag{4.41}$$

with

$$C_7 = C_6 \left(e^{(\delta + L_2^{\varepsilon})\tau} + \frac{e^{(\delta + L_2^{\varepsilon})\tau} - 1}{\delta + L_2^{\varepsilon}} k L_2^{\varepsilon} \right)^{T_x}. \tag{4.42}$$

For $t \in [s, \bar{s}]$, the function $y^{\bar{\phi}}$ satisfies

$$y^{\bar{\phi}}(\bar{s}) = 1+\eta \text{ and } y^{\bar{\phi}}(t) = 1 + \eta + O(-\varepsilon \log \varepsilon),$$

because the length of the interval $[s, \bar{s}]$ is of order $O(-\varepsilon \log \varepsilon)$ and the derivative of $y^{\bar{\phi}}$ is bounded; c.f. Remark 4.3. On the other hand, since $s = T_x + O(-\varepsilon \log \varepsilon)$, $\bar{s} = T_x + O(-\varepsilon \log \varepsilon)$, $x(t) \geq 1 + \eta$ for $t \in [t_3, T_x]$, and $y^{\bar{\phi}}(t) = x(t) + O(-\varepsilon \log \varepsilon)$ for $t \in [0, T_x]$, we know by (4.20) that for $t \in [s, \bar{s}]$,

$$y^{\bar{\phi}}(t - \tau) \geq \left(\frac{1}{\varepsilon} \right)^{2\varepsilon}$$

and

$$k\beta_{\varepsilon}(y^{\bar{\phi}}(t-\tau)) y^{\bar{\phi}}(t-\tau) = O(-\varepsilon \log \varepsilon).$$

Therefore, from (4.1) we know that for $t \in [s, \bar{s}]$ the function $y^{\bar{\phi}}$ is decreasing and

$$\begin{aligned} \left| \frac{dy^{\bar{\phi}}(t)}{dt} \right| &= \left| -(\delta + \beta_{\varepsilon}(y^{\bar{\phi}}(t))) y^{\bar{\phi}}(t) + k\beta_{\varepsilon}(y^{\bar{\phi}}(t-\tau)) y^{\bar{\phi}}(t-\tau) \right| \\ &\geq |-\delta(1+\eta) + O(-\varepsilon \log \varepsilon)| \\ &\geq \frac{\delta(1+\eta)}{2}. \end{aligned}$$

Here we have assumed that ε is in the interval $(0, \sigma_{10})$, where σ_{10} is chosen so that for each $\varepsilon \in (0, \sigma_{10})$, the inequality

$$-\delta(1+\eta) + O(-\varepsilon \log \varepsilon) < -\frac{\delta(1+\eta)}{2}$$

holds. Applying the mean value theorem to the function $y^{\bar{\phi}}$ yields the existence of $\rho \in [s, \bar{s}]$ such that

$$|y^{\bar{\phi}}(\bar{s}) - y^{\bar{\phi}}(s)| = |(y^{\bar{\phi}})'(\rho)(\bar{s} - s)| \geq \frac{\delta(1+\eta)}{2} |\bar{s} - s|$$

or, by (4.41),

$$\begin{aligned} |\bar{s} - s| &\leq \frac{2}{\delta(1+\eta)} |y^{\bar{\phi}}(\bar{s}) - y^{\bar{\phi}}(s)| = \frac{2}{\delta(1+\eta)} |y^{\phi}(s) - y^{\bar{\phi}}(s)|, \\ &\leq \frac{2C_7}{\delta(1+\eta)} ||\phi - \bar{\phi}||. \end{aligned} \tag{4.43}$$

Our ultimate goal is to derive an estimate of $|y^{\bar{\phi}}_{\bar{s}}(\theta) - y^{\phi}_{s}(\theta)|$ where $\theta \in [-\tau, 0]$. Indeed, we have

$$|y^{\bar{\phi}}_{\bar{s}}(\theta) - y^{\phi}_{s}(\theta)| \leq |y^{\bar{\phi}}_{\bar{s}}(\theta) - y^{\bar{\phi}}_{s}(\theta)| + |y^{\bar{\phi}}_{s}(\theta) - y^{\phi}_{s}(\theta)|. \tag{4.44}$$

The first term of the right-hand side is bounded by

$$\int_{s+\theta}^{\bar{s}+\theta} \frac{dy^{\bar{\phi}}(t)}{dt} dt \leq M_2 |\bar{s} - s|,$$

where M_2 is the maximum value of the derivative of the function $y^{\bar{\phi}}$; c.f. Remark 4.3. The second term of (4.44) is bounded by $C_7 ||\phi - \bar{\phi}||$. Thus from (4.44), we have

$$|y^{\bar{\phi}}_{\bar{s}}(\theta) - y^{\phi}_{s}(\theta)| \leq C_7 \left(1 + \frac{2M_2}{\delta(1+\eta)}\right) ||\phi - \bar{\phi}||. \tag{4.45}$$

Using (4.21), we conclude from (4.29), (4.31), (4.33), (4.36), (4.38), (4.40), and (4.42) that

$$\lim_{\varepsilon \to 0} L^{\varepsilon}_{R} = \lim_{\varepsilon \to 0} C_7 \left(1 + \frac{2M_2}{\delta(1+\eta)}\right) = 0 < 1.$$

Therefore we conclude that there exists $\varepsilon_2 < \min\{\varepsilon_1, \sigma_6, \sigma_7, \sigma_8, \sigma_9, \sigma_{10}\}$ so that for each $\varepsilon \in (0, \varepsilon_2)$, the Lipschitz constant L^{ε}_{R} of the map R is less than 1. This completes our proof. □

For $L^{\varepsilon}_{R} < 1$, the return map R is contractive and there exists a unique fixed point ϕ in A_η. Thus we have demonstrated the existence of a unique slowly oscillating periodic solution for (4.1). The stability and exponential attractivity of this unique periodic orbit can be established using the standard techniques developed in [31, 32, 33, 34, 36].

5. Asymptotic expansions for the periodic solution. In the previous section we used fixed point theory to prove that there exists a unique periodic orbit for (4.1). We now carry out a quantitative analysis of this periodic solution as $\varepsilon < \varepsilon_2$.

Since the map R is contractive and the Lipschitz constant L_R^ε is exponentially decaying as $\varepsilon \to 0$, we are able to give an asymptotic expansion for this particular solution for $t \in [-\tau, 0]$ with error bound beyond all integer orders of ε.

If we take the initial function given by $\phi = 1 + \eta$, then we have a solution $y^{1+\eta}(\cdot)$ which is not periodic. But by Lemma 4.2, we have $y^{1+\eta}(t) = x(t) + O(-\varepsilon \log \varepsilon)$ for $t \in [0, T_x]$, and a $T_{1+\eta} > 0$ such that

$$y^{1+\eta}_{T_{1+\eta}}(0) = 1 + \eta, \quad y^{1+\eta}_{T_{1+\eta}}(\theta) > 1 + \eta, \qquad \theta \in [-\tau, 0).$$

It is obvious that $y^{1+\eta}_{T_{1+\eta}}(\theta) \in A_\eta$.

Assume that y is the periodic solution to (4.1) and satisfies $y(\theta) \in A_\eta$ for $\theta \in [-\tau, 0]$. Suppose also that $y(\theta)$ has the following asymptotic expansion:

$$y(\theta) = \sum_{i=0}^{\infty} \phi_i(\theta), \ \theta \in [-\tau, 0]. \tag{5.1}$$

The function ϕ_0 is given by $y^{1+\eta}_{T_{1+\eta}}$, and $\phi_i, i \geq 1$, with the norm $||\phi|| = \max_{-\tau \leq \theta \leq 0} |\phi(\theta)|$, will be constructed below. Let $y^{\phi_0}_{T_0}$ denote the image of the return map R at ϕ_0, i.e.,

$$y^{\phi_0}_{T_0}(\theta) = R(\phi_0) = F_\varepsilon(T_0, \phi_0), \ \theta \in [-\tau, 0],$$

where $T_0 > 0$ satisfies

$$y^{\phi_0}_{T_0}(0) = 1 + \eta, \ y^{\phi_0}_{T_0}(\theta) > 1 + \eta, \ \theta \in [-\tau, 0).$$

Similarly, by induction, we set

$$\phi_1 = R(\phi_0) - \phi_0 = y^{\phi_0}_{T_0} - \phi_0,$$

$$y^{\phi_1}_{T_1}(\theta) = R(\phi_1) = F_\varepsilon(T_1, \phi_1),$$

$$\phi_n(\theta) = R^n(\phi_0) - R^{n-1}(\phi_0) \text{ for } n \geq 2,$$

$$y^{\phi_n}_{T_n}(\theta) = R(\phi_n) = F_\varepsilon(T_n, \phi_n) \text{ for } n \geq 2,$$

where T_n satisfies

$$y^{\phi_n}_{T_n}(0) = 1 + \eta, \ y^{\phi_n}_{T_n}(\theta) > 1 + \eta, \ \theta \in [-\tau, 0), \ n \geq 1.$$

Thus we have

$$\begin{aligned} |\phi_n(\theta) - \phi_{n-1}(\theta)| &\leq L_R^\varepsilon |\phi_{n-1}(\theta) - \phi_{n-2}(\theta)| \\ &\leq (L_R^\varepsilon)^{n-1} |\phi_1(\theta) - \phi_0(\theta)|. \end{aligned}$$

Therefore, $y(\theta) = \sum_{i=0}^{\infty} \phi_i(\theta)$ is uniformly convergent for $\theta \in [-\tau, 0]$ and it is the fixed point of R.

We now give an asymptotic expansion for the period of the periodic solution y. Using (4.43), we have

$$|T_i - T_{i-1}| \le \frac{2C_7}{\delta(1+\eta)}||\phi_i - \phi_{i-1}||,$$

which means that the series

$$T_0 + \sum_{j=1}^{\infty}(T_j - T_{j-1})$$

is absolutely convergent to some constant, say, T_ε. Since L_R is exponentially decaying as $\varepsilon \to 0$, it is easy to see that the value of T_ε is dominated by T_0 in the sense that $T_\varepsilon - T_0$ is exponentially small as $\varepsilon \to 0$. Likewise the value of $y(\theta)$ in (5.1) is dominated by ϕ_0 with an exponential error bound as $\varepsilon \to 0$. Thus when $t \in [0, T_0]$, we know that the periodic solution $y(t)$ is also dominated by $y^{\phi_0}(t)$. Therefore the estimate of $y^{\phi_0}(t)$ and T_0 becomes significant. From Lemma 4.2 we have the following rough result for $y^{\phi_0}(t)$ and T_0:

$$y^{\phi_0}(t) = x(t) + O(-\varepsilon \log \varepsilon),\ T_0 = T_x + O(-\varepsilon \log \varepsilon).$$

We now give refined estimates for $y^{\phi_0}(t)$ and T_0 using the above information. As in the proof of Lemma 4.2, we split the interval $[0, T_0]$ into subintervals and estimate $y^{\phi_0}(t)$ on each subinterval successively. We demonstrate this process on the first subinterval for the purpose of illustration. Remember that the initial data are taken to be ϕ_0 which is greater than $1+\eta$ when t lies in the interval $[-\tau, 0)$. Let $t_1^{\phi_0}, \eta_1$, and η_2 be the values as defined in the proof of Lemma 4.2. Thus $t_1^{\phi_0}$ satisfy $y^{\phi_0}(t_1^{\phi_0}) = 1$. Integrating (4.1) from 0 to t, $t \in [0, t_1^{\phi_0}]$, gives

$$\begin{aligned} y^{\phi_0}(t) - y^{\phi_0}(0) = &-\delta \int_0^t y^{\phi_0}(s)ds - \int_0^t \beta_\varepsilon(y^{\phi_0}(s))y^{\phi_0}(s)ds \\ &+ k\int_0^t \beta_\varepsilon(y^{\phi_0}(s-\tau))y^{\phi_0}(s-\tau)ds. \end{aligned} \tag{5.2}$$

Since $t_1^{\phi_0} = t_1 + O(-\varepsilon \log \varepsilon)$ and $t_1 < \tau$, it is easy to see that the last term of the right-hand side of (5.2) is small and of $O(\varepsilon)$. Next we claim that

$$\int_0^t \beta_\varepsilon(y^{\phi_0}(s))y^{\phi_0}(s)ds = O(\varepsilon), \quad t \in [0, t_1^{\phi_0}]. \tag{5.3}$$

Indeed when $t \in [0, t_1^{\phi_0}]$, we have $k\beta y^{\phi_0}(t-\tau)y^{\phi_0}(t-\tau) = O(\varepsilon)$. Then from (4.1) we have

$$-\alpha(1+\eta) \le \frac{dy^{\phi_0}(t)}{dt} = -[\beta_\varepsilon(y^{\phi_0}(t)) + \delta]y^{\phi_0}(t) + O(\varepsilon) \le -\delta + O(\varepsilon). \tag{5.4}$$

Thus from (5.4) and the fact that

$$\begin{aligned} \left|\int_0^t \beta_\varepsilon(y^{\phi_0}(s))y^{\phi_0}(s)\frac{dy^{\phi_0}}{ds}ds\right| &\le \left|\int_0^{t_1^{\phi_0}} \beta_\varepsilon(y^{\phi_0}(s))y^{\phi_0}(s)\frac{dy^{\phi_0}}{ds}ds\right| \\ &= \left|\int_{1+\eta}^1 \frac{\beta_0 u}{1+u^{1/\varepsilon}}du\right| \\ &= O(\varepsilon), \end{aligned}$$

we know that $\int_0^t \beta_\varepsilon(y^{\phi_0}(s))y^{\phi_0}(s)ds$ is also of $O(\varepsilon)$ and the claim (5.3) is true. It follows then from (5.2) that for $t \in [0, t_1^{\phi_0}]$,

$$y^{\phi_0}(t) \ = -\delta \int_0^t y^{\phi_0}(t)dt + 1 + \eta + O(\varepsilon).$$

Using Gronwall's inequality, we obtain

$$y^{\phi_0}(t) = (1 + \eta + O(\varepsilon))e^{-\delta t},$$

which implies

$$y^{\phi_0}(t) = x(t) + O(\varepsilon),\ t \in [0, t_1^{\phi_0}]. \tag{5.5}$$

Continuing the above process, we can prove that (5.5) holds in the entire interval $[0, T_0]$. Furthermore, we also have

$$T_0 = T_x + O\left(\varepsilon\right),$$

which completes our refined estimate.

REFERENCES

[1] M. Adimy, F. Crauste, and S. Ruan, *A mathematical study of the hematopoiesis process with applications to chronic myelogenous leukemia*, SIAM J. Appl. Math., 65 (2005), pp. 1328–1352.

[2] M. Adimy and L. Pujo-Menjouet, *A mathematical model describing cellular division with a proliferating phase duration depending on the maturity of cells*, Electron J. Differential Equations, 107 (2003), pp. 1–14.

[3] J. Bélair, M. C. Mackey, and J. M. Mahaffy, *Age-structured and two-delay models for erythropoiesis*, Math. Biosci., 128 (1995), pp. 317–346.

[4] S. Bernard, J. Bélair, and M. C. Mackey, *Sufficient conditions for stability of linear differential equations with distributed delay*, Discrete Contin. Dyn. Syst. Ser. B, 1 (2001), pp. 233–256.

[5] A.-M. Buckle, R. Mottram, A. Pierce, G. S. Lucas, N. Russel, J. A. Miyan, and A. D. Whetton, *The effects of bcr-abl protein tyrosine kinase on maturation and proliferation of primitive haemotopoietic cells*, Molecular Medicine, 6 (2000), pp. 892–902.

[6] F. J. Burns and I. F. Tannock, *On the existence of a G_0 phase in the cell cycle*, Cell Tissue Kinet., 3 (1970), pp. 321–334.

[7] C. J. Eaves and A. C. Eaves, *Stem cell kinetics*, Baillieres Clinical Haematology, 10 (1997), pp. 233–257.

[8] P. Fortin and M. C. Mackey, *Periodic chronic myelogenous leukemia: Spectral analysis of blood cell counts and etiological implications*, Brit. J. Haematol., 104 (1999), pp. 336–345.

[9] T. Haferlach, M. Winkemann, C. Nickenig, M. Meeder, L. Rammpetersen, R. Schoch, M. Nickelsen, K. Webermatthiesen, B. Schlegelberger, C. Schoch, W. Gassman, and H. Loffler, *Which compartments are involved in Philadelphia-chromosome positive chronic myeloid leukaemia? An answer at the single cell level by combining may-grunwald-giemsa staining and fluorescence in situ hybridization techniques*, Brit. J. Haematol., 97 (1997), pp. 99–106.

[10] C. Haurie, D. C. Dale, and M. C. Mackey, *Cyclical neutropenia and other periodic hematological diseases: A review of mechanisms and mathematical models*, Blood, 92 (1998), pp. 2629–2640.

[11] T. Hearn, C. Haurie, and M. C. Mackey, *Cyclical neutropenia and the peripheral control of white blood cell production*, J. Theor. Biol., 192 (1998), pp. 167–181.

[12] X. Jiang, C. J. Eaves, and A. C. Eaves, *IL-3 and G-CSF gene expression in primitive PH+CD34(+) cells from patients with chronic myeloid leukemia (CML)*, Blood, 90 (1997), pp. 1745–1745.

[13] L. Kold-Andersen and M. C. Mackey, *Resonance in periodic chemotheray: A case study of acute myelogenous leukemia*, J. Theor. Biol., 209 (2001), pp. 113–130.

[14] L. G. Lajtha, *On DNA Labeling in the Study of the Dynamics of Bone Marrow Cell Populations*, Grune & Stratton, New York, 1959, pp. 173–182.
[15] M. C. Mackey, *Unified hypothesis of the origin of aplastic anemia and periodic hematopoiesis*, Blood, 51 (1978), pp. 941–956.
[16] M. C. Mackey, *Dynamic haematological disorders of stem cell origin*, in Biophysical and Biochemical Information Transfer in Recognition, J. G. Vassileva-Popova and E. V. Jensen, eds., Plenum Publishing, New York, 1979, pp. 373–409.
[17] M. C. Mackey, *Periodic auto-immune hemolytic anemia: An induced dynamical disease*, Bull. Math. Biol., 41 (1979), pp. 829–834.
[18] M. C. Mackey, *Mathematical models of hematopoietic cell replication and control*, in The Art of Mathematical Modeling: Case Studies in Ecology, Physiology & Biofluids, H. G. Othmer, F. R. Adler, M. A. Lewis, and J. C. Dalton, eds., Prentice–Hall, Upper Saddle River, NJ, 1997, pp. 149–178.
[19] M. C. Mackey, *Cell kinetic status of hematopoietic stem cells*, Cell Prolif., 34 (2001), pp. 71–83.
[20] M. C. Mackey and P. Dörmer, *Continuous maturation of proliferating erythroid precursors*, Cell Tissue Kinet., 15 (1982), pp. 381–392.
[21] M. C. Mackey and R. Rudnicki, *Global stability in a delayed partial differential equation describing cellular replication*, J. Math. Biol., 33 (1994), pp. 89–109.
[22] M. C. Mackey and R. Rudnicki, *A new criterion for the global stability of simultaneous cell replication and maturation processes*, J. Math. Biol., 38 (1999), pp. 195–219.
[23] C. Ou and J. Wu, *Periodic solutions of delay differential equations with a small parameter: Existence, stability and asymptotic expansion*, J. Dynam. Differential Equations, 3 (2004), pp. 605–628.
[24] L. Pujo-Menjouet, S. Bernard, and M. C. Mackey, *Long period oscillations in a G_0 model of hematopoietic stem cells*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 312–332.
[25] L. Pujo-Menjouet and M. C. Mackey, *Contribution to the study of periodic chronic myelogenous leukemia*, C.R. Biologies, 327 (2004), pp. 235–244.
[26] S. I. Rubinow and J. L. Lebowitz, *A mathematical model of neutrophil production and control in normal man*, J. Math. Biol., 1 (1975), pp. 187–225.
[27] M. Santillan, J. Bélair, J. M. Mahaffy, and M. C. Mackey, *Regulation of platelet production: The normal response to perturbation and cyclical platelet disease*, J. Theor. Biol., 206 (2000), pp. 585–903.
[28] J. A. Smith and L. Martin, *Do cells cycle?*, Proc. Natl. Acad. Sci. USA, 70 (1973), pp. 1263–1267.
[29] J. Swinburne and M. C. Mackey, *Cyclical thrombocytopenia: Characterization by spectral analysis and a review*, J. Theor. Med., 2 (2000), pp. 81–91.
[30] N. Takahashi, I. Miura, K. Saitoh, and A. B. Miura, *Lineage involvment of stem cells bearing the Philadelphia chromosome in chronic myeloid leukemia in the chronic phase as shown by a combination of fluorescence-activated cell sorting and fluorescence in situ hybridization*, Blood, 92 (1998), pp. 4758–4763.
[31] H.-O. Walther, *Contracting return maps for monotone delayed feedback*, Discrete Contin. Dynam. Systems, 7 (2001), pp. 259–274.
[32] H.-O. Walther, *Contracting return maps for some delay differential equations*, in Topics in Functional Differential and Difference Equations, Fields Inst. Commun. 29, T. Faria and P. Freitas, eds., AMS, Providence, RI, 2001, pp. 349–360.
[33] H.-O. Walther, *Stable periodic motion for a system with state dependent delay*, Differential Integral Equations, 15 (2002), pp. 923–944.
[34] H.-O. Walther, *Stable periodic motion of a system using echo for position control*, J. Dynam. Differential Equations, 15 (2003), pp. 143–223.
[35] G. F. Webb, *Theory of Nonlinear Age-Dependent Population Dynamics*, Monographs Textbooks Pure Appl. Math. 89, Dekker, New York, 1985.
[36] J. Wu, *Stable phase-locked periodic solutions in a delay differential system*, J. Differential Equations, 194 (2003), pp. 237–286.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 188–209

LONG-TIME EXISTENCE OF QUASILINEAR WAVE EQUATIONS EXTERIOR TO STAR-SHAPED OBSTACLES VIA ENERGY METHODS*

JASON METCALFE† AND CHRISTOPHER D. SOGGE‡

Abstract. We establish long-time existence results for quasilinear wave equations in the exterior of star-shaped obstacles. To do so, we prove an analogue of the mixed-norm estimates of Keel, Smith, and Sogge for the perturbed wave equation. The arguments that are presented rely only upon the invariance of the wave operator under translations and spatial rotations.

Key words. systems of nonlinear wave equations, exterior domains, almost global existence, star-shaped

AMS subject classification. 35L70

DOI. 10.1137/050627149

1. Introduction.

The purpose of this article is to establish long-time existence results for quasilinear wave equations in the exterior of a star-shaped obstacle. The proofs that are presented rely upon the classical invariance of the wave operator under translations and spatial rotations. These techniques use only energy methods, and thus we are optimistic about their potential use in other applications. A key step in completing the proof is to establish a weighted $L^2_t L^2_x$-estimate for the perturbed equation that is analogous to the one of Keel, Smith, and Sogge [8] for the free wave equation.

Let us more explicitly describe the initial value boundary value problem that we will study. We begin by fixing an obstacle $\mathcal{K} \subset \mathbb{R}^n$ that is compact, has smooth boundary, and is star-shaped with respect to the origin. The latter condition means that there is a smooth positive function ψ on S^{n-1} so that

$$\mathcal{K} = \{(r,\omega) \,:\, \psi(\omega) - r \geq 0\}.$$

Here, we have expanded x in polar coordinates as $x = r\omega$, $(r,\omega) \in [0,\infty) \times S^{n-1}$.

For such a fixed $\mathcal{K}$, we examine the quasilinear wave equation

$$\begin{cases} \Box u = Q(du, d^2u), \quad (t,x) \in \mathbb{R}_+ \times \mathbb{R}^n \backslash \mathcal{K}, \\ u(t,\,\cdot\,)|_{\partial\mathcal{K}} = 0, \\ u(0,\,\cdot\,) = f, \quad \partial_t u(0,\,\cdot\,) = g. \end{cases} \tag{1.1}$$

Here and throughout, $\Box = (\partial_t^2 - \Delta)$ denotes the standard d'Alembertian.

The nonlinearity $Q(du, d^2u)$ in (1.1) is quadratic in its arguments and is linear in d^2u. We can expand

$$Q(du, d^2u) = B(du) + \sum_{0 \leq \alpha,\beta,\gamma \leq n} B^{\alpha\beta}_{\gamma} \partial_\gamma u \partial_\alpha \partial_\beta u, \tag{1.2}$$

*Received by the editors March 18, 2005; accepted for publication (in revised form) November 29, 2005; published electronically April 12, 2006. The authors were supported in part by the NSF.
http://www.siam.org/journals/sima/38-1/62714.html

†Department of Mathematics, University of California Berkeley, Berkeley, CA 94720-3840 (metcalfe@math.berkeley.edu).

‡Department of Mathematics, Johns Hopkins University, Baltimore, MD 21218 (sogge@jhu.edu).

where $B(du)$ is a quadratic form and the $B_\gamma^{\alpha\beta}$ are real constants. We assume the symmetry condition

$$B_\gamma^{\alpha\beta} = B_\gamma^{\beta\alpha}. \tag{1.3}$$

By scaling, we note that it suffices to choose $\mathcal{K} \subset \{|x| < 1\}$, and we will make this assumption throughout.

In order to solve (1.1), the data must be assumed to satisfy the relevant compatibility conditions. Briefly, this means that if we set $J_k u = \{\partial_x^\alpha u : 0 \le |\alpha| \le k\}$ and if u is a formal H^m solution for some fixed m, then we can write $\partial_t^k u(0, \cdot) = \psi_k(J_k f, J_{k-1} g)$, $0 \le k \le m$, for compatibility functions ψ_k which depend on Q, $J_k f$, and $J_{k-1} g$. The compatibility condition for $(f, g) \in H^m \times H^{m-1}$ states that ψ_k must vanish on $\partial\mathcal{K}$ when $0 \le k \le m-1$. Additionally, $(f, g) \in C^\infty$ are said to satisfy the compatibility condition to infinite order if this condition holds for all m. For a more detailed exposition on compatibility conditions, see, e.g., [7].

In describing the main results, we will use the notation $\{\Omega\} = \{x_i\partial_j - x_j\partial_i : 1 \le i < j \le n\}$ to denote the generators of the spatial rotations. We will also use $\{Z\} = \{\partial_k, \Omega : 0 \le k \le n\}$ to denote the generators of translations and spatial rotations.

Our main results are as follows. The first states that small-data solutions to (1.1) exist almost globally if $n = 3$.

THEOREM 1.1. *Assume that the star-shaped obstacle $\mathcal{K} \subset \mathbb{R}^3$ and the nonlinearity $Q(du, d^2u)$ are as above. Suppose that the initial data $(f, g) \in C^\infty(\mathbb{R}^3\backslash\mathcal{K})$ satisfy the compatibility condition to infinite order. Then, there are constants $\kappa, \varepsilon_0 > 0$ and an integer $N > 0$ so that for all $\varepsilon < \varepsilon_0$ and data satisfying*

$$\sum_{|\mu|\le N} \|Z^\mu \nabla_x f\|_{L^2(\mathbb{R}^3\backslash\mathcal{K})} + \sum_{|\mu|\le N} \|Z^\mu g\|_{L^2(\mathbb{R}^3\backslash\mathcal{K})} \le \varepsilon, \tag{1.4}$$

(1.1) *has a unique solution $u \in C^\infty([0, T_\varepsilon] \times \mathbb{R}^3\backslash\mathcal{K})$ with*

$$T_\varepsilon = \exp(\kappa/\varepsilon). \tag{1.5}$$

This bound on the lifespan of solutions in $n = 3$ is sharp, as is illustrated by finite propagation speed and the counterexamples of John [5] and Sideris [21] in the boundaryless case.

The second main result states that small-data solutions exist globally in higher dimensions.

THEOREM 1.2. *Suppose $n \ge 4$. Assume that the star-shaped obstacle $\mathcal{K} \subset \mathbb{R}^n$ and the nonlinearity $Q(du, d^2u)$ are as above. Suppose that the initial data $(f, g) \in C^\infty(\mathbb{R}^n\backslash\mathcal{K})$ satisfy the compatibility condition to infinite order. Then, there are a constant $\varepsilon_0 > 0$ and an integer $N > 0$ so that for all $\varepsilon < \varepsilon_0$ and data satisfying*

$$\sum_{|\mu|\le N} \|Z^\mu \nabla_x f\|_{L^2(\mathbb{R}^n\backslash\mathcal{K})} + \sum_{|\mu|\le N} \|Z^\mu g\|_{L^2(\mathbb{R}^n\backslash\mathcal{K})} \le \varepsilon, \tag{1.6}$$

(1.1) *has a unique solution $u \in C^\infty([0, \infty) \times \mathbb{R}^n\backslash\mathcal{K})$.*

While we have stated only the theorems for scalar wave equations, as the proofs rely only upon energy methods, straightforward modifications would yield the results for multiple speed systems of wave equations. In order not to further complicate the

notation, we will prove only the scalar case. A more detailed exposition concerning the multiple speed case will be available in a forthcoming paper on null-form wave equations.

Theorem 1.1 was first proved by Keel, Smith, and Sogge [9]. It is an analogue of the results concerning boundaryless wave equations of John and Klainerman [6] and Klainerman and Sideris [11]. Theorem 1.2 was previously shown by the authors [17]. This generalizes the work on wave equations in higher dimensions previously completed by Metcalfe [13], Shibata and Tsutsumi [20], and Hayashi [3]. It is also worth pointing out the following works for related problems involving null-form nonlinearities: Keel, Smith, and Sogge [7]; Metcalfe and Sogge [16]; and Metcalfe, Nakamura, and Sogge [14, 15]. The techniques in this paper appear to allow for some simplifications of these proofs, and this will be explored in a subsequent paper. The arguments, however, are more involved as they require the use of the scaling vector field and decay estimates of Klainerman and Sideris [11].

The techniques used to prove Theorem 1.1 represent an improvement over those in [9] in a number of ways. Most importantly, the proofs in this article make no reference to the fundamental solution of the wave equation or to the sharp Huygens' principle. Thus, it is believed that these techniques will be more suitable for other applications. For example, one might compare the methods of Sideris and Tu [24] to those used in Sideris [22, 23]. Additionally, we are not required to use the scaling vector field $L = t\partial_t + r\partial_r$. On a lesser note, we remark that the proofs herein seem to require less regularity of the initial data and less regularity of the boundary of the obstacle. As neither proof takes care to minimize such regularity, there is much possibility for further improvement in this direction. The proof of Theorem 1.2 improves upon the techniques of previous works in similar ways.

It is interesting to note that our arguments never make explicit use of the well-known decay of local energy. See, e.g., Lax, Morawetz, and Phillips [12]. We do, however, rely upon a geometrical condition that is sufficient to ensure such estimates. This condition is used in ways that are reminiscent of those of Morawetz [18] in proving said decay estimates.

A key estimate which is common to many of the previous studies of wave equations in exterior domains was established by Keel, Smith, and Sogge [8] and states that for $n \geq 3$

$$(\log(2+T))^{-1/2}\Big(\|\langle x\rangle^{-1/2}\nabla_{t,x}\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n)} + \|\langle x\rangle^{-3/2}\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n)}\Big) \lesssim \|\nabla_{t,x}\phi(0,\,\cdot\,)\|_2 + \int_0^T \|\Box\phi(s,\,\cdot\,)\|_2\,ds. \tag{1.7}$$

The proof is easily modified to yield the second estimate

$$\|\langle x\rangle^{-1/2-}\nabla_{t,x}\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n)} + \|\langle x\rangle^{-3/2-}\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n)} \lesssim \|\nabla_{t,x}\phi(0,\,\cdot\,)\|_2 + \int_0^T \|\Box\phi(s,\,\cdot\,)\|_2\,ds. \tag{1.8}$$

Here, we are using the notation $\langle x\rangle = \langle r\rangle = (1+|x|^2)^{1/2}$. We are also using the notation $\langle x\rangle^{-1/2-}$ and $\langle x\rangle^{-3/2-}$ to indicate that (1.8) holds with the weights replaced, respectively, by $\langle x\rangle^{-1/2-\delta}$ and $\langle x\rangle^{-3/2-\delta}$ for any $\delta > 0$. Moreover, we are using $A \lesssim B$ to indicate $A \leq CB$ for some positive, unspecified constant C.

These estimates are related to an earlier one of Strauss [27] and were used by Keel, Smith, and Sogge [8] to give a proof of almost global existence to semilinear

wave equations in exterior domains. Using this estimate, long-time existence was established using the $O(1/|x|)$ decay of the wave equation rather than the more standard $O(1/t)$ decay which is much more difficult to prove when there is a boundary. Metcalfe [13] completed the analogous result for higher dimensions using an estimate of the form (1.8) and arguments that are reminiscent of [8].

It should be noted that estimates similar to (1.7) and (1.8) which hold in all dimensions have been shown by Hidano and Yokoyama [4]. The above estimates should also be compared to the Morawetz identities (see, e.g., [19])

$$\||x|^{-1/2}\not\nabla\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n)} + \||x|^{-3/2}\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n)} \lesssim \|\nabla_{t,x}\phi(0,\,\cdot\,)\|_2 + \int_0^T \|\Box\phi(s,\,\cdot\,)\|_2\,ds, \quad n \geq 4,$$

and

$$\||x|^{-1/2}\not\nabla\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^3)} + \|\phi(\,\cdot\,,0)\|_{L^2_t([0,T])} \lesssim \|\nabla_{t,x}\phi(0,\,\cdot\,)\|_2 + \int_0^T \|\Box\phi(s,\,\cdot\,)\|_2\,ds, \quad n = 3,$$

which correspond to choosing $f(r) \equiv 1$ in the proof of Lemma 4.1. Here $\not\nabla$ denotes the angular portion of ∇_x.

The estimates (1.7) and (1.8) are proved by scaling a version of (1.8) where the norms in the left side are taken over $[0,T] \times \{|x| < 1\}$. These local versions are established either, in odd dimensions, by noticing that the backward light cones $s + |x| \in (j-1, j]$, $j = 1, 2, 3, \ldots$, have finite overlap or by an argument using Plancherel's identity (see Smith and Sogge [25]). Then, using techniques that resemble those of [25], one can show that an estimate for the Dirichlet-wave equation follows from those for the free equation.

The estimates (1.7) and (1.8) are, however, insufficient to give a proof of long-time existence for quasilinear equations as there is a loss of regularity in the right side. In order to get around this, previous works have had to rely on pointwise estimates that involve direct estimation of the fundamental solution of the free wave equation.

Recently, Rodnianski [26, Appendix] has given a new proof of an estimate related to (1.8). This new proof relies only upon energy methods. A main topic of this paper is the further study of this argument. In particular, we show that Rodnianski's argument can be used to prove (1.7) and (1.8). Moreover, we show that this argument can be used to directly prove an estimate for the Dirichlet-wave equation if the obstacle is assumed to be star-shaped. Thus, we will not rely on the cutoff methods used previously. Last, this new geometric argument, unlike the previously established proofs, lends itself well to establishing similar weighted estimates for perturbed equations. With such estimates for the perturbed equation, one can prove Theorems 1.1 and 1.2 using the arguments of [8].

The mixed-norm estimates for the perturbed equation in Theorem 5.1 give a partial answer to questions raised in Alinhac [1] concerning the adaptability of the Keel–Smith–Sogge estimates to more general settings. During final preparations of this article, it was learned that Alinhac [2] had independently obtained a Keel–Smith–Sogge-type estimate for the perturbed wave equation using different (although related) techniques. This argument, however, requires assumptions on the perturbation that are less favorable in the current setting. In particular, it is required that the perturbation decay in t. When there is a boundary, such decay is quite difficult to prove, and

we are using the mixed-norm estimates in place of such decay. Thus, it is essential that we require the perturbation terms to have decay only in $|x| = r$.

Before proceeding, we fix some notation. Throughout the paper, we will use the Einstein convention where repeated indices are summed. We will use Greek indices $\alpha, \beta, \gamma, \delta$ when the indices are to run from $0, \dots, n$. We will use Latin indices a, b when the implicit summations run from $1, \dots, n$. We will let $g_{\alpha\beta} = \mathrm{diag}(-1, 1, \dots, 1)$ be the Minkowski metric, and $\langle \cdot, \cdot \rangle$ will occasionally be used to denote the Euclidean inner product on $\mathbb{R}^n$. Unless explicitly stated to the contrary, L^2 norms are taken over $\mathbb{R}^n \backslash \mathcal{K}$. We will let $S_T = [0, T] \times \mathbb{R}^n \backslash \mathcal{K}$ denote a time strip of height T. We will use the notation $t = x_0$, $\partial_t = \partial_0$ interchangeably. And, when convenient, we will use $' = \partial = \nabla_{t,x} = (\partial_t, \nabla_x)$ to denote the full space-time gradient. We will use D to denote the Levi–Civita connection of $g_{\alpha\beta}$, but as this metric is flat, we have the correspondence $D^\alpha = \partial^\alpha$.

This paper is organized as follows. In the next section, we will give the weighted Sobolev inequality from which we easily obtain the required $O(1/|x|^{(n-1)/2})$ decay for solutions to the wave equation. In the third section, we prove the basic energy estimates that will be used in the proofs of long-time existence. In the fourth section, we give the new geometrical proof of the mixed-norm estimates of Keel, Smith, and Sogge. This argument follows that of Rodnianski [26, Appendix] quite closely. In the following section, we show that the energy methods used to prove the mixed-norm estimates are stable under small perturbations. In the final two sections, we give the proofs of Theorems 1.1 and 1.2, respectively.

2. Sobolev estimates. In this section, we give the now standard weighted Sobolev estimate from which one can obtain the necessary $O(1/|x|^{(n-1)/2})$ decay in order to show our long-time existence results. See [10].

LEMMA 2.1. *Suppose that $h \in C^\infty(\mathbb{R}^n)$. Then, for $R \geq 1$,*

$$\|h\|_{L^\infty(R/2<|x|<R)} \lesssim R^{-(n-1)/2} \sum_{|\mu|+|\nu| \leq \frac{n+2}{2}} \|\Omega^\mu \partial_x^\nu h\|_{L^2(R/4<|x|<2R)}. \tag{2.1}$$

3. Energy estimates. In this section, we will collect the energy estimates that we will require. These results are rather standard. We will be concerned with solutions $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^n \backslash \mathcal{K})$ of the Dirichlet-wave equation

$$\begin{cases} \Box_h \phi = F, \\ \phi|_{\partial\mathcal{K}} = 0, \end{cases} \tag{3.1}$$

where

$$\Box_h \phi = -\partial^\alpha \partial_\alpha \phi + h^{\alpha\beta} \partial_\alpha \partial_\beta \phi = (\partial_t^2 - \Delta)\phi + \sum_{\alpha,\beta=0}^n h^{\alpha\beta}(t,x) \partial_\alpha \partial_\beta \phi.$$

We shall assume that the $h^{\alpha\beta}$ satisfy the symmetry conditions

$$h^{\alpha\beta} = h^{\beta\alpha} \tag{3.2}$$

as well as the size conditions

$$\sum_{\alpha,\beta=0}^n |h^{\alpha\beta}(t,x)| \leq \delta \ll 1. \tag{3.3}$$

We define the energy form associated with $\Box_h$

$$e_0 = e_0(\phi) = (\partial_0\phi)^2 + \frac{1}{2}\partial^\gamma\phi\partial_\gamma\phi - g_{0\gamma}h^{\gamma\delta}\partial_\delta\phi\partial_0\phi - \frac{1}{2}h^{\gamma\delta}\partial_\gamma\phi\partial_\delta\phi. \tag{3.4}$$

Our most basic estimate involves

$$E_M(t) = E_M(\phi)(t) = \int_{\mathbb{R}^n\setminus\mathcal{K}} \sum_{j=0}^{M} e_0(\partial_t^j\phi)(t,x)\,dx.$$

LEMMA 3.1. *Fix $M = 0, 1, 2, \ldots$ and assume that the perturbation terms $h^{\alpha\beta}$ are as above. Suppose also that $\phi \in C^\infty$ solves* (3.1) *and that for every t, $\phi(t,x) = 0$ for large $|x|$. Then,*

$$E_M(T) \lesssim E_M(0) + \sum_{j,k=0}^{M}\int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}} |(\partial_0\partial_t^k\phi)(\Box_h\partial_t^j\phi)|\,dx\,dt \tag{3.5}$$

$$+ \sum_{j,k=0}^{M}\int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}} \Big(|(\partial_\alpha h^{\alpha\beta})(\partial_0\partial_t^j\phi)(\partial_\beta\partial_t^k\phi)| + |(\partial_0 h^{\alpha\beta})(\partial_\alpha\partial_t^j\phi)(\partial_\beta\partial_t^k\phi)|\Big)\,dx\,dt.$$

We first note that since $\partial_t^j\phi$ satisfies the Dirichlet boundary conditions for $1 \le j \le M$ it suffices to prove the result for $M = 0$. To proceed with the proof, we must define the other components of the energy-momentum vector. For $k = 1, \ldots, n$, we set

$$e_k = e_k(\phi) = \partial_k\phi\partial_0\phi - g_{k\gamma}h^{\gamma\delta}\partial_\delta\phi\partial_0\phi. \tag{3.6}$$

Calculating the divergence of this energy-momentum vector, we see that

$$D^\alpha e_\alpha = -\partial_0\phi\Box_h\phi - (\partial_\alpha h^{\alpha\delta})\partial_\delta\phi\partial_0\phi + \frac{1}{2}(\partial_0 h^{\gamma\delta})\partial_\gamma\phi\partial_\delta\phi. \tag{3.7}$$

If we integrate in the spatial components and apply the divergence theorem, it follows that

$$\partial_t\int_{\mathbb{R}^n\setminus\mathcal{K}} e_0\,dx + \int_{\partial\mathcal{K}} e_a n^a\,d\sigma = \int_{\mathbb{R}^n\setminus\mathcal{K}} \partial_0\phi\Box_h\phi\,dx \tag{3.8}$$

$$+ \int_{\mathbb{R}^n\setminus\mathcal{K}} \left((\partial_\alpha h^{\alpha\delta})\partial_\delta\phi\partial_0\phi - \frac{1}{2}(\partial_0 h^{\gamma\delta})\partial_\gamma\phi\partial_\delta\phi\right)\,dx,$$

where $\vec{n}$ is the outward unit normal to $\mathcal{K}$ and $d\sigma$ is the surface measure on $\partial\mathcal{K}$.

Since ∂_t preserves the Dirichlet boundary condition, we have that $\partial_t\phi$ vanishes on $\partial\mathcal{K}$ and that the integrand of the second term in the left side of (3.8) vanishes. If we integrate the remaining terms over a time interval $[0,T]$, (3.5) follows easily. □

Next, we will need energy estimates that also involve spatial derivatives.

LEMMA 3.2. *Suppose that the $h^{\alpha\beta}$ are as above with δ chosen sufficiently small. Then, if ϕ solves* (3.1) *and if $N = 0, 1, 2, \dots$ is fixed,*

$$\begin{aligned}\sum_{|\mu|\le N} \|\partial^\mu_{t,x}\phi'(T,\,\cdot\,)\|_2 &\lesssim \sum_{j\le N} \|\partial_t^j \phi'(0,\,\cdot\,)\|_2 \\ &+ \sum_{j,k=0}^N \left(\int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} |(\partial_0\partial_t^k\phi)(\Box_h\partial_t^j\phi)|\,dx\,dt\right)^{1/2} \\ &+ \sum_{j,k=0}^N \left[\int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} \Big(|(\partial_\alpha h^{\alpha\beta})(\partial_0\partial_t^j\phi)(\partial_\beta\partial_t^k\phi)| + |(\partial_0 h^{\alpha\beta})(\partial_\alpha\partial_t^j\phi)(\partial_\beta\partial_t^k\phi)|\Big)\,dx\,dt\right]^{1/2} \\ &+ \sum_{|\mu|\le N-1} \|\Box\partial^\mu_{t,x}\phi(T,\,\cdot\,)\|_2. \end{aligned} \tag{3.9}$$

Since

$$\frac{1}{2}|\phi'(t,x)|^2 \le e_0(t,x) \le 2|\phi'(t,x)|^2$$

for δ in (3.3) sufficiently small, this follows from (3.9) and a standard elliptic regularity argument. The interested reader can see, e.g., Lemma 2.3 of [16] or Theorem 5.2 of [9]. □

Finally, we will need energy estimates that involve the generators of spatial rotations as well as derivatives.

LEMMA 3.3. *Fix $N = 0, 1, 2, \dots$ and set*

$$Y_N(t) = \sum_{|\mu|\le N} \int e_0(Z^\mu\phi)(t,x)\,dx.$$

Suppose that (3.3) *holds for δ sufficiently small. Then*

$$\begin{aligned}\partial_t Y_N(t) &\lesssim \sum_{|\mu|,|\nu|\le N} \int_{\mathbb{R}^n\setminus\mathcal{K}} |(\partial_0 Z^\mu\phi)(\Box_h Z^\nu\phi)|\,dx \\ &+ \sum_{|\mu|,|\nu|\le N} \int_{\mathbb{R}^n\setminus\mathcal{K}} \Big(|(\partial_\gamma h^{\gamma\delta})(\partial_0 Z^\mu\phi)(\partial_\delta Z^\nu\phi)| + |(\partial_0 h^{\gamma\delta})(\partial_\gamma Z^\mu\phi)(\partial_\delta Z^\nu\phi)|\Big)\,dx \\ &+ \sum_{|\mu|\le N+1} \|\partial^\mu u'(s,\,\cdot\,)\|^2_{L^2(|x|<1)}. \end{aligned} \tag{3.10}$$

In order to prove (3.10), we argue as in the proof of Lemma 3.1 and find that

$$\begin{aligned}\partial_t Y_N &\lesssim \sum_{|\mu|,|\nu|\le N} \int_{\mathbb{R}^n\setminus\mathcal{K}} |(\partial_0 Z^\mu\phi)(\Box_h Z^\nu\phi)|\,dx \\ &+ \sum_{|\mu|,|\nu|\le N} \int_{\mathbb{R}^n\setminus\mathcal{K}} \Big(|(\partial_\gamma h^{\gamma\delta})(\partial_0 Z^\mu\phi)(\partial_\delta Z^\nu\phi)| + |(\partial_0 h^{\gamma\delta})(\partial_\gamma Z^\mu\phi)(\partial_\delta Z^\nu\phi)|\Big)\,dx \\ &+ \int_{\partial\mathcal{K}} |e_k^N n^k|\,d\sigma, \end{aligned}$$

where $\vec{n}$ is the outward normal at a given point of $\partial\mathcal{K}$ and the $e_k^N = \sum_{|\mu|\le N} e_k^N (Z^\mu\phi)(t,x)$ are as in (3.6). Since $\mathcal{K} \subset \{|x|<1\}$ and since

$$\sum_{|\mu|\le N} |Z^\mu\phi(t,x)| \lesssim \sum_{|\mu|\le N} |\partial^\mu\phi(t,x)|, \quad x\in\partial\mathcal{K},$$

we have

$$\int_{\partial\mathcal{K}} |e_k^N n^k|\, d\sigma \lesssim \int_{\{x\in\mathbb{R}^n\setminus\mathcal{K}\,:\,|x|<1\}} \sum_{|\mu|\le N+1} |\partial^\mu\phi'(t,x)|^2\, dx,$$

which completes the proof. □

4. Geometric approach to $L_t^2L_x^2$-estimates. In this section, we show that the estimates (1.7) and (1.8) hold in the exterior of a star-shaped obstacle. In the next section, we will show that these estimates also hold for the perturbed wave equation. For clarity of exposition, we begin here with the proofs for the standard d'Alembertian. These estimates will be shown directly using energy methods and result from straightforward modifications of Rodnianski's argument [26, Appendix].

LEMMA 4.1. *Suppose that $\mathcal{K}$ is as above and $n\ge 3$. Suppose also that $\phi\in C^\infty$ satisfies $\phi|_{\partial\mathcal{K}}=0$ and that ϕ vanishes for large $|x|$ for every t. Then, we have*

$$(\log(2+T))^{-1/2}\Big(\|\langle x\rangle^{-1/2}\phi'\|_{L_t^2L_x^2(S_T)} + \|\langle x\rangle^{-3/2}\phi\|_{L_t^2L_x^2(S_T)}\Big) \lesssim \|\phi'(0,\cdot)\|_2 + \int_0^T \|\Box\phi(s,\cdot)\|_2\, ds \tag{4.1}$$

and

$$\|\langle x\rangle^{-1/2-}\phi'\|_{L_t^2L_x^2(S_T)} + \|\langle x\rangle^{-3/2-}\phi\|_{L_t^2L_x^2(S_T)} \lesssim \|\phi'(0,\cdot)\|_2 + \int_0^T \|\Box\phi(s,\cdot)\|_2\, ds \tag{4.2}$$

for any $T>0$. The implicit constants in (4.1) *and* (4.2) *are independent of $\mathcal{K}$.*

By Duhamel's principle, we shall need only the homogeneous case. So, let ϕ be a solution to

$$\begin{cases} \Box\phi = (\partial_t^2-\Delta)\phi = 0, \\ \phi|_{\partial\mathcal{K}} = 0. \end{cases} \tag{4.3}$$

We let $Q_{\alpha\beta}$ denote its energy-momentum tensor

$$Q_{\alpha\beta}[\phi] = \partial_\alpha\phi\partial_\beta\phi - \frac{1}{2}g_{\alpha\beta}\partial^\gamma\phi\partial_\gamma\phi. \tag{4.4}$$

It is well known that $Q_{\alpha\beta}$ is divergence free. That is,

$$D^\alpha Q_{\alpha\beta}[\phi] = 0.$$

In order to get the weighted estimates, we define the momentum density

$$P_\alpha[\phi, X] = Q_{\alpha\beta}[\phi]X^\beta \tag{4.5}$$

by contracting $Q_{\alpha\beta}[\phi]$ with the radial vector field

$$X = f(r)\partial_r \tag{4.6}$$

(and thus, $X^a = \frac{f(r)}{r}x^a$ for $a = 1, \dots, n$ and $X^0 = 0$). One can check that this satisfies

$$D^\alpha P_\alpha[\phi, X] = \frac{1}{2} Q_{\alpha\beta}[\phi]\pi^{\alpha\beta},$$

where

$$\pi_{\alpha\beta} = D_\alpha X_\beta + D_\beta X_\alpha$$

is the deformation tensor of X.

A direct calculation then yields that

$$D^\alpha P_\alpha[\phi, X] = f'(r)|\partial_r\phi|^2 + \frac{f(r)}{r}|\not\nabla\phi|^2 - \frac{1}{2}\mathrm{tr}\pi\partial^\gamma\phi\partial_\gamma\phi, \tag{4.7}$$

where, as you can check,

$$\mathrm{tr}\pi = f'(r) + (n-1)\frac{f(r)}{r}. \tag{4.8}$$

Here, $\not\nabla\phi$ denotes the angular portion of the spatial gradient $\nabla_x\phi$. At this point, we define the modified momentum density

$$\tilde{P}_\alpha[\phi, X] = P_\alpha[\phi, X] + \frac{1}{2}\mathrm{tr}\pi\phi\partial_\alpha\phi - \frac{1}{4}\partial_\alpha(\mathrm{tr}\pi)|\phi|^2, \tag{4.9}$$

which satisfies

$$D^\alpha \tilde{P}_\alpha[\phi, X] = f'(r)|\partial_r\phi|^2 + \frac{f(r)}{r}|\not\nabla\phi|^2 - \frac{1}{4}\Delta(\mathrm{tr}\pi)|\phi|^2. \tag{4.10}$$

If we integrate this identity over a time strip $[0,T]\times\mathbb{R}^n\backslash\mathcal{K}$ and apply the divergence theorem, we see that

$$\begin{aligned}\int_{\mathbb{R}^n\backslash\mathcal{K}} \tilde{P}_0[\phi, X](0)\,dx - \int_{\mathbb{R}^n\backslash\mathcal{K}} \tilde{P}_0[\phi, X](T)\,dx - \int_0^T\int_{\partial\mathcal{K}} \tilde{P}_a[\phi, X](t)n^a\,d\sigma\,dt \\ = \int_0^T\int_{\mathbb{R}^n\backslash\mathcal{K}} \left(f'(r)|\partial_r\phi|^2 + \frac{f(r)}{r}|\not\nabla\phi|^2 - \frac{1}{4}\Delta(\mathrm{tr}\pi)|\phi|^2\right)\,dx\,dt,\end{aligned} \tag{4.11}$$

where $\vec{n}$ is the outward unit normal to $\mathcal{K}$ and $d\sigma$ is the surface measure on $\partial\mathcal{K}$. Here

$$\int_{\mathbb{R}^n\backslash\mathcal{K}} \tilde{P}_0[\phi, X](0)\,dx = \int_{\mathbb{R}^n\backslash\mathcal{K}} \left(X^a\partial_t\phi(0)\partial_a\phi(0) + \frac{1}{2}\mathrm{tr}\pi\phi(0)\partial_t\phi(0)\right)\,dx.$$

There is an identical expression for the time T piece on the left side of (4.11), since $\mathrm{tr}\pi$ is independent of t. If one chooses $f(r)$ so that $|f(r)| \lesssim 1$ and $|f'(r)| \lesssim \frac{1}{r}$, it follows from (4.8) and the Schwarz inequality that

$$\begin{aligned}\left|\int_{\mathbb{R}^n\backslash\mathcal{K}} \tilde{P}_0[\phi, X](0)\,dx\right| &\lesssim \left(\|r^{-1}\phi(0)\|_2 + \|\nabla_{t,x}\phi(0)\|_2\right)\|\nabla_{t,x}\phi(0)\|_2 \\ &\lesssim \|\nabla_{t,x}\phi(0)\|_2^2.\end{aligned}$$

The last inequality above follows from the Hardy inequality. Analogous estimates hold for the $\tilde{P}_0[\phi, X](T)$ term. Thus, by conservation of energy, this term is also controlled by $\|\nabla_{t,x}\phi(0)\|_2^2$.

Since $\phi|_{\partial\mathcal{K}} = 0$ and since ∂_t preserves the Dirichlet boundary condition, for the remaining boundary term, we have

$$-\int_0^T \int_{\partial\mathcal{K}} \tilde{P}_a[\phi, X](t) \cdot n^a \, d\sigma \, dt = -\int_0^T \int_{\partial\mathcal{K}} \frac{f(r)}{r} \left(\partial_{\vec{n}}\phi \partial_\beta \phi x^\beta - \frac{1}{2}|\nabla\phi|^2 \langle x, \vec{n}\rangle \right) d\sigma \, dt.$$

Here, $\partial_{\vec{n}}\phi = \langle \vec{n}, \nabla_x \rangle \phi$ denotes differentiation with respect to the outward normal to $\mathcal{K}$. Since $\phi = 0$ on $\partial\mathcal{K}$, we have that $\partial_\beta \phi = \partial_{\vec{n}}\phi \, n_\beta$. And, thus, we see that

$$-\int_0^T \int_{\partial\mathcal{K}} \tilde{P}_a[\phi, X](t) \cdot n^a \, d\sigma \, dt = -\frac{1}{2}\int_0^T \int_{\partial\mathcal{K}} \frac{f(r)}{r} (\partial_{\vec{n}}\phi)^2 \langle x, \vec{n}\rangle \, d\sigma \, dt.$$

This term is then easily seen to be negative as $\langle x, \vec{n}\rangle > 0$ for star-shaped $\mathcal{K}$.

Combining (4.11) and these estimates for the boundary terms, we see that

$$\int_0^T \int_{\mathbb{R}^n \setminus \mathcal{K}} \left(f'(r)|\partial_r \phi|^2 + \frac{f(r)}{r}|\not\nabla \phi|^2 - \frac{1}{4}\Delta(\mathrm{tr}\pi)|\phi|^2 \right) dx \, dt \lesssim \|\nabla_{t,x}\phi(0)\|_2^2. \tag{4.12}$$

At this point, we choose the weight function

$$f(r) = \frac{r}{\rho + r} \tag{4.13}$$

for some $\rho > 0$. It is easy to check that

$$\Delta(\mathrm{tr}\pi) = -\frac{1}{r(\rho+r)^4}\Big((n-1)(n-3)r^2 + 2(n^2-2n-2)r\rho + (n+1)(n-1)\rho^2 \Big) < 0 \tag{4.14}$$

for $n \geq 3$. Indeed, each term above is nonpositive. This, therefore, gives the a priori estimate

$$\int_0^T \int_{\mathbb{R}^n \setminus \mathcal{K}} \left(\frac{\rho}{(\rho+r)^2}|\partial_r\phi|^2 + \frac{1}{r+\rho}|\not\nabla\phi|^2 + \frac{\rho}{(\rho+r)^4}|\phi|^2 \right) dx \, dt \lesssim \|\nabla_{t,x}\phi(0)\|_2^2. \tag{4.15}$$

The implicit constant is independent of ρ.

By choosing $\rho = 1$, this yields the estimate

$$\int_0^T \int_{|x|\leq 1} \Big(|\nabla_x \phi|^2 + |\phi|^2 \Big) \, dx \, dt \lesssim \|\nabla_{t,x}\phi(0)\|_2^2. \tag{4.16}$$

Similarly, if we choose $\rho = 2^k$ for an integer $k \geq 0$, we get

$$\int_0^T \int_{2^{k-1}\leq|x|\leq 2^k} \Big(\frac{|\nabla_x\phi|^2}{r} + \frac{|\phi|^2}{r^3} \Big) \, dx \, dt \lesssim \|\nabla_{t,x}\phi(0)\|_2^2. \tag{4.17}$$

If we combine (4.16) and (4.17) and sum over $k \geq 1$, we see immediately that

$$\|\langle r\rangle^{-1/2-}\nabla_x\phi\|_{L_t^2 L_x^2([0,T]\times\mathbb{R}^n\setminus\mathcal{K})} + \|\langle x\rangle^{-3/2-}\phi\|_{L_t^2 L_x^2([0,T]\times\mathbb{R}^n\setminus\mathcal{K})} \lesssim \|\nabla_{t,x}\phi(0)\|_2. \tag{4.18}$$

The same argument also yields

$$(\log(2+T))^{-1/2}\Big(\|\langle r\rangle^{-1/2}\nabla_x\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n\setminus\mathcal{K})} + \|\langle x\rangle^{-3/2}\phi\|_{L^2_tL^2_x([0,T]\times\mathbb{R}^n\setminus\mathcal{K})}\Big) \lesssim \|\nabla_{t,x}\phi(0)\|_2. \tag{4.19}$$

Indeed, (4.19) follows trivially from the energy inequality and a Hardy inequality if the norms on the left side are taken over $|x| \geq T$. Thus, we need only sum over the $O(\log(2+T))$ choices of k with $2^{k-1} \lesssim T$.

It remains to see that a similar bound holds for $\partial_t\phi$. To do so, we define another modified momentum density

$$\bar{P}_\alpha[\phi, X] = P_\alpha[\phi, X] + \frac{n-1}{2}\frac{f(r)}{r}\phi\partial_\alpha\phi - \frac{n-1}{4}\partial_\alpha\Big(\frac{f(r)}{r}\Big)|\phi|^2. \tag{4.20}$$

Calculating the divergence, we have

$$D^\alpha\bar{P}_\alpha[\phi, X] = f'(r)(\partial_r\phi)^2 + \frac{f(r)}{r}|\not\nabla\phi|^2 - \frac{1}{2}f'(r)\partial^\gamma\phi\partial_\gamma\phi - \frac{n-1}{4}\Delta\Big(\frac{f(r)}{r}\Big)|\phi|^2. \tag{4.21}$$

Thus, if we integrate both sides of (4.21) over a time strip, apply the divergence theorem, and use similar arguments for controlling the boundary terms, it follows that

$$\int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}} f'(r)(\partial_r\phi)^2 + \frac{f(r)}{r}|\not\nabla\phi|^2 - \frac{1}{2}f'(r)\partial^\gamma\phi\partial_\gamma\phi - \frac{n-1}{4}\Delta\left(\frac{f(r)}{r}\right)|\phi|^2 \lesssim \|\nabla_x\phi(0)\|_2^2. \tag{4.22}$$

For f as in (4.13), we have $f'(r) < f(r)/r$ and

$$\Delta\left(\frac{f(r)}{r}\right) = \frac{-(n-3)r - (n-1)\rho}{r(r+\rho)^3} < 0 \tag{4.23}$$

for $n \geq 3$. Since $\partial^\gamma\phi\partial_\gamma\phi = -(\partial_t\phi)^2 + (\partial_r\phi)^2 + |\not\nabla\phi|^2$, we see from (4.22) that

$$\int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}} \frac{\rho}{(r+\rho)^2}(\partial_t\phi)^2\,dx\,dt \lesssim \|\nabla_{t,x}\phi(0)\|_2^2.$$

By choosing $\rho = 1$ and $\rho = 2^k$ for $k \geq 1$, we get the estimates analogous to (4.16) and (4.17) for $\partial_t\phi$. Summing over k as above and combining this with (4.18) and (4.19) immediately yield (4.1) and (4.2). □

It is worth noting that the proof shows that there is additional decay for certain derivatives of the solution ϕ, namely, for $\not\nabla\phi$. Indeed, from (4.15), one can see that the $\log(2+T)$ factor in (4.1) is not necessary for the angular derivatives. This extra decay is something that has been exploited in other estimates and applications. See, e.g., Alinhac [1].

5. $L^2_tL^2_x$-estimate for the perturbed equation. The goal of this section is to show that the methods of the previous section can be adapted to give similar bounds for perturbed wave equations. This is the main new estimate of this article.

THEOREM 5.1. *Suppose $\mathcal{K}$ is as above and that $n \geq 3$. Let $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^n \backslash \mathcal{K})$ be a solution of (3.1). Suppose that $h^{\alpha\beta}$ satisfies (3.2) and (3.3) for a small choice of δ. Then, if $\Box_h \phi = F$, we have*

$$\begin{aligned}(5.1)\quad &\|\langle x\rangle^{-1/2-}\phi'\|_{L^2_t L^2_x(S_T)} + (\log(2+T))^{-1/2}\|\langle x\rangle^{-1/2}\phi'\|_{L^2_t L^2_x(S_T)} \\ &\quad + \|\langle x\rangle^{-3/2-}\phi\|_{L^2_t L^2_x(S_T)} + (\log(2+T))^{-1/2}\|\langle x\rangle^{-3/2}\phi\|_{L^2_t L^2_x(S_T)} \\ &\quad \lesssim \|\phi'(0,\cdot)\|_2 + \left(\int_0^T \int_{\mathbb{R}^n\backslash\mathcal{K}} \left(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\right)|F|\, dx\, dt\right)^{1/2} \\ &\quad\quad + \left[\int_0^T \int_{\mathbb{R}^n\backslash\mathcal{K}} \left(|\partial h| + \frac{|h|}{r}\right)|\nabla_{t,x}\phi|\left(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\right) dx\, dt\right]^{1/2}\end{aligned}$$

for any $T \geq 0$.

Here, in an abuse of notation, we are setting

$$|h| = \sum_{\alpha,\beta=0}^n |h^{\alpha\beta}(t,x)|, \quad |\partial h| = \sum_{\alpha,\beta,\gamma=0}^n |\partial_\gamma h^{\alpha\beta}(t,x)|.$$

Although we shall not use this, we point out that this estimate holds when the regularity of the boundary of $\mathcal{K}$ is merely C^1. Also, the implicit constants in (5.1) are independent of the choice of star-shaped obstacle $\mathcal{K}$.

To prove (5.1), we set

$$(5.2)\qquad Q_{\alpha\beta}[\phi] = \partial_\alpha\phi\partial_\beta\phi - \frac{1}{2}g_{\alpha\beta}\partial^\gamma\phi\partial_\gamma\phi - g_{\alpha\gamma}h^{\gamma\delta}\partial_\delta\phi\partial_\beta\phi + \frac{1}{2}g_{\alpha\beta}h^{\gamma\delta}\partial_\gamma\phi\partial_\delta\phi.$$

Then it is straightforward to check that

$$D^\alpha Q_{\alpha\beta}[\phi] = -(\partial_\beta\phi)F - (\partial_\gamma h^{\gamma\delta})\partial_\delta\phi\partial_\beta\phi + \frac{1}{2}(\partial_\beta h^{\gamma\delta})\partial_\gamma\phi\partial_\delta\phi.$$

As above, we contract this with a radial vector field $X = f(r)\partial_r$, define $P_\alpha[\phi, X] = Q_{\alpha\beta}[\phi]X^\beta$, and compute that

$$\begin{aligned}D^\alpha P_\alpha[\phi,X] &= f'(r)(\partial_r\phi)^2 + \frac{f(r)}{r}|\not\nabla\phi|^2 - \frac{1}{2}\mathrm{tr}\pi\partial^\gamma\phi\partial_\gamma\phi - F(\partial_r\phi)f(r) \\ &\quad - (\partial_\gamma h^{\gamma\delta})\partial_\delta\phi\partial_r\phi f(r) + \frac{1}{2}(\partial_r h^{\gamma\delta})\partial_\gamma\phi\partial_\delta\phi f(r) - x_a h^{a\delta}\partial_\delta\phi\partial_r\phi\frac{f'(r)}{r} \\ &\quad\quad + x_a h^{a\delta}\partial_\delta\phi\partial_r\phi\frac{f(r)}{r^2} - h^{a\delta}\partial_\delta\phi\partial_a\phi\frac{f(r)}{r} + \frac{1}{2}(\mathrm{tr}\pi)h^{\gamma\delta}\partial_\gamma\phi\partial_\delta\phi,\end{aligned}$$

where $\mathrm{tr}\pi$ is given as in (4.8).

By modifying this momentum density as in (4.20) and setting

$$\begin{aligned}(5.3)\quad \bar{P}_\alpha[\phi,X] &= P_\alpha[\phi,X] + \frac{n-1}{2}\Big(\frac{f(r)}{r}\Big)\phi\partial_\alpha\phi \\ &\quad - \frac{n-1}{4}\partial_\alpha\left(\frac{f(r)}{r}\right)|\phi|^2 - \frac{n-1}{2}\left(\frac{f(r)}{r}\right)g_{\alpha\gamma}h^{\gamma\beta}\phi\partial_\beta\phi,\end{aligned}$$

it follows that

$$
\begin{aligned}
(5.4)\qquad D^\alpha \bar P_\alpha[\phi, X] &= f'(r)(\partial_r\phi)^2 + \frac{f(r)}{r}|\nabla\!\!\!/\phi|^2 - \frac{1}{2}f'(r)\partial^\gamma\phi\partial_\gamma\phi \\
&\quad - \frac{n-1}{4}\Delta\left(\frac{f(r)}{r}\right)|\phi|^2 - F(\partial_r\phi)f(r) - \frac{n-1}{2}F\frac{\phi}{r}f(r) \\
&\quad - (\partial_\gamma h^{\gamma\delta})\partial_\delta\phi\left(\partial_r\phi + \frac{n-1}{2}\frac{\phi}{r}\right)f(r) + \frac{1}{2}(\partial_r h^{\gamma\delta})\partial_\gamma\phi\partial_\delta\phi f(r) \\
&\quad - x_a h^{a\delta}\partial_\delta\phi\left(\partial_r\phi + \frac{n-1}{2}\frac{\phi}{r}\right)\frac{f'(r)}{r} \\
&\quad + x_a h^{a\delta}\partial_\delta\phi\left(\partial_r\phi + \frac{n-1}{2}\frac{\phi}{r}\right)\frac{f(r)}{r^2} \\
&\quad - h^{a\delta}\partial_\delta\phi\partial_a\phi\frac{f(r)}{r} + \frac{1}{2}f'(r)h^{\gamma\delta}\partial_\gamma\phi\partial_\delta\phi.
\end{aligned}
$$

Integrating both sides of (5.4) in a time strip S_T yields

$$
\begin{aligned}
(5.5)\qquad \int_{\mathbb{R}^n\setminus\mathcal{K}} \bar P_0[\phi, X](0)\,dx - \int_{\mathbb{R}^n\setminus\mathcal{K}} \bar P_0[\phi, X](T)\,dx - \int_0^T\int_{\partial\mathcal{K}} \bar P_a[\phi, X]n^a\,d\sigma\,dt \\
= \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}} D^\alpha \bar P_\alpha[\phi, X]\,dx\,dt.
\end{aligned}
$$

For f as given by (4.13), we have $|f(r)| \lesssim 1$ and $|f'(r)| \lesssim \frac{1}{r}$. Thus, we see that

$$
\begin{aligned}
\left|\int_{\mathbb{R}^n\setminus\mathcal{K}} \bar P_0[\phi, X](0)\,dx\right| &= \left|\int_{\mathbb{R}^n\setminus\mathcal{K}}\Big(\partial_t\phi(0)\partial_r\phi(0)f(r) - g_{0\gamma}h^{\gamma\delta}(0)\partial_\delta\phi(0)\partial_r\phi(0)f(r)\right. \\
&\qquad \left. + \frac{n-1}{2}\frac{f(r)}{r}\phi(0)\partial_t\phi(0) - \frac{n-1}{2}\frac{f(r)}{r}g_{0\gamma}h^{\gamma\beta}\phi(0)\partial_\beta\phi(0)\Big)\,dx\right| \\
&\lesssim \|\nabla_{t,x}\phi(0)\|_2^2.
\end{aligned}
$$

For the last inequality, we are, as above, using the Schwarz inequality and the Hardy inequality. We are also using the bound (3.3).

A similar estimate holds for the $\bar P_0[\phi, X](T)$ term. And, thus, by the energy inequality (3.5),

$$
\begin{aligned}
\left|\int_{\mathbb{R}^n\setminus\mathcal{K}} \bar P_0[\phi, X](T)\,dx\right| &\lesssim \|\nabla_{t,x}\phi(T)\|_2^2 \\
&\lesssim \|\nabla_{t,x}\phi(0)\|_2^2 + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}} |\partial_t\phi||F|\,dx\,dt \\
&\quad + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}}\Big(|(\partial_\alpha h^{\alpha\beta})\partial_t\phi\partial_\beta\phi| + |(\partial_t h^{\alpha\beta})\partial_\alpha\phi\partial_\beta\phi|\Big)\,dx\,dt.
\end{aligned}
$$

For the remaining boundary terms, we use the fact that the Dirichlet boundary conditions permit us to write $\partial_a\phi = \partial_{\vec n}\phi\, n_a$ on $\partial\mathcal{K}$. Thus, if δ in (3.3) is small enough,

we have

$$
\begin{aligned}
-\int_0^T \int_{\partial\mathcal{K}} \bar{P}_a[\phi, X] n^a \, d\sigma \, dt \\
&= -\frac{1}{2}\int_0^T \int_{\partial\mathcal{K}} \frac{f(r)}{r}\Big((\partial_{\vec{n}}\phi)^2\langle x,\vec{n}\rangle - (\partial_{\vec{n}}\phi)^2(h^{ab}n_a n_b)\langle x,\vec{n}\rangle\Big)\, d\sigma \, dt \\
&\le -\frac{1}{4}\int_0^T \int_{\partial\mathcal{K}} \frac{f(r)}{r}(\partial_{\vec{n}}\phi)^2\langle x,\vec{n}\rangle \, d\sigma \, dt \le 0.
\end{aligned}
$$

For the first inequality, we are using (3.3). For the second inequality, we use the fact that $\langle x, \vec{n}\rangle > 0$ for star-shaped $\mathcal{K}$.

Using these bounds in (5.5), fixing f as in (4.13), and applying (5.4), it follows that

$$
\begin{aligned}
\int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} & f'(r)(\partial_r\phi)^2 + \frac{f(r)}{r}|\not\nabla\phi|^2 - \frac{1}{2}f'(r)\partial^\gamma\phi\partial_\gamma\phi - \frac{n-1}{4}\Delta\left(\frac{f(r)}{r}\right)|\phi|^2 \\
&- (\partial_\gamma h^{\gamma\delta})\partial_\delta\phi\left(\partial_r\phi + \frac{n-1}{2}\frac{\phi}{r}\right)f(r) + \frac{1}{2}(\partial_r h^{\gamma\delta})\partial_\gamma\phi\partial_\delta\phi f(r) \\
&- x_a h^{a\delta}\partial_\delta\phi\left(\partial_r\phi + \frac{n-1}{2}\frac{\phi}{r}\right)\frac{f'(r)}{r} + x_a h^{a\delta}\partial_\delta\phi\left(\partial_r\phi + \frac{n-1}{2}\frac{\phi}{r}\right)\frac{f(r)}{r^2} \\
&- h^{a\delta}\partial_\delta\phi\partial_a\phi\frac{f(r)}{r} + \frac{1}{2}f'(r)h^{\gamma\delta}\partial_\gamma\phi\partial_\delta\phi \, dx \, dt \\
&\lesssim \|\nabla_{t,x}\phi(0)\|_2^2 + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}}\left(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\right)|F| \, dx \, dt \\
&\quad + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}}\Big(|(\partial_\alpha h^{\alpha\beta})\partial_t\phi\partial_\beta\phi| + |(\partial_t h^{\alpha\beta})\partial_\alpha\phi\partial_\beta\phi|\Big)\, dx \, dt.
\end{aligned}
$$

Using (4.23), this yields

$$
\begin{aligned}
\int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} & \frac{\rho}{(r+\rho)^2}(\partial_r\phi)^2 + \frac{1}{r+\rho}|\not\nabla\phi|^2 + \frac{\rho}{(r+\rho)^2}(\partial_t\phi)^2 \\
&+ \frac{\rho}{r(r+\rho)^3}|\phi|^2 \, dx \, dt \\
&\lesssim \|\nabla_{t,x}\phi(0)\|_2^2 + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}}\Big(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\Big)|F| \, dx \, dt \\
&\quad + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}}\Big(|\partial h| + \frac{|h|}{r}\Big)|\nabla_{t,x}\phi|\Big(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\Big)\, dx \, dt,
\end{aligned} \tag{5.6}
$$

since $f'(r) < f(r)/r$. Thus, it follows that

$$
\begin{aligned}
\int_0^T \int_{|x|\le 1} & \Big(|\nabla_x\phi|^2 + |\partial_t\phi|^2 + |\phi|^2\Big)\, dx \, dt \\
&\lesssim \|\nabla_{t,x}\phi(0)\|_2^2 + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}}\left(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\right)|F| \, dx \, dt \\
&\quad + \int_0^T\int_{\mathbb{R}^n\setminus\mathcal{K}}\left(|\partial h| + \frac{|h|}{r}\right)|\nabla_{t,x}\phi|\left(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\right)\, dx \, dt
\end{aligned} \tag{5.7}
$$

for the choice $\rho = 1$ (since $0 \in \mathcal{K}$ and thus $1/r$ is bounded on the complement of $\mathcal{K}$) and

$$\begin{aligned}(5.8)\quad \int_0^T \int_{2^{k-1}\le|x|\le 2^k} \left(\frac{|\nabla_x \phi|^2}{r} + \frac{|\partial_t \phi|^2}{r} + \frac{|\phi|^2}{r^3}\right) dx\, dt \\ \lesssim \|\nabla_{t,x}\phi(0)\|_2^2 + \int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} \left(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\right)|F|\, dx\, dt \\ + \int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} \left(|\partial h| + \frac{|h|}{r}\right)|\nabla_{t,x}\phi| \left(|\nabla_{t,x}\phi| + \frac{|\phi|}{r}\right) dx\, dt\end{aligned}$$

for $\rho = 2^k$. If we sum over $k \ge 1$ as above, (5.1) follows, which completes the proof. □

If we use an elliptic regularity argument as above, the following lemma holds.

LEMMA 5.2. *Suppose that $\mathcal{K}$ is as above and $n \ge 3$. Let $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^n\setminus\mathcal{K})$ be a solution of* (3.1). *Suppose that $h^{\alpha\beta}$ satisfies* (3.2) *and* (3.3) *for a small choice of δ. Then, we have*

$$\begin{aligned}(5.9)\quad (\log(2+T))^{-1/2} \sum_{|\mu|\le N} \|\langle x\rangle^{-1/2}\partial^\mu_{t,x}\phi'\|_{L^2_t L^2_x(S_T)} \\ + \sum_{|\mu|\le N} \|\langle x\rangle^{-1/2-}\partial^\mu_{t,x}\phi'\|_{L^2_t L^2_x(S_T)} \\ \lesssim \sum_{j\le N} \|\partial_t^j \phi'(0,\cdot)\|_2 + \sum_{j,k\le N} \left(\int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} \left(|\nabla_{t,x}\partial_t^j\phi| + \frac{|\partial_t^j\phi|}{r}\right)|\Box_h \partial_t^k \phi|\, dx\, dt\right)^{1/2} \\ + \sum_{j,k\le N} \left[\int_0^T \int_{\mathbb{R}^n\setminus\mathcal{K}} \left(|\partial h| + \frac{|h|}{r}\right)|\nabla_{t,x}\partial_t^j\phi| \left(|\nabla_{t,x}\partial_t^k\phi| + \frac{|\partial_t^k\phi|}{r}\right) dx\, dt\right]^{1/2} \\ + \sum_{|\mu|\le N-1} \|\Box \partial^\mu_{t,x}\phi\|_{L^2_t L^2_x(S_T)}\end{aligned}$$

for any $T > 0$ and $N = 0, 1, 2, \ldots$.

Indeed, in order to obtain (5.9), we argue inductively where (5.1) is the base case ($N = 0$). We then notice that

$$\begin{aligned}(5.10)\quad \sum_{|\mu|\le N} \|\langle x\rangle^{-1/2-}\partial^\mu_{t,x}\phi'\|_{L^2_t L^2_x(S_T)} \le \sum_{|\mu|\le N-1} \|\langle x\rangle^{-1/2-}\partial^\mu_{t,x}\partial_x^2\phi\|_{L^2_t L^2_x(S_T)} \\ + \sum_{|\mu|\le N-1} \|\langle x\rangle^{-1/2-}\partial^\mu_{t,x}(\partial_t\phi)'\|_{L^2_t L^2_x(S_T)} + \|\langle x\rangle^{-1/2-}\phi'\|_{L^2_t L^2_x(S_T)}.\end{aligned}$$

The estimate for the last term on the right follows trivially from (5.1). Since ∂_t preserves the Dirichlet boundary condition, we can use the inductive hypothesis to bound the second term on the right.

In order to bound the first term on the right side of (5.10), we will use elliptic regularity. To see this, we fix a smooth cutoff function β with $\beta \equiv 1$ for $1/2 < |x| < 1$ and $\beta \equiv 0$ outside of $1/4 \le |x| \le 2$. Applying elliptic regularity to $\beta(x/R)\phi(t,x)$,

we see that

$$\sum_{|\mu|\le N-1} \|\partial_{t,x}^{\mu}\partial_x^2\phi(t,\cdot)\|_{L_x^2(\{|x|\in[R/2,R]\})} \lesssim \sum_{|\mu|\le N-1} \|\partial_{t,x}^{\mu}\Delta\phi(t,\cdot)\|_{L_x^2(\{|x|\in[R/4,2R]\})} + \sum_{|\mu|\le N-1} \|\partial_{t,x}^{\mu}\phi'(t,\cdot)\|_{L_x^2(\{|x|\in[R/4,2R]\})} + \|\langle x\rangle^{-1}\phi(t,\cdot)\|_{L_x^2(\{|x|\in[R/4,2R]\})} \tag{5.11}$$

for $R \ge 2$. Similarly

$$\sum_{|\mu|\le N-1} \|\partial_{t,x}^{\mu}\partial_x^2\phi(t,\cdot)\|_{L_x^2(\{|x|\le 1\})} \lesssim \sum_{|\mu|\le N-1} \|\partial_{t,x}^{\mu}\Delta\phi(t,\cdot)\|_{L_x^2(\{|x|\le 2\})} + \sum_{|\mu|\le N-1} \|\partial_{t,x}^{\mu}\phi'(t,\cdot)\|_{L_x^2(\{|x|\le 2\})},$$

where we have used the fact that the Dirichlet boundary conditions allow us to control ϕ locally by ϕ'. By multiplying both sides of (5.11) by $R^{-1/2-}$, summing over $R = 2^k$, $k = 1, 2, \ldots$, and integrating in time, we see that

$$\sum_{|\mu|\le N-1} \|\langle x\rangle^{-1/2-}\partial_{t,x}^{\mu}\partial_x^2\phi\|_{L_t^2L_x^2(S_T)} \lesssim \sum_{|\mu|\le N-1} \|\langle x\rangle^{-1/2-}\partial_{t,x}^{\mu}\Delta\phi\|_{L_t^2L_x^2(S_T)} + \sum_{|\mu|\le N-1} \|\langle x\rangle^{-1/2-}\partial_{t,x}^{\mu}\phi'\|_{L_t^2L_x^2(S_T)} + \|\langle x\rangle^{-3/2-}\phi\|_{L_t^2L_x^2(S_T)}. \tag{5.12}$$

The estimates for the last two terms can, again, be obtained by the inductive hypothesis and (5.1), respectively.

For the first term in the right side of (5.12), we simply notice that

$$\sum_{|\mu|\le N-1} \|\langle x\rangle^{-1/2-}\partial_{t,x}^{\mu}\Delta\phi\|_{L_t^2L_x^2(S_T)} \lesssim \sum_{|\mu|\le N-1} \|\langle x\rangle^{-1/2-}\partial_{t,x}^{\mu}\partial_t^2\phi\|_{L_t^2L_x^2(S_T)} + \sum_{|\mu|\le N-1} \|\partial_{t,x}^{\mu}\Box\phi\|_{L_t^2L_x^2(S_T)}.$$

As ∂_t preserves the boundary condition, we may use the inductive hypothesis to see that the first term on the right is bounded by the right side of (5.9). As a similar argument may be used to obtain the estimate for the first term on the left of (5.9), the proof of Lemma 5.2 is complete. □

Similarly, if as above (Lemma 3.3), we repeat the argument with ϕ replaced by $Z^{\mu}\phi$ for some multi-index μ, we see that the following holds.

LEMMA 5.3. *Suppose that $\mathcal{K}$ is as above and $n \geq 3$. Let $\phi \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^n \backslash \mathcal{K})$ be a solution of (3.1). Suppose that $h^{\alpha\beta}$ satisfies (3.2) and (3.3) for a small choice of δ. Then, we have*

$$
\begin{aligned}
(5.13)\qquad & (\log(2+T))^{-1/2} \sum_{|\mu|\leq N} \|\langle x\rangle^{-1/2} Z^\mu \phi'\|_{L^2_t L^2_x(S_T)} \\
& \qquad + \sum_{|\mu|\leq N} \|\langle x\rangle^{-1/2-} Z^\mu \phi'\|_{L^2_t L^2_x(S_T)} \\
\lesssim & \sum_{|\mu|\leq N} \|Z^\mu \phi'(0,\,\cdot\,)\|_2 + \sum_{|\mu|,|\nu|\leq N} \left(\int_0^T \int_{\mathbb{R}^n\backslash\mathcal{K}} \left(|\nabla_{t,x} Z^\mu \phi| + \frac{|Z^\mu \phi|}{r} \right) |\Box_h Z^\nu \phi| \, dx\, dt \right)^{1/2} \\
& + \sum_{|\mu|,|\nu|\leq N} \left[\int_0^T \int_{\mathbb{R}^n\backslash\mathcal{K}} \left(|\partial h| + \frac{|h|}{r} \right) |\nabla_{t,x} Z^\mu \phi| \left(|\nabla_{t,x} Z^\nu \phi| + \frac{|Z^\nu \phi|}{r} \right) dx\, dt \right]^{1/2} \\
& \qquad + \sum_{|\mu|\leq N+1} \|\partial_x^\mu \phi'\|_{L^2_t L^2_x([0,T]\times\{|x|<1\})}
\end{aligned}
$$

for any $T > 0$ and $N = 0, 1, 2, \ldots$.

In particular, notice that the last term of (5.13) can be controlled using (5.9). The same is true for the boundary term of (3.10).

6. Almost global existence for $n = 3$. In this section, we shall prove Theorem 1.1. Since (5.9) and (5.13) have been established, this proof will resemble that of the semilinear case [8], which is much easier than the subsequent proofs for the quasilinear case. We shall use an iteration argument to solve (1.1) and to show that the solution satisfies

$$
\begin{aligned}
(6.1)\qquad & \sum_{|\mu|\leq 15} \Big(\|\partial^\mu u'(t,\,\cdot\,)\|_2 + (\log(2+t))^{-1/2} \|\langle x\rangle^{-1/2} \partial^\mu u'\|_{L^2_t L^2_x(S_t)} \Big) \\
& \qquad + \sum_{|\mu|\leq 14} \Big(\|Z^\mu u'(t,\,\cdot\,)\|_2 + (\log(2+t))^{-1/2} \|\langle x\rangle^{-1/2} Z^\mu u'\|_{L^2_t L^2_x(S_t)} \Big) \leq C\varepsilon
\end{aligned}
$$

for $0 \leq t \leq T_\varepsilon$ and for uniform constant C.

Here, we let $u_{-1} \equiv 0$ and then recursively define u_k, $k = 0, 1, 2, \ldots$, to solve

$$
(6.2)\qquad \begin{cases} \Box u_k(t,x) = Q(du_{k-1}, d^2 u_k), \quad (t,x) \in [0,T] \times \mathbb{R}^3 \backslash \mathcal{K}, \\ u_k|_{\partial\mathcal{K}} = 0, \\ u_k(0,\,\cdot\,) = f, \quad \partial_t u_k(0,\,\cdot\,) = g. \end{cases}
$$

We set

$$
\begin{aligned}
(6.3)\qquad M_k(T) = \sup_{0\leq t\leq T} & \left[\sum_{|\mu|\leq 15} \Big(\|\partial^\mu u_k'(t,\,\cdot\,)\|_2 + (\log(2+t))^{-1/2} \|\langle x\rangle^{-1/2} \partial^\mu u_k'\|_{L^2_t L^2_x(S_t)} \Big) \right. \\
& \left. + \sum_{|\mu|\leq 14} \Big(\|Z^\mu u_k'(t,\,\cdot\,)\|_2 + (\log(2+t))^{-1/2} \|\langle x\rangle^{-1/2} Z^\mu u_k'\|_{L^2_t L^2_x(S_t)} \Big) \right].
\end{aligned}
$$

Clearly, by (1.3), the standard energy inequality, and (5.9) and (5.13) (with $h^{\alpha\beta} \equiv 0$), there is a uniform constant C_0 so that

$$M_0(T) \le C_0\varepsilon$$

for any T. Here, C_0 can be chosen to be larger than the implicit constants of (3.9), (3.10), (5.9), and (5.13). For $\varepsilon < \varepsilon_0$ sufficiently small and for κ in (1.5) small, we will show inductively that for $k = 1, 2, 3, \dots$

$$M_k(T_\varepsilon) \le 10C_0\varepsilon. \tag{6.4}$$

By (1.4), (3.9), (3.10), (5.9), and (5.13), we have

(6.5)

$$\begin{aligned}
M_k(T_\varepsilon) \le 4C_0\varepsilon &+ \sum_{|\mu|,|\nu|\le 15} \Big(\int_0^{T_\varepsilon}\int_{\mathbb{R}^3\setminus\mathcal{K}} \Big(|\nabla_{t,x}\partial^\mu u_k| + \frac{|\partial^\mu u_k|}{r}\Big)|\partial^\nu \Box_h u_k|\,dx\,dt\Big)^{1/2} \\
&+ \sum_{|\mu|,|\nu|\le 15} \Big(\int_0^{T_\varepsilon}\int_{\mathbb{R}^3\setminus\mathcal{K}} \Big(|\nabla_{t,x}\partial^\mu u_k| + \frac{|\partial^\mu u_k|}{r}\Big)|[\partial^\nu, \Box_h]u_k|\,dx\,dt\Big)^{1/2} \\
&+ \sum_{|\mu|,|\nu|\le 15} \Big[\int_0^{T_\varepsilon}\int_{\mathbb{R}^3\setminus\mathcal{K}} \Big(|\partial h| + \frac{|h|}{r}\Big)|\nabla_{t,x}\partial^\mu u_k|\Big(|\nabla_{t,x}\partial^\nu u_k| + \frac{|\partial^\nu u_k|}{r}\Big)\,dx\,dt\Big]^{1/2} \\
&+ \sum_{|\mu|,|\nu|\le 14} \Big(\int_0^{T_\varepsilon}\int_{\mathbb{R}^3\setminus\mathcal{K}} \Big(|\nabla_{t,x}Z^\mu u_k| + \frac{|Z^\mu u_k|}{r}\Big)|Z^\nu \Box_h u_k|\,dx\,dt\Big)^{1/2} \\
&+ \sum_{|\mu|,|\nu|\le 14} \Big(\int_0^{T_\varepsilon}\int_{\mathbb{R}^3\setminus\mathcal{K}} \Big(|\nabla_{t,x}Z^\mu u_k| + \frac{|Z^\mu u_k|}{r}\Big)|[Z^\nu, \Box_h]u_k|\,dx\,dt\Big)^{1/2} \\
&+ \sum_{|\mu|,|\nu|\le 14} \Big[\int_0^{T_\varepsilon}\int_{\mathbb{R}^3\setminus\mathcal{K}} \Big(|\partial h| + \frac{|h|}{r}\Big)|\nabla_{t,x}Z^\mu u_k|\Big(|\nabla_{t,x}Z^\mu u_k| + \frac{|Z^\mu u_k|}{r}\Big)\,dx\,dt\Big]^{1/2} \\
&+ \sup_{0\le t\le T_\varepsilon}\Big[\sum_{|\mu|\le 14}\|\partial^\mu_{t,x}\Box u_k(t,\cdot)\|_2\Big] + \sum_{|\mu|\le 14}\|\partial^\mu_{t,x}\Box u_k\|_{L^2_tL^2_x(S_{T_\varepsilon})}.
\end{aligned}$$

Here, we set $h^{\gamma\delta} = -\sum_{0\le\beta\le 3} B^{\gamma\delta}_\beta \partial_\beta u_{k-1}$. We note that

$$\begin{aligned}
&\sum_{|\mu|\le 15}\Big(|\partial^\mu \Box_h u_k| + \sum_{|\mu|\le 15}|[\partial^\mu, \Box_h]u_k|\Big) \\
&\qquad \lesssim \sum_{|\mu|\le 7}|\partial^\mu u'_{k-1}| \sum_{|\nu|\le 15}|\partial^\nu u'_k| + \sum_{|\mu|\le 8}|\partial^\mu u'_k| \sum_{|\nu|\le 15}|\partial^\nu u'_{k-1}| \\
&\qquad\qquad + \sum_{|\mu|\le 7}|\partial^\mu u'_{k-1}| \sum_{|\nu|\le 15}|\partial^\nu u'_{k-1}|.
\end{aligned}$$

Similarly,

$$\sum_{|\mu|\le 14}\Big(|Z^\mu \Box_h u_k| + \sum_{|\mu|\le 14}|[Z^\mu,\Box_h]u_k|\Big)$$

$$\lesssim \sum_{|\mu|\le 7}|Z^\mu u'_{k-1}|\sum_{|\nu|\le 14}|Z^\nu u'_k| + \sum_{|\mu|\le 8}|Z^\mu u'_k|\sum_{|\nu|\le 14}|Z^\nu u'_{k-1}|$$

$$+\sum_{|\mu|\le 7}|Z^\mu u'_{k-1}|\sum_{|\nu|\le 14}|Z^\nu u'_{k-1}|.$$

If we use this, (2.1), the Schwarz inequality, and a Hardy inequality, it follows that the second, third, and fourth terms in (6.5) are controlled by

$$\sum_{|\mu|\le 9}\|\langle x\rangle^{-1/2}Z^\mu u'_{k-1}\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}\sum_{|\mu|\le 15}\|\langle x\rangle^{-1/2}\partial^\mu u'_k\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}$$

$$\times \sup_{0\le t\le T_\varepsilon}\sum_{|\mu|\le 15}\|\partial^\mu u'_k(t,\,\cdot\,)\|_2^{1/2}$$

$$+\left(\sum_{|\mu|\le 10}\|\langle x\rangle^{-1/2}Z^\mu u'_k\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}+\sum_{|\mu|\le 9}\|\langle x\rangle^{-1/2}Z^\mu u'_{k-1}\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}\right)$$

$$\times\sum_{|\mu|\le 15}\|\langle x\rangle^{-1/2}\partial^\mu u'_{k-1}\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}\left(\sup_{0\le t\le T_\varepsilon}\sum_{|\mu|\le 15}\|\partial^\mu u'_k(t,\,\cdot\,)\|_2^{1/2}\right). \tag{6.6}$$

In particular, for the fourth term in (6.5), notice that for the given choice of h, we have

$$\left(|\partial h|+\frac{|h|}{r}\right)\lesssim\sum_{|\mu|\le 1}|\partial^\mu u'_{k-1}|.$$

Here, we use that $\frac{1}{r}$ is bounded on $\mathbb{R}^3\backslash\mathcal{K}$ since $0\in\mathcal{K}$. Thus, by (2.1) and the Schwarz inequality, we have

$$\sum_{|\mu|,|\nu|\le 15}\int_{2^{j-1}}^{2^j}\left(|\partial h|+\frac{|h|}{r}\right)|\nabla_{t,x}\partial^\mu u_k(t,x)|\left(|\nabla_{t,x}\partial^\nu u_k(t,x)|+\frac{|\partial^\nu u_k(t,x)|}{r}\right)\,dx$$

$$\lesssim\sum_{|\mu|\le 3}\|\langle x\rangle^{-1/2}Z^\mu u'_{k-1}(t,\,\cdot\,)\|_{L^2(2^{j-2}\le|x|\le 2^{j+1})}\sum_{|\mu|\le 15}\|\langle x\rangle^{-1/2}\partial^\mu u'_k(t,\,\cdot\,)\|_{L^2(2^{j-1}\le|x|\le 2^j)}$$

$$\times\left(\sum_{|\mu|\le 15}\|\partial^\nu u'_k(t,\,\cdot\,)\|_2+\left\|\frac{1}{r}u_k(t,\,\cdot\,)\right\|_2\right).$$

A similar bound holds over $|x|\in[0,1]$. If we sum over j and integrate in t, it follows that the fourth term in (6.5) is indeed controlled by the first term in (6.6). Here, we use a Hardy inequality to gain the control $\|(1/r)u_k(t,\,\cdot\,)\|_2\lesssim\|u'_k(t,\,\cdot\,)\|_2$. Similar arguments yield the given bounds for the second and third terms in (6.5).

By the inductive hypothesis, we see that (6.6) is controlled by

$$C_1\varepsilon^{1/2}(\log(2+T_\varepsilon))^{1/2}M_k(T_\varepsilon)+C_1\varepsilon^{3/2}. \tag{6.7}$$

This provides the necessary bound for the second, third, and fourth terms in (6.5).

Similarly, the fifth, sixth, and seventh terms in (6.5) are bounded by

$$\begin{aligned}\sum_{|\mu|\le 9}\|\langle x\rangle^{-1/2}Z^\mu u'_{k-1}\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}&\sum_{|\mu|\le 14}\|\langle x\rangle^{-1/2}Z^\mu u'_k\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}\\ &\times\sup_{0\le t\le T_\varepsilon}\sum_{|\mu|\le 14}\|Z^\mu u'_k(t,\,\cdot\,)\|_2^{1/2}\\ +\left(\sum_{|\mu|\le 10}\|\langle x\rangle^{-1/2}Z^\mu u'_k\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}+\sum_{|\mu|\le 9}\|\langle x\rangle^{-1/2}Z^\mu u'_{k-1}\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}\right)&\\ \times\sum_{|\mu|\le 14}\|\langle x\rangle^{-1/2}Z^\mu u'_{k-1}\|^{1/2}_{L^2_tL^2_x(S_{T_\varepsilon})}&\left(\sup_{0\le t\le T_\varepsilon}\sum_{|\mu|\le 14}\|Z^\mu u'_k(t,\,\cdot\,)\|_2^{1/2}\right),\end{aligned}$$

which as above is controlled by

$$C_2\varepsilon^{1/2}(\log(2+T_\varepsilon))^{1/2}M_k(T_\varepsilon)+C_2\varepsilon^{3/2}. \tag{6.8}$$

By similar arguments, it follows that the last two terms of (6.5) are controlled by

$$C_3\Big(\varepsilon M_k(T_\varepsilon)+\varepsilon^2+\varepsilon(\log(2+T_\varepsilon))^{1/2}M_k(T_\varepsilon)+\varepsilon^2(\log(2+T_\varepsilon))^{1/2}\Big). \tag{6.9}$$

If we have

$$(C_1+C_2+C_3)\varepsilon^{1/2}(\log(2+T_\varepsilon))^{1/2}\le\frac{1}{2},$$

which indeed is the case if κ is chosen to be smaller than $\frac{1}{2(C_1+C_2+C_3)^2}$, we can bootstrap the terms in (6.7), (6.8), and (6.9) involving $M_k(T_\varepsilon)$. Thus, we see that

$$M_k(T_\varepsilon)\le 2[4C_0\varepsilon+(C_1+C_2+C_3\kappa^{1/2}(\log 2)^{1/2})\varepsilon^{3/2}].$$

Thus, if ε is small enough, we obtain (6.4) as desired.

If we define

$$\begin{aligned}A_k(T)=\sup_{0\le t\le T}\Bigg[&\sum_{|\mu|\le 14}\Big(\|\partial^\mu(u'_k-u'_{k-1})(t,\,\cdot\,)\|_2\\ &+(\log(2+t))^{-1/2}\|\langle x\rangle^{-1/2}\partial^\mu(u'_k-u'_{k-1})\|_{L^2_tL^2_x(S_t)}\Big)\\ &+\sum_{|\mu|\le 13}\Big(\|Z^\mu(u'_k-u'_{k-1})(t,\,\cdot\,)\|_2\\ &+(\log(2+t))^{-1/2}\|\langle x\rangle^{-1/2}Z^\mu(u'_k-u'_{k-1})\|_{L^2_tL^2_x(S_t)}\Big)\Bigg],\end{aligned}$$

similar arguments can be used to show that

$$A_k(T_\varepsilon)\le\frac{1}{2}A_{k-1}(T_\varepsilon).$$

Thus, we have that u_k converges to a solution of (1.1) satisfying (6.1), which completes the proof.

7. Global existence in higher dimensions. In this last section, we provide a few remarks that explain how to modify the proof in the previous section in order to obtain a proof of Theorem 1.2. Indeed, it is possible to show via iteration that a solution exists and satisfies

$$\sum_{|\mu|\le n+10}\Big(\|\partial^\mu u'(t,\,\cdot\,)\|_2+\|\langle x\rangle^{-(n-1)/4}\partial^\mu u'\|_{L^2_tL^2_x(S_t)}\Big)+\sum_{|\mu|\le n+9}\Big(\|Z^\mu u'(t,\,\cdot\,)\|_2+\|\langle x\rangle^{-(n-1)/4}Z^\mu u'\|_{L^2_tL^2_x(S_t)}\Big)\le C\varepsilon \tag{7.1}$$

for any $T>0$. Here, we argue as in the previous section. When we apply the weighted Sobolev estimate, we get weights $\langle x\rangle^{-(n-1)/4}$. When $n\ge 4$, we have $(n-1)/4>1/2$, and thus, we may apply the bound for the first term in the left side of (5.1) rather than that for the second term. Since the first term in the left side of (5.1) does not require the loss of a $\log(2+T)^{1/2}$, we see immediately that we have no restriction on T_ε, which proves the desired global existence result.

Acknowledgments. It is a pleasure to thank J. Sterbenz for pointing out the work of Rodnianski. The authors are also grateful to the anonymous referees for their thorough reading of this manuscript and for their helpful suggestions.

REFERENCES

[1] S. Alinhac, *Remarks on energy inequalities for wave and Maxwell equations on a curved background*, Math. Ann., 329 (2004), pp. 707–722.
[2] S. Alinhac, *On the Morawetz/KSS Inequality for the Wave Equation on a Curved Background*, preprint, 2005.
[3] N. Hayashi, *Global existence of small solutions to quadratic nonlinear wave equations in an exterior domain*, J. Funct. Anal., 131 (1995), pp. 302–344.
[4] K. Hidano and K. Yokoyama, *A remark on the almost global existence theorems of Keel, Smith, and Sogge*, Funkcial. Ekvac., 48 (2005), pp. 1–34.
[5] F. John, *Blow-up for quasilinear wave equations in three space dimensions*, Comm. Pure Appl. Math., 34 (1981), pp. 29–51.
[6] F. John and S. Klainerman, *Almost global existence to nonlinear wave equations in three dimensions*, Comm. Pure Appl. Math., 37 (1984), pp. 443–455.
[7] M. Keel, H. Smith, and C. D. Sogge, *Global existence for a quasilinear wave equation outside of star-shaped domains*, J. Funct. Anal., 189 (2002), pp. 155–226.
[8] M. Keel, H. Smith, and C. D. Sogge, *Almost global existence for some semilinear wave equations*, J. Anal. Math., 87 (2002), pp. 265–279.
[9] M. Keel, H. Smith, and C. D. Sogge, *Almost global existence for quasilinear wave equations in three space dimensions*, J. Amer. Math. Soc., 17 (2004), pp. 109–153.
[10] S. Klainerman, *The null condition and global existence to nonlinear wave equations*, in Nonlinear Systems of Partial Differential Equations in Applied Mathematics, Part 1, Lectures in Appl. Math. 23, AMS, Providence, RI, 1986, pp. 293–326.
[11] S. Klainerman and T. Sideris, *On almost global existence for nonrelativistic wave equations in* $3d$, Comm. Pure Appl. Math., 49 (1996), pp. 307–321.
[12] P. D. Lax, C. S. Morawetz, and R. S. Phillips, *Exponential decay of solutions of the wave equation in the exterior of a star-shaped obstacle*, Comm. Pure Appl. Math., 16 (1963), pp. 477–486.
[13] J. Metcalfe, *Global existence for semilinear wave equations exterior to nontrapping obstacles*, Houston J. Math., 30 (2004), pp. 259–281.
[14] J. Metcalfe, M. Nakamura, and C. D. Sogge, *Global existence of solutions to multiple speed systems of quasilinear wave equations in exterior domains*, Forum Math., 17 (2005), pp. 133–168.
[15] J. Metcalfe, M. Nakamura, and C. D. Sogge, *Global existence of quasilinear, nonrelativistic wave equations satisfying the null condition*, Japan. J. Math., to appear.

[16] J. Metcalfe and C. D. Sogge, *Hyperbolic trapped rays and global existence of quasilinear wave equations*, Invent. Math., 159 (2005), pp. 75–117.
[17] J. Metcalfe and C. D. Sogge, *Global existence for Dirichlet-wave equations with quadratic nonlinearities in high dimensions*, Math. Ann., submitted; also available online from http://arxiv.org/abs/math.AP/0404420, 2004.
[18] C. S. Morawetz, *The decay of solutions of the exterior initial-boundary problem for the wave equation*, Comm. Pure Appl. Math., 14 (1961), pp. 561–568.
[19] C. S. Morawetz, *Time decay for the nonlinear Klein-Gordon equations*, Proc. Roy. Soc. Ser. A, 306 (1968), pp. 291–296.
[20] Y. Shibata and Y. Tsutsumi, *On a global existence theorem of small amplitude solutions for nonlinear wave equations in an exterior domain*, Math. Z., 191 (1986), pp. 165–199.
[21] T. C. Sideris, *Global behavior of solutions to nonlinear wave equations in three dimensions*, Comm. Partial Differential Equations, 8 (1983), pp. 1291–1323.
[22] T. C. Sideris, *Nonresonance and global existence of prestressed nonlinear elastic waves*, Ann. of Math., 151 (2000), pp. 849–874.
[23] T. C. Sideris, *The null condition and global existence of nonlinear elastic waves*, Invent. Math., 123 (1996), pp. 323–342.
[24] T. C. Sideris and S. Y. Tu, *Global existence for systems of nonlinear wave equations in* $3D$ *with multiple speeds*, SIAM J. Math. Anal., 33 (2001), pp. 477–488.
[25] H. Smith and C. D. Sogge, *Global Strichartz estimates for nontrapping perturbations of the Laplacian*, Comm. Partial Differential Equations, 25 (2000), pp. 2171–2183.
[26] J. Sterbenz, *Angular regularity and Strichartz estimates for the wave equation, with an appendix by I. Rodnianski*, Int. Math. Res. Not. (2005), pp. 187–231.
[27] W. A. Strauss, *Dispersal of waves vanishing on the boundary of an exterior domain*, Comm. Pure Appl. Math., 28 (1975), pp. 265–278.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 210–232

NAVIER–STOKES EQUATIONS WITH NAVIER BOUNDARY CONDITIONS FOR A BOUNDED DOMAIN IN THE PLANE*

JAMES P. KELLIHER†

Abstract. We consider solutions to the Navier–Stokes equations with Navier boundary conditions in a bounded domain Ω in $\mathbb{R}^2$ with a C^2-boundary Γ. Navier boundary conditions can be expressed in the form $\omega(v) = (2\kappa - \alpha)v \cdot \boldsymbol{\tau}$ and $v \cdot \mathbf{n} = 0$ on Γ, where v is the velocity, $\omega(v)$ the vorticity, $\mathbf{n}$ a unit normal vector, $\boldsymbol{\tau}$ a unit tangent vector, and α is in $L^\infty(\Gamma)$. These boundary conditions were studied in the special case where $\alpha = 2\kappa$ by J.-L. Lions and P.-L. Lions. We establish the existence, uniqueness, and regularity of such solutions, extending the work of Clopeau, Mikelić, and Robert and of Lopes Filho, Nussenzveig Lopes, and Planas, which was restricted to simply connected domains and nonnegative α.

Assuming a particular bound on the growth of the L^p-norms of the initial vorticity with p (*Yudovich vorticity*), and also assuming additional smoothness on Γ and α, we obtain a uniform-in-time bound on the rate of convergence in $L^2(\Omega)$ of solutions to the Navier–Stokes equations with Navier boundary conditions to the solution to the Euler equations in the vanishing viscosity limit. We also show that for smoother initial velocities, the solutions to the Navier–Stokes equations with Navier boundary conditions converge uniformly in time in $L^2(\Omega)$, and L^2 in time in $H^1(\Omega)$, to the solution to the Navier–Stokes equations with the usual no-slip boundary conditions as we let α grow large uniformly on the boundary.

Key words. Navier–Stokes equations, vanishing viscosity limit

AMS subject classifications. 76D05, 76B99

DOI. 10.1137/040612336

1. Introduction.

Let Ω be a bounded domain in $\mathbb{R}^2$ with a C^2-boundary Γ consisting of a finite number of connected components, and let $\mathbf{n}$ and $\boldsymbol{\tau}$ be unit normal and tangent vectors, respectively, to Γ. We follow the convention that $\mathbf{n}$ is an outward normal vector and that the ordered pair $(\mathbf{n}, \boldsymbol{\tau})$ gives the standard orientation to $\mathbb{R}^2$. Define the rate-of-strain tensor,

$$D(v) = \frac{1}{2}\left[\nabla v + (\nabla v)^T\right].$$

We consider the existence, uniqueness, and regularity of solutions to the Navier–Stokes equations with *Navier boundary conditions*. These boundary conditions, introduced by Navier in [19] and derived by Maxwell in [18] from the kinetic theory of gases (see [12]), assume that the tangential "slip" velocity, rather than being zero, is proportional to the tangential stress. With a factor of proportionality a in $L^\infty(\Gamma)$, we can express Navier boundary conditions for a sufficiently regular vector field v as

$$v \cdot \mathbf{n} = 0 \text{ and } 2\nu(\mathbf{n} \cdot D(v)) \cdot \boldsymbol{\tau} + av \cdot \boldsymbol{\tau} = 0 \text{ on } \Gamma. \tag{1.1}$$

We will find it more convenient, however, to let $\alpha = a/\nu$, and write these boundary conditions in the form

$$v \cdot \mathbf{n} = 0 \text{ and } 2(\mathbf{n} \cdot D(v)) \cdot \boldsymbol{\tau} + \alpha v \cdot \boldsymbol{\tau} = 0 \text{ on } \Gamma. \tag{1.2}$$

*Received by the editors July 27, 2004; accepted for publication (in revised form) December 8, 2005; published electronically April 12, 2006. The author was supported in part by a Hubert S. Wall Memorial Fellowship in Mathematics at the University of Texas at Austin during the course of this work.

http://www.siam.org/journals/sima/38-1/61233.html

†Department of Mathematics, Brown University, 151 Thayer Street, Providence, RI 02912 (kelliher@math.brown.edu).

(We give an equivalent form of Navier boundary conditions in Corollary 4.2.)

The reason for preferring the second form for the boundary conditions is that, in the vanishing viscosity limit, we will hold α fixed as we let ν approach zero, and we will show that the solution to the Navier–Stokes equations with Navier boundary conditions converges to a solution to the Euler equations. (See, however, the comment at the end of section 8.)

J.-L. Lions in [15, pp. 87–98] and P.-L. Lions in [16, pp. 129–131] consider the following boundary conditions, which we call *Lions* boundary conditions:

$$v \cdot \mathbf{n} = 0 \text{ and } \omega(v) = 0 \text{ on } \Gamma,$$

where $\omega(v) = \partial_1 v^2 - \partial_2 v^1$ is the vorticity of v. Lions boundary conditions are the special case of Navier boundary conditions in which $\alpha = 2\kappa$, as we show in Corollary 4.3.

J.-L. Lions, in Theorem 6.10 on page 88 of [15], proves existence and uniqueness of a solution to the Navier–Stokes equations in the special case of Lions boundary conditions but includes the assumption that the initial vorticity is bounded. With the same assumption of bounded initial vorticity, the existence and uniqueness are established in Theorem 4.1 of [5] for Navier boundary conditions, under the restriction that α is nonnegative (and in $C^2(\Gamma)$). This is the usual restriction, which is imposed to ensure the conservation of energy. Mathematically, negative values of α present no real difficulty, so we do not make that restriction (except in section 9). The only clear gain from removing the restriction, however, is that it allows us to view Lions boundary conditions as a special case of Navier boundary conditions for more than just convex domains (nonnegative curvature).

P.-L. Lions establishes an energy inequality on page 130 of [16] that can be used in place of the usual one for no-slip boundary conditions. He argues that existence and uniqueness can then be established for initial velocity in $L^2(\Omega)$—and no additional assumption on the initial vorticity—exactly as was done for no-slip boundary conditions in the earlier sections of his text. As we will show, P.-L. Lions's energy inequality applies to Navier boundary conditions in general, which gives us the same existence and uniqueness theorem as for no-slip boundary conditions. (P.-L. Lions's comment on the regularity of $\partial_t u$ does not follow as in [16], though, because (4.18) of [16] is not valid for general Navier boundary conditions.) We include a proof of existence and uniqueness in section 6 that closely parallels the classical proofs due to Leray as they appear in [15] and [22]. In section 7, we extend the existence, uniqueness, regularity, and convergence results of [5] and [17] to multiply connected domains.

It is shown in [17] that if the initial vorticity is in $L^p(\Omega)$ for some $p > 2$, then after extracting a subsequence, solutions to the Navier–Stokes equations with Navier boundary conditions converge in $L^\infty([0,T]; L^2(\Omega))$ to a solution to the Euler equations (with the usual boundary condition of tangential velocity on the boundary) as $\nu \to 0$. This extends a result in [5] for initial vorticity in $L^\infty(\Omega)$, and because the solution to the Euler equations is unique in this case, it follows that the convergence is strong in $L^\infty([0,T]; L^2(\Omega))$—that is, does not require the extraction of a subsequence.

The convergence in [17] also generalizes the similar convergence established for the special case of Lions boundary conditions on page 131 of [16] (though not including the case $p = 2$). The main difficulty faced in making this generalization is establishing a bound on the L^p-norms of the vorticity, a task that is much easier for Lions boundary conditions (see pages 91–92 of [15] or page 131 of [16]). In contrast, nearly all of [5] and [17], including the structure of the existence proofs, is directed toward establishing an analogous bound.

The methods of proof in [5] and [17] do not yield a bound on the rate of convergence. With the assumptions in [17], such a bound is probably not possible. We can, however, make an assumption that is weaker than that of [5] but stronger than that of [17] and achieve a bound on the rate of convergence. Specifically, we assume, as in [14], that the L^p-norms of the initial vorticity grow sufficiently slowly with p (Definition 8.2) and establish the bound given in Theorem 8.4. To achieve this result, we also assume additional regularity on α and Γ.

The bound on the convergence rate in $L^\infty([0,T];L^2(\Omega))$ in Theorem 8.4 is the same as that obtained for $\Omega = \mathbb{R}^2$ in [14]. In particular, when α is nonnegative, it gives a bound on the rate of convergence for initial vorticity in $L^\infty(\Omega)$ that is proportional to

$$(\nu t)^{\frac{1}{2}} \exp\bigl(-C\|\omega^0\|_{L^2\cap L^\infty} t\bigr),$$

where C is a constant depending on Ω and α, and ω^0 is the initial vorticity. This is essentially the same bound on the convergence rate as that for $\Omega = \mathbb{R}^2$ appearing in [3].

Another interesting question is whether solutions to the Navier–Stokes equations with Navier boundary conditions converge to a solution to the Navier–Stokes equations with the usual no-slip boundary conditions if we let the function α grow large. We show in section 9 that such convergence does take place for initial velocity in $H^3(\Omega)$ and Γ in C^3 when we let α approach $+\infty$ uniformly on Γ. This type of convergence is, in a sense, an inverse of the derivation of the Navier boundary conditions from no-slip boundary conditions for rough boundaries discussed in [10] and [11].

In [13], Kato gives necessary and sufficient conditions for the vanishing viscosity limit of solutions to the Navier–Stokes equations with no-slip boundary conditions to converge to a strong solution to the Euler equations. In particular, he shows that the vanishing viscosity limit holds if and only if the L^2-norm of the gradient of the velocity in a boundary layer of width proportional to the viscosity vanishes sufficiently rapidly as the viscosity goes to zero. For Navier boundary conditions, it is easy to show that this norm on the boundary layer converges sufficiently rapidly, and because we have established the vanishing viscosity limit, it follows that Kato's conditions all hold, thus completing, in a sense, Kato's program for Navier boundary conditions. We describe this in more detail in section 10.

We follow the convention that C is always an unspecified constant that may vary from expression to expression, even across an inequality (but not across an equality). When we wish to emphasize that a constant depends, at least in part, upon the parameters $x_1, \dots, x_n$, we write $C(x_1, \dots, x_n)$. To distinguish between unspecified constants, we use C and C'.

For vectors u and v in $\mathbb{R}^2$, by $u \cdot \nabla v$ we mean the vector whose jth component is $u^i \partial_i v^j$. For 2×2 matrices A and B we define $A \cdot B = A^{ij} B^{ij}$, so $\nabla u \cdot \nabla v = \partial_j u^i \partial_j v^i$. Here, as everywhere in this paper, we follow the common convention that repeated indices are summed—whether or not one is a superscript and one is a subscript.

For the vector u and the scalar function ψ we define

$$u^\perp = (-u^2, u^1), \qquad \nabla^\perp \psi = (-\partial_2 \psi, \partial_1 \psi), \qquad \omega(u) = \partial_1 u^2 - \partial_2 u^1.$$

If X is a function space and k a positive integer, we define $(X)^k$ to be

$$\{(f_1, \dots, f_k) : f_1 \in X, \dots, f_k \in X\}.$$

For instance, $(H^1(\Omega))^2$ is the set of all vector fields, each of whose components lies in $H^1(\Omega)$. To avoid excess notation, however, we always suppress the superscript k when it is clear from the context whether we are dealing with scalar-, vector-, or matrix-valued functions.

We will make use of the following generalization of Gronwall's lemma. The succinct form of the proof is due to Tehranchi [21].

LEMMA 1.1 (Osgood's lemma). *Let L be a measurable nonnegative function and γ a nonnegative locally integrable function, each defined on the domain $[t_0, t_1]$. Let $\mu\colon [0,\infty) \to [0,\infty)$ be a continuous nondecreasing function, with $\mu(0) = 0$. Let $a \geq 0$, and assume that for all t in $[t_0, t_1]$,*

$$L(t) \leq a + \int_{t_0}^{t} \gamma(s)\mu(L(s))\, ds. \tag{1.3}$$

If $a > 0$, then

$$\int_a^{L(t)} \frac{ds}{\mu(s)} \leq \int_{t_0}^{t} \gamma(s)\, ds.$$

If $a = 0$ and $\int_0^\infty ds/\mu(s) = \infty$, then $L \equiv 0$.

Proof. We have

$$\begin{aligned}\int_a^{L(t)} \frac{dx}{\mu(x)} &\leq \int_a^{a+\int_{t_0}^t \gamma(u)\mu(L(u))\,du} \frac{dx}{\mu(x)} \\ &\leq \int_{t_0}^{t} \frac{\gamma(s)\mu(L(s))\,ds}{\mu(a+\int_{t_0}^s \gamma(u)\mu(L(u))\,du)} \leq \int_{t_0}^{t} \gamma(s)\,ds.\end{aligned}$$

The last inequality follows from (1.3), since μ is nondecreasing. ∎

We have stated Lemma 1.1 in the form that it appears on page 92 of [4]. This lemma is equivalent to a theorem of Bihari [2], though with an assumption only of measurability of μ rather than continuity; see, for example, Theorem 5.1 on pages 40–41 of [1].[1] An early form of the inequality appears in the work of Osgood [20], who assumes that $a = 0$, $\gamma \equiv 1$, and the bound is on $|L(t)|$ in (1.3); because of this, Lemma 1.1 is often referred to as Osgood's lemma. See also the historical discussion in section 2.14 of [6].

2. Function spaces. Let

$$E(\Omega) = \left\{ v \in (L^2(\Omega))^2 : \operatorname{div} v \in L^2(\Omega) \right\}, \tag{2.1}$$

as in [22], with the inner product

$$(u, v)_{E(\Omega)} = (u, v) + (\operatorname{div} u, \operatorname{div} v).$$

We will use the following theorem, which is Theorem 1.2 on page 7 of [22], several times.

LEMMA 2.1. *There exists a continuous linear operator $\gamma_{\mathbf{n}}$ mapping $E(\Omega)$ into $H^{-1/2}(\Gamma)$ such that*

$$\gamma_{\mathbf{n}} v = \text{ the restriction of } v \cdot \mathbf{n} \text{ to } \Gamma \text{ for every } v \text{ in } (\mathcal{D}(\overline{\Omega}))^2.$$

[1] The inequality in equation (5.2) of Theorem 5.1 of [1] should be $\leq$ instead of $\geq$.

Also, the following form of the divergence theorem holds for all vector fields v in $E(\Omega)$ and scalar functions h in $H^1(\Omega)$:

$$\int_\Omega v \cdot \nabla h + \int_\Omega (\operatorname{div} v) h = \int_\Gamma \gamma_{\mathbf{n}} v \cdot \gamma_0 h.$$

We always suppress the trace function γ_0 in our expressions, and we write $v \cdot \mathbf{n}$ in place of $\gamma_{\mathbf{n}} v$.

Define the following function spaces as in [5]:

$$\begin{aligned} H &= \left\{ v \in (L^2(\Omega))^2 : \operatorname{div} v = 0 \text{ in } \Omega \text{ and } v \cdot \mathbf{n} = 0 \text{ on } \Gamma \right\}, \\ V &= \left\{ v \in (H^1(\Omega))^2 : \operatorname{div} v = 0 \text{ in } \Omega \text{ and } v \cdot \mathbf{n} = 0 \text{ on } \Gamma \right\}, \\ \mathcal{W} &= \left\{ v \in V \cap H^2(\Omega) : v \text{ satisfies (1.2)} \right\}. \end{aligned} \tag{2.2}$$

We give $\mathcal{W}$ the H^2-norm, H the L^2-inner product and norm, which we symbolize by $(\cdot, \cdot)$ and $\|\cdot\|_{L^2(\Omega)}$, and V the H^1-inner product,

$$(u, v)_V = \sum_i (\partial_i u, \partial_i v),$$

and associated norm. This norm is equivalent to the H^1-norm, because Poincaré's inequality,

$$\|v\|_{L^p(\Omega)} \leq C(\Omega, p) \|\nabla v\|_{L^p(\Omega)} \tag{2.3}$$

for all p in $[1, \infty]$, holds for all v in V.

Ladyzhenskaya's inequality,

$$\|v\|_{L^4(\Omega)} \leq C(\Omega) \|v\|^{1/2}_{L^2(\Omega)} \|\nabla v\|^{1/2}_{L^2(\Omega)}, \tag{2.4}$$

also holds for all v in V, though the constant in the inequality is domain dependent, unlike the constant for the classical version of the space V.

We will also frequently use the following inequality, which follows from the standard trace theorem, Sobolev interpolation, and Poincaré's inequality:

$$\|v\|_{L^2(\Gamma)} \leq C(\Omega) \|v\|^{1/2}_{L^2(\Omega)} \|\nabla v\|^{1/2}_{L^2(\Omega)} \leq C(\Omega) \|v\|_V \tag{2.5}$$

for all v in V.

3. Hodge decomposition of H. Only simply connected domains are considered in [5] and [17]. To handle multiply connected domains we will need a portion of the Hodge decomposition of $L^2(\Omega)$. We briefly summarize the pertinent facts, drawing mostly from Appendix I of [22].

We assume that Ω is connected, for if it has multiple components we perform the decomposition separately on each component. Let $\Gamma_1, \ldots, \Gamma_{N+1}$ be the components of the boundary Γ with Γ_{N+1} bounding the unbounded component of Ω^C. Let $\Sigma_1, \ldots, \Sigma_N$ be one-manifolds with boundary that generate $H_1(\Omega, \Gamma; \mathbb{R})$, the one-dimensional real homology class of Ω relative to its boundary Γ.

We can decompose the space H into two subspaces, $H = H_0 \oplus H_c$, where

$$\begin{aligned} H_0 &= \{ v \in H : \text{all internal fluxes are zero} \}, \\ H_c &= \{ v \in H : \omega(v) = 0 \}. \end{aligned}$$

An internal flux is a value of $\int_{\Sigma_i} v \cdot \mathbf{n}$. Then $H_0 = H_c^\perp$.

Define ψ_i, $i = 1, \ldots, N$, to be the solution to $\Delta\psi_i = 0$ on Ω, $\psi_i = C_i$ on Γ_i, and $\psi_i = 0$ on all other components of Γ, where C_i is a nonzero constant. By elliptic regularity, ψ_i is in $H^2(\Omega)$ (apply, for instance, Theorem 8.12 on page 176 of [8]). Thus, $h_i := \nabla^\perp \psi_i$ is in $H^1(\Omega)$ and is divergence-free since $\operatorname{div} \nabla^\perp = 0$, and $h_i \cdot \mathbf{n} = 0$ since ψ_i is locally constant along Γ; that is, h_i is in V. The vectors $(h_1, \ldots, h_N)$ form an orthogonal basis for $H_c \subseteq V$, which we can assume is orthonormal by choosing (C_i) appropriately.

If v is in V then v is also in H so there exist a unique u in H_0 and h in H_c such that $v = u + h$; also, $(u, h)_H = 0$. But h is in V; hence, u also lies in V. This shows that $V = (V \cap H_0) \oplus H_c$, though this is not an orthogonal decomposition of V.

Given v in H we construct an associated stream function ψ in $H^1(\Omega)$ as follows. Fix a point a on $\partial\Omega$. For any x in $\overline{\Omega}$, and let γ be a smooth curve in $\overline{\Omega}$ from a to x. Along the curve γ let $\boldsymbol{\tau}$ be a unit tangent vector in the direction of γ and $\mathbf{n}$ be the unit normal vector for which $(\mathbf{n}, \boldsymbol{\tau})$ gives the standard orientation to $\mathbb{R}^2$. Then one can show that the function ψ defined by

$$\psi(x) = -\int_\gamma v \cdot \mathbf{n}\, ds \tag{3.1}$$

is independent of the choice of γ and of the set of generators, and that $v = \nabla^\perp \psi$. (The salient fact is that $v \cdot \mathbf{n}$ integrates to zero along any generator of the first (nonrelative) homology because $v \cdot \mathbf{n} = 0$ along $\partial\Omega$.)

On the boundary component containing a, ψ is zero, because $v \cdot \mathbf{n} = 0$ on Γ. On the other boundary components, ψ is constant, because the internal fluxes are independent of the path. In the special case where v is in H_0, all the internal fluxes are zero, so ψ is zero on all of Γ. From the way that we defined the basis (h_k) for H_c, it is clear that the projection into H_c of a vector lying in H is uniquely determined by the value of its stream function on the boundary.

The following is due to Yudovich.

LEMMA 3.1. *For any p in $[2, \infty)$ and any v in H_0 with $\omega(v)$ in $L^p(\Omega)$,*

$$\|\nabla v\|_{L^p(\Omega)} \le C(\Omega) p \, \|\omega(v)\|_{L^p(\Omega)} .$$

Proof. Let v be in H_0 with $\omega(v)$ in $L^p(\Omega)$. Then, as noted above, the associated stream function ψ vanishes on Γ. Applying Corollary 1 of [24] with the operator $L = \Delta$ and $r = 0$ gives

$$\|\nabla v\|_{L^p(\Omega)} \le \|\psi\|_{H^{2,p}(\Omega)} \le C(\Omega) p \, \|\Delta\psi\|_{L^p(\Omega)} = C(\Omega) p \, \|\omega(v)\|_{L^p(\Omega)} . \qquad \square$$

For Ω simply connected, $H = H_0$, and Lemma 3.1 applies to all of H. The critical feature of Lemma 3.1 is that the dependence of the inequality on p is made explicit, a fact we will exploit in the proof of Theorem 8.4.

With the assumption of additional regularity on Γ, we have the following result for velocity fields in H.

COROLLARY 3.2. *Assume that Γ is $C^{2,\epsilon}$ for some $\epsilon > 0$. Then for any p in $[2, \infty)$ and any v in H with $\omega(v)$ in $L^p(\Omega)$,*

$$\|\nabla v\|_{L^p(\Omega)} \le C(\Omega) p \, \|\omega(v)\|_{L^p(\Omega)} + C'(\Omega) \, \|v\|_{L^2(\Omega)} ,$$

the constants $C(\Omega)$ and $C'(\Omega)$ being independent of p.

Proof. Because Γ is $C^{2,\epsilon}$, it follows from elliptic regularity theory that each ψ_i is in $C^{2,\epsilon}(\overline{\Omega})$ (apply, for instance, Theorem 6.14 on page 101 of [8]). Thus, each basis element $h_i = \nabla^\perp \psi_i$ for H_c is in $C^{1,\epsilon}(\overline{\Omega})$ and so ∇h_i is in $L^\infty(\Omega)$.

Let v be in H with $\omega(v)$ in $L^p(\Omega)$, and let $v = u + h$, where u is in H_0 and h is in H_c. Let $h = \sum_{i=1}^N c_i h_i$ and $r = \|h\|_{L^2(\Omega)} = (\sum_i c_i^2)^{1/2}$. Then

$$\begin{aligned}\|\nabla h\|_{L^p(\Omega)} &\le \sum_{i=1}^N |c_i| \, \|\nabla h_i\|_{L^p(\Omega)} \le \sum_{i=1}^N r|\Omega|^{1/p} \, \|\nabla h_i\|_{L^\infty(\Omega)} \\ &\le r \max\left\{1, |\Omega|^{1/2}\right\} \sum_{i=1}^N \|\nabla h_i\|_{L^\infty(\Omega)} \le C \, \|h\|_{L^2(\Omega)} \, .\end{aligned}$$

But $H_0 = H_c^\perp$, so $\|v\|^2_{L^2(\Omega)} = \|u\|^2_{L^2(\Omega)} + \|h\|^2_{L^2(\Omega)}$, and thus $\|h\|_{L^2(\Omega)} \le \|v\|_{L^2(\Omega)}$. Therefore,

$$\begin{aligned}\|\nabla v\|_{L^p(\Omega)} &\le \|\nabla u\|_{L^p(\Omega)} + \|\nabla h\|_{L^p(\Omega)} \\ &\le C(\Omega) p \, \|\omega(v)\|_{L^p(\Omega)} + C'(\Omega) \, \|v\|_{L^2(\Omega)}\end{aligned}$$

by virtue of Lemma 3.1. ∎

4. Vorticity on the boundary. Let κ be the curvature of Γ. Then κ is continuous because Γ is C^2, and if we parameterize each component of Γ by arc length, s, it follows that

$$\frac{\partial \mathbf{n}}{\partial \boldsymbol{\tau}} := \frac{d\mathbf{n}}{ds} = \kappa \boldsymbol{\tau}.$$

LEMMA 4.1. *If u and v are in $(H^2(\Omega))^2$ with $u \cdot \mathbf{n} = v \cdot \mathbf{n} = 0$ on Γ, then*

$$(v \cdot \nabla u) \cdot \mathbf{n} = -\kappa u \cdot v, \tag{4.1}$$

$$(\mathbf{n} \cdot \nabla v) \cdot \boldsymbol{\tau} = \omega(v) + (\boldsymbol{\tau} \cdot \nabla v) \cdot \mathbf{n} = \omega(v) - \kappa v \cdot \boldsymbol{\tau}, \tag{4.2}$$

$$(\mathbf{n} \cdot D(v)) \cdot \boldsymbol{\tau} = \frac{1}{2}\omega(v) - \kappa v \cdot \boldsymbol{\tau}. \tag{4.3}$$

Proof. Because $u \cdot \mathbf{n}$ has a constant value (of zero) along Γ,

$$0 = \frac{\partial}{\partial \boldsymbol{\tau}}(u \cdot \mathbf{n}) = \frac{\partial u}{\partial \boldsymbol{\tau}} \cdot \mathbf{n} + u \cdot \frac{\partial \mathbf{n}}{\partial \boldsymbol{\tau}} = (\boldsymbol{\tau} \cdot \nabla u) \cdot \mathbf{n} + \kappa u \cdot \boldsymbol{\tau},$$

so $(\boldsymbol{\tau} \cdot \nabla u) \cdot \mathbf{n} = -\kappa u \cdot \boldsymbol{\tau}$. But v is parallel to $\boldsymbol{\tau}$, so (4.1) follows by linearity. The identity in (4.3) is Lemma 2.1 of [5], and (4.2) is established similarly. ∎

COROLLARY 4.2. *A vector v in $V \cap H^2(\Omega)$ satisfies Navier boundary conditions (that is, lies in $\mathcal{W}$) if and only if*

$$\omega(v) = (2\kappa - \alpha) v \cdot \boldsymbol{\tau} \text{ and } v \cdot \mathbf{n} = 0 \text{ on } \Gamma. \tag{4.4}$$

Also, for all v in $\mathcal{W}$ and u in V,

$$(\mathbf{n} \cdot \nabla v) \cdot u = (\kappa - \alpha) v \cdot u \text{ on } \Gamma. \tag{4.5}$$

Proof. Let v be in $V \cap H^2(\Omega)$. Then from (4.3),

$$2(\mathbf{n} \cdot D(v)) \cdot \boldsymbol{\tau} + 2\kappa(v \cdot \boldsymbol{\tau}) = \omega(v). \tag{4.6}$$

If v satisfies Navier boundary conditions, then (4.4) follows by subtracting $2(\mathbf{n}\cdot D(v))\cdot \boldsymbol{\tau} + \alpha v \cdot \boldsymbol{\tau} = 0$ from (4.6). Conversely, substituting the expression for $\omega(v)$ in (4.4) into (4.6) gives $2(\mathbf{n} \cdot D(v)) \cdot \boldsymbol{\tau} + \alpha v \cdot \boldsymbol{\tau} = 0$.

If v is in $\mathcal{W}$, then from (4.2),

$$(\mathbf{n} \cdot \nabla v) \cdot \boldsymbol{\tau} = \omega(v) - \kappa v \cdot \boldsymbol{\tau} = (2\kappa - \alpha)v \cdot \boldsymbol{\tau} - \kappa v \cdot \boldsymbol{\tau} = (\kappa - \alpha)v \cdot \boldsymbol{\tau},$$

and (4.5) follows from this, since u is parallel to τ on Γ. ∎

COROLLARY 4.3. *For initial velocity in $H^2(\Omega)$, Lions boundary conditions are the special case of Navier boundary conditions in which*

$$\alpha = 2\kappa.$$

That is, any solution of (NS) with Navier boundary conditions where $\alpha = 2\kappa$ is also a solution to (NS) with Lions boundary conditions.

5. Weak formulations. We give two equivalent formulations of a weak solution to the Navier–Stokes equations with Navier boundary conditions, in analogy with Problems 3.1 and 3.2 on pages 190–191 of [22].

For all u in $\mathcal{W}$ and v in V,

$$\begin{aligned} \int_\Omega \Delta u \cdot v &= \int_\Omega (\operatorname{div} \nabla u^i) v^i = \int_\Gamma (\nabla u^i \cdot \mathbf{n}) v^i - \int_\Omega \nabla u^i \cdot \nabla v^i \\ &= \int_\Gamma (\mathbf{n} \cdot \nabla u) \cdot v - \int_\Omega \nabla u \cdot \nabla v = \int_\Gamma (\kappa - \alpha) u \cdot v - \int_\Omega \nabla u \cdot \nabla v, \end{aligned} \tag{5.1}$$

where we used (4.5) of Corollary 4.2. This motivates our first formulation of a weak solution.

DEFINITION 5.1. *Given viscosity $\nu > 0$ and initial velocity u^0 in H, u in $L^2([0,T];V)$ is a weak solution to the Navier–Stokes equations (without forcing) if $u(0) = u^0$ and*

$$(NS) \qquad \frac{d}{dt}\int_\Omega u \cdot v + \int_\Omega (u \cdot \nabla u) \cdot v + \nu \int_\Omega \nabla u \cdot \nabla v - \nu \int_\Gamma (\kappa - \alpha) u \cdot v = 0$$

for all v in V. (We make sense of the initial condition $u(0) = u^0$ as in [22].*)*

This formulation of a weak solution is equivalent to that in (2.11) and (2.12) of [5]. This follows from the identity

$$2\int_\Omega D(u) \cdot D(v) = \int_\Omega \nabla u \cdot \nabla v - \int_\Gamma \kappa u \cdot v,$$

which holds for all u and v in V. To establish this identity, let u and v be in $V \cap H^2(\Omega)$ and observe that $2D(u) \cdot D(v) = \nabla u \cdot \nabla v + \nabla u \cdot (\nabla v)^T$. Also, because u is divergence-free, $\nabla u \cdot (\nabla v)^T = \partial_i u^j \partial_j v^i = \partial_j(\partial_i u^j v^i) = \operatorname{div}(v \cdot \nabla u)$. Then, using (4.1), the identity follows from

$$\int_\Omega \nabla u \cdot (\nabla v)^T = \int_\Omega \operatorname{div}(v \cdot \nabla u) = \int_\Gamma (v \cdot \nabla u) \cdot \mathbf{n} = -\int_\Gamma \kappa u \cdot v$$

and the density of $H^2(\Omega) \cap V$ in V.

Our second formulation of a weak solution will be identical to that of Problem 3.2 on page 191 of [22], except that the operator A of [22] will also include the boundary integral of (5.1). Accordingly, we define the operators A and B by

$$(Au, v)_{V,V'} = \int_\Omega \nabla u \cdot \nabla v - \int_\Gamma (\kappa - \alpha) u \cdot v,$$
$$(Bu, v)_{V,V'} = \int_\Omega (u \cdot \nabla u) \cdot v$$

for all u and v in V.

By (2.5),

$$|(Au, v)_{V,V'}| \leq \|u\|_V \|v\|_V + C \|u\|_{L^2(\Gamma)} \|v\|_{L^2(\Gamma)} \leq C \|u\|_V \|v\|_V . \tag{5.2}$$

Thus, $A : L^2([0,T]; V) \to L^2([0,T]; V')$, and, as it does for the classical version of the space V (for which vectors are zero on Γ), $B : L^2([0,T]; V) \to L^1([0,T]; V')$. Thus, if u is a solution as in Definition 5.1, then $-\nu Au - Bu$ lies in $L^1([0,T]; V')$ and

$$\frac{d}{dt} \langle u, v \rangle = (-\nu Au - Bu, v)_{V,V'}$$

for all v in V. It follows from Lemma 1.1 on page 169 of [22] that u is in $C([0,T]; H)$. This not only makes sense of the initial condition $u(0) = u^0$ but also shows that the following formulation of a weak solution is equivalent to that of Definition 5.1.

DEFINITION 5.2. *Given viscosity $\nu > 0$ and initial velocity u^0 in H, u in $L^2([0,T]; V)$ is a weak solution to the Navier–Stokes equations if $u(0) = u^0$ and*

$$\begin{cases} u' \in L^1([0,T]; V'), \\ u' + \nu Au + Bu = 0 \text{ on } (0,T), \\ u(0) = u^0, \end{cases}$$

where $u' := \partial_t u$.

From here on we will refer to either of the formulations in Definitions 5.1 and 5.2 as (NS).

6. Existence and uniqueness. We can obtain existence and uniqueness of a solution to (NS) assuming only that the initial velocity is in H.

THEOREM 6.1. *Assume that Γ is C^2 and α is in $L^\infty(\Gamma)$. Let u^0 be in H and let $T > 0$. Then there exists a solution u to (NS). Moreover, u is in $L^2([0,T]; V) \cap C([0,T]; H)$, u' is in $L^2([0,T]; V')$, and we have the energy inequality*

$$\|u(t)\|_{L^2(\Omega)} \leq e^{C(\alpha)\nu t} \|u^0\|_{L^2(\Omega)}. \tag{6.1}$$

The constant $C(\alpha)$ is zero if α is nonnegative on Γ.

Sketch of proof. Existence of a solution to (NS) proceeds as in the first proof of existence in [15, pp. 75–77], though using the analogue of the energy inequality on page 130 of [16]. Using (4.5), we have, formally,

$$\frac{1}{2} \frac{d}{dt} \|u\|^2_{L^2(\Omega)} + \nu \|\nabla u\|^2_{L^2(\Omega)} = \nu \int_\Omega (\mathbf{n} \cdot \nabla u) \cdot u \leq C\nu \|u\|^2_{L^2(\Gamma)} , \tag{6.2}$$

where $C = \sup_\Gamma |\kappa - \alpha|$. Arguing exactly as in [16], it follows that

$$\frac{d}{dt}\|u\|^2_{L^2(\Omega)} + \nu\|\nabla u\|^2_{L^2(\Omega)} \le C\nu\|u\|^2_{L^2(\Omega)}.$$

Integrating over time gives

$$\begin{aligned}\|u(t)\|^2_{L^2(\Omega)} + \nu\int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)}\,ds \\ \le \|u^0\|^2_{L^2(\Omega)} + C\nu\int_0^t \|u(s)\|^2_{L^2(\Omega)}\,ds.\end{aligned} \tag{6.3}$$

The energy bound, (6.1), then follows from Gronwall's lemma. (If α is nonnegative, then, in fact, energy is nonincreasing—in the absence of forcing—so $C(\alpha) = 0$. This follows from the equation preceding (2.16) of [5].)

The proofs of regularity in time and space and of uniqueness proceed exactly as in the proof of Theorem 3.2 on page 199 of [22], though in the proof of uniqueness we must account for the presence of the boundary integral in (NS). ∎

7. Additional regularity. If we assume extra regularity on the initial velocity, that regularity will be maintained for all time. This is crucial for establishing the vanishing viscosity limit in section 8, where we must impose stronger regularity on the initial velocity to obtain existence of a solution to the Euler equations.

THEOREM 7.1. *Assume that* Γ *is* $C^{2,1/2+\epsilon}$ *and that* α *is in* $H^{1/2+\epsilon}(\Gamma)+C^{1/2+\epsilon}(\Gamma)$ *for some* $\epsilon > 0$. *Let* u^0 *be in* $\mathcal{W}$ *with initial vorticity* ω^0, *and let* u *be the unique solution to* (NS) *given by Theorem* 6.1 *with corresponding vorticity* ω. *Let* $T > 0$. *Then*

$$u' \in L^2([0,T];V) \cap C([0,T];H).$$

If, in addition, ω^0 *is in* $L^\infty(\Omega)$ *(so* u^0 *is compatible), then*

$$u \in C([0,T];H^2(\Omega)),\ \omega \in C([0,T];H^1(\Omega)) \cap L^\infty([0,T]\times\Omega).$$

Proof. Regularity of u'. We prove that u' lies in $L^2([0,T];V) \cap L^\infty([0,T];H)$ in three steps as in the proof of Theorem 3.5 on pages 202–204 of [22]. In this proof, Temam uses a Galerkin approximation sequence (u_m) to the solution u. We employ the same sequence, though using the basis of Corollary A.3 rather than that of [22]; this is the only change to step (i).

No change to step (ii) of Temam's proof is required, because the bound on $\|u_m'(0)\|^2_{L^2}$ in (3.88) of [22], which does not involve boundary integrals, still holds.

In step (iii), an additional term of

$$\nu\int_\Gamma (\kappa - \alpha)|u_m'|^2$$

appears on the right side of (3.94) of Temam's proof, which we bound by

$$C\nu\|u_m'\|_{L^2(\Omega)}\|\nabla u_m'\|_{L^2(\Omega)} \le \frac{\nu}{2}\|\nabla u_m'\|^2_{L^2(\Omega)} + C\nu\|u_m'\|^2_{L^2(\Omega)}.$$

Then (3.95) of Temam's proof becomes

$$\frac{d}{dt}\|u_m'(t)\|^2_{L^2(\Omega)} \le \phi_m(t)\|u_m'(t)\|^2_{L^2(\Omega)},$$

where

$$\phi_m(t) = \left(\frac{2}{\nu} + C\nu\right) \|u_m(t)\|^2_{L^2(\Omega)},$$

and the proof that u' lies in $L^2([0,T];V) \cap L^\infty([0,T];H)$ is completed as in [22].

Regularity of u and ω. To establish the regularity of u and ω we follow the bootstrap argument in the second half of the proof of Theorem 2.3 in [5]. For completeness, we give a full account of this argument here, adapting it to multiply connected domains.

Because u' is in $L^2([0,T];V)$ we can argue as in the paragraph preceding Definition 5.2, with u' playing the role of u, that u' is in $C([0,T];H)$. The membership of u' in $L^2([0,T];V)$ also gives u in $H^1([0,T];V)$; by one-dimensional Sobolev embedding, u is then in $C^{1/2}([0,T];V)$ and hence in $C([0,T];V)$. It follows that $u \cdot \nabla u$ is in $C([0,T];L^q(\Omega))$ for all q in $[1,2)$ (see, for instance, Theorem 1.4.4.2 on page 28 of [9]).

Now let $\beta > 0$ and let $\Phi := -u \cdot \nabla u - u' + \beta u$. Then Φ is in $C([0,T];L^q(\Omega))$ for all q in $[1,2)$ by our observations above. Because of the additional regularity we have imposed on Γ and on α over that assumed in Theorem 6.1, $g := (2\kappa - \alpha)u \cdot \boldsymbol{\tau}$ lies in $C([0,T];H^{1/2}(\Gamma))$ (see, for instance, Theorem 1.4.1.1 on page 21 and Theorem 1.4.4.2 on page 28 of [9]).

Let $a : V \times V \to \mathbb{R}$ be defined by $a(u,v) = (Au + \beta u, v)_{V,V'}$ and require that $\beta > 0$ be sufficiently large that $a(v,v) \geq \|v\|^2_V$ for all v in V. This is possible because α is in $L^\infty(\Gamma) \subseteq H^{1/2+\epsilon}(\Gamma) + C^{1/2+\epsilon}(\Gamma)$ by one-dimensional Sobolev embedding. Also by (5.2) we see that $|a(u,v)| \leq C \|u\|_V \|v\|_V$. Applying the Lax–Millgram lemma, we find that there exists a unique w in V such that $a(w,v) = \Phi$ for all v in V. By Definition 5.2, $w = u(t)$ for all t in $[0,T]$. By Definition 5.1, $u(t)$ is also the unique variational solution at time t in $[0,T]$ to Stokes's problem,

$$\begin{cases} -\nu\Delta u + \nabla p + \beta u = \Phi & \text{in } \Omega, \\ \operatorname{div} u = 0 & \text{in } \Omega, \\ \omega(u) = g & \text{on } \Gamma. \end{cases}$$

Formally, the vorticity formulation of the above system is

$$\begin{cases} -\nu\Delta\omega + \beta\omega = \omega(\Phi) & \text{in } \Omega, \\ \omega = g & \text{on } \Gamma. \end{cases}$$

Because $\omega(\Phi)$ is in $C([0,T];H^{-1,q}(\Omega))$, standard elliptic theory gives a unique solution ω in $C([0,T];H^{1,q}(\Omega))$. Because u is in $C([0,T];H)$, there exists an associated stream function φ in $C([0,T] \times \Omega)$ that is constant on each component of Γ at time t in $[0,T]$; this follows directly from (3.1). Letting ψ be the unique solution to

$$\begin{cases} \Delta\psi = \omega & \text{in } \Omega, \\ \psi = \varphi & \text{on } \Gamma, \end{cases}$$

it follows that $u = \nabla^\perp\psi$, because u and $\nabla^\perp\psi$ have the same vorticity ($\operatorname{curl}\nabla^\perp = \Delta$) and their stream functions (namely, φ and ψ) share the same value on Γ.

By Theorem 2.5.1.1 on page 128 of [9], ψ is in $C([0,T];H^{3,q}(\Omega))$; u, then, is in $C([0,T];H^{2,q}(\Omega))$ and hence in $C([0,T] \times \Omega)$ by Sobolev embedding. But u is in $C([0,T];V)$, so $u \cdot \nabla u$ and also Φ are in $C([0,T];L^2(\Omega))$. Passing through the same argument again, this time with $q = 2$, gives u in $C([0,T];(H^2(\Omega))^2)$. □

With Theorem 7.1, we have a replacement for Theorem 2.3 of [5] that applies regardless of the sign of α. Since the nonnegativity of α is used nowhere else in [5] and [17], all the results of both of those papers apply to simply connected domains as well regardless of the sign of α—with the regularity we have assumed on Γ and α.

To remove the restriction on the domain being simply connected, it remains only to show that Lemmas 2 and 3 of [17] remain valid for multiply connected domains. We show this for Lemma 2 of [17] in Theorem A.2. Because, however, for multiply connected domains there is no longer a unique vector field in $\mathcal{W}$ with a given vorticity, we must define a *vector field* to be compatible, rather than its *vorticity*, as was done in [17].

DEFINITION 7.2. *A vector field v in $\mathcal{W}$ is called compatible if $\omega(v)$ is in $L^\infty(\Omega)$.*

As for Lemma 3 of [17], we need only use Corollary 3.2 to replace the term $\|\omega(\cdot,t)\|_{L^p(\Omega)}^{1-\theta}$ with $(\|\omega(\cdot,t)\|_{L^p(\Omega)}+\|u(\cdot,t)\|_{L^2(\Omega)})^{1-\theta}$ in the proof of Lemma 3 in [17]. Lemma 3 of [17] then follows with no other changes in the proof—only the value of the constant C changes.

We thus have the following theorem, which is only a slight modification of Proposition 1 of [17], and which applies to multiply connected domains and unsigned α.

THEOREM 7.3. *Assume that Γ and α are as in Theorem 7.1. Let q be in $(2,\infty]$, and assume that u^0 is in V with initial vorticity ω^0 in $L^p(\Omega)$ for some p in $[q,\infty]$. Let $T>0$. Then there exists a unique solution u to (NS) with corresponding vorticity ω, and for all p in $[q,\infty]$,*

$$\|\omega(t)\|_{L^p} \le \|\omega^0\|_{L^p} + C_0 \tag{7.1}$$

a.e. in $[0,T]$. The constant C_0, which is independent of p, is given by

$$C_0 = C(T,\alpha,\kappa,q)e^{C(\alpha)\nu T}\max\{|\Omega|^{1/2},1\}\left(\|u^0\|_{L^2(\Omega)}+\|\omega^0\|_{L^q(\Omega)}\right).$$

Also, u is in $L^\infty([0,T];C(\overline{\Omega}))\cap L^\infty([0,T];V)$, the norm of u in this space being bounded over any finite range of viscosity ν.

Proof. Approximate u^0 by a sequence of compatible vector fields via Theorem A.2, and let u_n be the corresponding solutions to (NS) given by Theorem 7.1. The argument in the proof of Lemma 3 of [17] can be used to bound $\Lambda = \|(2\kappa-\alpha)u_n\cdot\boldsymbol{\tau}\|_{L^\infty(\Omega)}$ in terms of $\|\omega^0\|_{L^q(\Omega)}$, and this in turn gives the bound $\|\omega_n(t)\|_{L^p}\le\|\omega^0\|_{L^p}+C_0$. This bound holds for the solution u in the limit, as in the proof of Proposition 1 in [17]. (The constant $C(T,\alpha,\kappa,q)$ approaches infinity as q approaches 2, so it is not possible to extend this result to $p=2$.)

Finally, using Sobolev interpolation, (2.3), and Corollary 3.2,

$$\begin{aligned}\|u(t)\|_{C(\overline{\Omega})} &\le C\,\|u(t)\|_{L^2(\Omega)}^{\theta}\,\|u(t)\|_{H^{1,q}(\Omega)}^{1-\theta}\\ &\le C\,\|u(t)\|_{L^2(\Omega)}^{\theta}\,(\|\omega(t)\|_{L^q(\Omega)}+\|u(t)\|_{L^2(\Omega)})^{1-\theta},\end{aligned} \tag{7.2}$$

where $\theta=(q-2)/(2q-2)$. This norm is finite by (6.1), so u is also in $L^\infty([0,T];C(\overline{\Omega}))$ and its norm is uniformly bounded over any finite range of viscosity, as is its norm in $L^\infty([0,T];V)$. Explicitly,

$$\begin{aligned}\|u\|_{L^\infty([0,T];V)} &= \|\nabla u\|_{L^\infty([0,T];L^2(\Omega))} \le C\,\|\nabla u\|_{L^\infty([0,T];L^q(\Omega))}\\ &\le C(\|\omega\|_{L^\infty([0,T];L^q(\Omega))}+\|u\|_{L^\infty([0,T];L^2(\Omega))})\\ &\le C(T,\alpha,\kappa)e^{C(\alpha)\nu T},\end{aligned} \tag{7.3}$$

a bound we will use in section 8. In the second inequality above we used Corollary 3.2. ∎

8. Vanishing viscosity. In this section we bound the rate of convergence in $L^\infty([0,T];L^2(\mathbb{R}^2))$ of solutions to (NS) to the unique solution to the Euler equations for the class of (bounded or unbounded) *Yudovich vorticities*. To describe Yudovich vorticity, we need the following definition.

DEFINITION 8.1. *Let $\theta : [p_0,\infty) \to \mathbb{R}$ for some $p_0 > 1$. We say that θ is admissible if the function $\beta : (0,\infty) \to [0,\infty)$, defined for some $M > 0$ by*[2]

$$\beta(x) := \beta_M(x) := x \inf\left\{(M^\epsilon x^{-\epsilon}/\epsilon)\theta(1/\epsilon) : \epsilon \text{ in } (0, 1/p_0]\right\}, \tag{8.1}$$

satisfies

$$\int_0^1 \frac{dx}{\beta(x)} = \infty. \tag{8.2}$$

Because $\beta_M(x) = M\beta_1(x/M)$, this definition is independent of the value of M, though the presence of M in the definition will turn out to be convenient. Also, β is a monotonically increasing continuous function, with $\lim_{x\to 0^+}\beta(x) = 0$.

Yudovich proves in [26] that for a bounded domain in $\mathbb{R}^n$, if $\|\omega^0\|_{L^p} \le \theta(p)$ for some admissible function θ, then at most one solution to the Euler equations exists. Because of this, we call the class of all such vorticities *Yudovich vorticity*.

DEFINITION 8.2. *We say that a vector field v has Yudovich vorticity if $p \mapsto \|\omega(v)\|_{L^p(\Omega)}$ is an admissible function.*

Examples of admissible bounds on vorticity are

$$\theta_0(p) = 1, \theta_1(p) = \log p, \dots, \theta_m(p) = \log p \cdot \log\log p \cdots \log^m p, \tag{8.3}$$

where $\log^m$ is log composed with itself m times. These admissible bounds are described in [26] (see also [14].) Roughly speaking, the L^p-norm of a Yudovich vorticity can grow in p only slightly faster than $\log p$ and still be admissible. Such growth in the L^p-norm arises, for example, from a point singularity of the type $\log\log(1/|x|)$.

DEFINITION 8.3. *Given an initial velocity u^0 in V, $\overline{u}$ in $L^2([0,T];V)$ is a weak solution to the Euler equations if $\overline{u}(0) = u^0$ and*

$$\frac{d}{dt}\int_\Omega \overline{u}\cdot v + \int_\Omega (\overline{u}\cdot\nabla\overline{u})\cdot v = 0$$

for all v in V.

The existence of a weak solution to the Euler equations under the assumption that the initial vorticity ω^0 is in $L^p(\Omega)$ for some $p > 1$ (a weaker assumption than that of Definition 8.3 when $1 < p < 2$) was proved in [25]. By the result in [26] mentioned above, the solutions are unique in the class of all such solutions $\overline{u}$ for which $\omega(\overline{u})$ and $\overline{u}'$ lie in $L^\infty_{loc}(\mathbb{R};L^p(\Omega))$ for all p in an interval $[p_0,\infty)$.

THEOREM 8.4. *Assume that Γ and α are as in Theorem 7.1. Fix $T > 0$, let u^0 be in V, and assume that ω^0 is in $L^p(\mathbb{R}^2)$ for all p in $[2,\infty)$, with $\|\omega^0\|_{L^p} \le \theta(p)$ for some admissible function θ. Let u be the solution to (NS) for $\nu > 0$ given by Theorem 7.3 and $\overline{u}$ be the unique weak solution to the Euler equations for which $\omega(\overline{u})$ and $\overline{u}'$ are in $L^\infty_{loc}(\mathbb{R};L^p(\Omega))$ for all p in $[2,\infty)$, $\overline{u}$ and u both having initial velocity u^0. Then*

$$u(t) \to \overline{u}(t) \text{ in } L^\infty([0,T];L^2(\Omega)\cap L^2(\Gamma)) \text{ as } \nu \to 0.$$

[2] The definition of β in (8.1) differs from that in [14] in that it directly incorporates the factor of p that appears in the Calderón–Zygmund inequality; in [14] this factor is included in the equivalent of (8.2).

Also, there exists a constant $R = C(T, \alpha, \kappa)$, such that if we define the function $f : [0, \infty) \to [0, \infty)$ by

$$\int_{R\nu}^{f(\nu)} \frac{dr}{\beta(r)} = CT,$$

where β is defined as in (8.1)*, then*

$$\begin{aligned} &\|u - \overline{u}\|_{L^\infty([0,T];L^2(\Omega))} \le f(\nu)^{1/2} \text{ and} \\ &\|u - \overline{u}\|_{L^\infty([0,T];L^2(\Gamma))} \le C'(T, \alpha, \kappa) f(\nu)^{1/4} \end{aligned} \tag{8.4}$$

for all ν in $(0, 1]$.

Proof. We let $w = u - \overline{u}$. It is possible to show that integral identity in Definition 5.1 extends to any v in $L^2([0, T]; V)$ in the form

$$\int_\Omega \partial_t u \cdot v + \int_\Omega (u \cdot \nabla u) \cdot v + \nu \int_\Omega \nabla u \cdot \nabla v - \nu \int_\Gamma (\kappa - \alpha) u \cdot v = 0$$

with a similar extension for the identity in Definition 8.3. Applying these identities with $v = w$ and subtracting give

$$\begin{aligned} &\int_\Omega w \cdot \partial_t w + \int_\Omega w \cdot (u \cdot \nabla w) + \int_\Omega w \cdot (w \cdot \nabla \overline{u}) \\ &\quad = \nu \int_\Gamma (\kappa - \alpha) u \cdot w - \nu \int_\Omega \nabla u \cdot \nabla w. \end{aligned} \tag{8.5}$$

Both $\partial_t u$ and $\partial_t \overline{u}$ are in $L^2([0, T]; V')$, so (see, for instance, Lemma 1.2 on page 176 of [22])

$$\int_\Omega w \cdot \partial_t w = \frac{1}{2} \frac{d}{dt} \|w\|^2_{L^2(\Omega)}.$$

Applying Lemma 2.1,

$$\begin{aligned} &\int_\Omega w \cdot (u \cdot \nabla w) \\ &\quad = \int_\Omega w^i u^j \partial_j w^i = \frac{1}{2} \int_\Omega u^j \partial_j \sum_i (w^i)^2 = \frac{1}{2} \int_\Omega u \cdot \nabla |w|^2 \\ &\quad = \frac{1}{2} \int_\Gamma (u \cdot \mathbf{n}) |w|^2 - \frac{1}{2} \int_\Omega (\operatorname{div} u) |w|^2 = 0, \end{aligned}$$

since $u \cdot \mathbf{n} = 0$ on Γ and $\operatorname{div} u = 0$ in Ω. Thus, integrating (8.5) over time,

$$\|w(t)\|^2_{L^2(\Omega)} \le K + 2 \int_0^t \int_\Omega |w|^2 |\nabla \overline{u}|, \tag{8.6}$$

where

$$\begin{aligned} K &= 2\nu \int_0^t \left[\int_\Gamma (\kappa - \alpha) u \cdot w - \int_\Omega \nabla u \cdot \nabla w \right] \\ &\le 2\nu \int_0^t \left[\int_\Gamma (\kappa - \alpha) u \cdot w + \int_\Omega \nabla u \cdot \nabla \overline{u} \right]. \end{aligned}$$

Applying (2.5) and then using (7.3) and its equivalent for solutions to the Euler equations (where the constant does not increase with time), we have

$$\begin{aligned}\left|\int_\Gamma (\kappa-\alpha)u\cdot w\right| &\le \|\kappa-\alpha\|_{L^\infty(\Gamma)}\,\|u\|_{L^2(\Gamma)}\,\|w\|_{L^2(\Gamma)}\\ &\le C\,\|u\|_V\,\|w\|_V \le C(T,\alpha,\kappa)e^{C(\alpha)\nu T}.\end{aligned}\tag{8.7}$$

By (7.3) we also have

$$\left|\int_\Omega \nabla u\cdot\nabla\overline{u}\right| \le \|\nabla u\|_{L^2(\Omega)}\,\|\nabla\overline{u}\|_{L^2(\Omega)} \le C(T,\alpha,\kappa)e^{C(\alpha)\nu T},\tag{8.8}$$

so

$$K \le C(T,\alpha,\kappa)e^{C(\alpha)\nu T}\nu \le R\nu\tag{8.9}$$

for all ν in $(0,1]$ for some constant R.

By (7.2), $\|u\|_{L^\infty([0,T]\times\Omega)} \le C$ for all ν in $(0,1]$. It is also true that $\overline{u}$ is in $L^\infty([0,T]\times\Omega)$ (arguing, for instance, exactly as in (7.2)). Thus,

$$M = \sup_{\nu\in(0,1]} \||w|^2\|_{L^\infty([0,T]\times\Omega)}$$

is finite.

Also, because the L^p-norms of vorticity are conserved for $\overline{u}$, we have, by Corollary 3.2,

$$2\,\|\nabla\overline{u}(t)\|_{L^p(\Omega)} \le Cp\|\omega^0\|_{L^p(\Omega)} + C\|u^0\|_{L^2(\Omega)} \le Cp(\theta(p)+1/p)\tag{8.10}$$

for all $p\ge 2$. Because θ is admissible, so is $p\mapsto C[\theta(p)+1/p]$, and its associated β function—call it $\overline{\beta}$—is bounded by a constant multiple of that associated to θ; that is, $\overline{\beta}\le C\beta$.

We now proceed as in [14]. Let s be in $[0,T]$, and let

$$A = |w(s,x)|^2, \quad B = |\nabla\overline{u}(s,x)|, \quad L(s) = \|w(s)\|_{L^2}^2\,.$$

Then

$$\begin{aligned}\int_{\mathbb{R}^2} |w(s,x)|^2|\nabla\overline{u}(s,x)|\,dx &= \int_{\mathbb{R}^2} AB = \int_{\mathbb{R}^2} A^\epsilon A^{1-\epsilon}B \le M^\epsilon\int_{\mathbb{R}^2} A^{1-\epsilon}B\\ &\le M^\epsilon\|A^{1-\epsilon}\|_{L^{1/(1-\epsilon)}}\|B\|_{L^{1/\epsilon}} = M^\epsilon\|A\|_{L^1}^{1-\epsilon}\|B\|_{L^{1/\epsilon}}\\ &= M^\epsilon L(s)^{1-\epsilon}\|\nabla\overline{u}(s)\|_{L^{1/\epsilon}} \le CM^\epsilon L(s)^{1-\epsilon}\frac{1}{\epsilon}(\theta(1/\epsilon)+\epsilon).\end{aligned}$$

Since this is true for all ϵ in $[1/p_0,\infty)$, it follows that

$$2\int_{\mathbb{R}^2} |\nabla\overline{u}(s,x)||w(s,x)|^2\,dx \le C\overline{\beta}(L(s)) \le C\beta(L(s)).$$

From (8.6) and (8.9), then, we have

$$L(t) \le R\nu + C\int_0^t \beta(L(r))\,dr.$$

By Lemma 1.1,

$$\int_{R\nu}^{L(t)} \frac{ds}{C\beta(s)} = \left(-\int_{L(t)}^{1} + \int_{R\nu}^{1} \right) \frac{ds}{C\beta(s)} \leq \int_0^t ds = t. \tag{8.11}$$

It follows that for all t in $(0, T]$,

$$\int_{R\nu}^{1} \frac{ds}{\beta(s)} \leq CT + \int_{L(t)}^{1} \frac{ds}{\beta(s)}. \tag{8.12}$$

As $\nu \to 0^+$, the left side of (8.12) becomes infinite because of (8.2); hence, so must the right side. But this implies that $L(t) \to 0$ as $\nu \to 0^+$ and that the convergence is uniform over $[0, T]$. It also follows from (8.11) that

$$\int_{R\nu}^{L(t)} \frac{dr}{\beta(r)} \leq Ct \tag{8.13}$$

and that, as $\nu \to 0$, $L(t) \to 0$ uniformly over any finite time interval. The rate of convergence given in $L^\infty([0, T]; L^2(\Omega))$ in (8.4) can be derived from (8.13) precisely as in [14].

By (2.5),

$$\begin{aligned} \|u - \overline{u}\|_{L^2(\Gamma)} = \|w\|_{L^2(\Gamma)} &\leq C \|\nabla w\|_{L^2(\Omega)}^{1/2} \|w\|_{L^2(\Omega)}^{1/2} \\ &\leq C(T, \alpha, \kappa) e^{C(\alpha)\nu T} L(t)^{1/4}, \end{aligned}$$

from which the convergence rate for $L^\infty([0, T]; L^2(\Gamma))$ in (8.4) follows. □

The convergence rate in $L^\infty([0, T]; L^2(\Omega))$ established in Theorem 8.4 is the same as that established for the entire plane in [14], except for the presence of the constant C and the value of the constant R, which now increases with time (linearly, when α is nonnegative).

In the important special cased of bounded initial vorticity, one obtains the bound

$$\|u - \overline{u}\|_{L^\infty([0,T];L^2(\mathbb{R}^2))} \leq M^{1/2} \left(\frac{R\nu}{M} \right)^{\frac{1}{2} e^{-\epsilon T}} \tag{8.14}$$

for all t in $[0, T]$ for which $\nu < (M/R)e^{-2}$. Here, the R and M are defined as in the proof of Theorem 8.4, and $\epsilon = C\|\omega^0\|_{L^2 \cap L^\infty}$. When α is nonnegative, R is proportional to t, and (8.14) is essentially the same bound obtained by Chemin in [3] working in all of $\mathbb{R}^2$.

One can also calculate explicit bounds for the sequence of admissible vorticities in (8.3), obtaining bounds similar to that of (8.14), but with iterated exponentials. In general, it is not possible to obtain an explicit bound. The important point, however, is that, as described in section 5 of [14], it is possible to obtain an arbitrarily poor bound on the convergence rate for properly chosen initial vorticity. This is because the function f, which was defined implicitly in terms of β, can, conversely, be used to define β, and we can choose f so that it approaches zero arbitrarily slowly. (It is an open and difficult question whether initial vorticities actually exist that achieve arbitrarily slow convergence.)

In Theorem 8.4, we held α constant in (1.2) and let $\nu \to 0$, which is equivalent to letting $a \to 0$ linearly with ν in (1.1). One could modify the proof of Theorem 8.4 in

an attempt to obtain the vanishing viscosity limit with slower than linear convergence of a to 0 by being explicit about the value of the constant C_0 in Theorem 7.3. This constant controls the bounds on both K and M in the proof of Theorem 8.4, which, along with the L^p-norms of the initial vorticity, ultimately determine the convergence rate. But C_0 increases to infinity with $\|\alpha\|_{L^\infty(\Gamma)}$, and the bounds on K and M each increase to infinity with C_0. The conclusion is that $\|\alpha\|_{L^\infty(\Gamma)}$ must be bounded over sufficiently small values of ν for the approach in the proof of Theorem 8.4 to remain valid. Thus, using our approach, we cannot significantly improve over the assumption that α remains fixed as $\nu \to 0$ in the vanishing viscosity limit.

9. No-slip boundary conditions. As long as α is nonvanishing, we can let $\gamma = 1/\alpha$ and re-express the Navier boundary conditions in (1.2) as

$$v \cdot \mathbf{n} = 0 \text{ and } 2\gamma(\mathbf{n} \cdot D(v)) \cdot \boldsymbol{\tau} + v \cdot \boldsymbol{\tau} = 0 \text{ on } \Gamma. \tag{9.1}$$

When γ is identically zero, we have the usual no-slip boundary conditions. An obvious question to ask is whether it is possible to arrange for γ to approach zero in such a manner that the corresponding solutions to the Navier–Stokes equations with Navier boundary conditions approach the solution to the Navier–Stokes equations with the usual no-slip boundary conditions in $L^\infty([0,T];L^2(\Omega))$.

Let u^0 be an initial velocity in V, and assume that $\gamma > 0$ lies in $L^\infty(\Gamma)$. Fix a $\nu > 0$ and let

$$\begin{aligned} u &= \text{the unique solution to the Navier–Stokes equations} \\ &\quad \text{with Navier boundary conditions for } \alpha = 1/\gamma \text{ and} \\ \widetilde{u} &= \text{the unique solution to the Navier–Stokes equations} \\ &\quad \text{with no-slip boundary conditions,} \end{aligned}$$

in each case with the same initial velocity u^0.

If we let γ approach 0 uniformly on the boundary, we automatically have some control over u on the boundary.

LEMMA 9.1. *For sufficiently small* $\|\gamma\|_{L^\infty(\Gamma)}$,

$$\|u\|_{L^2([0,T];L^2(\Gamma))} \leq \frac{\|u^0\|_{L^2(\Omega)}}{\sqrt{\nu}} \|\gamma\|^{1/2}_{L^\infty(\Gamma)}. \tag{9.2}$$

Proof. Assume that $\|\gamma\|_{L^\infty(\Gamma)}$ is sufficiently small that $\alpha > \kappa$ on Γ. Then, as in the proof of Theorem 6.1, we have

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \nu \|\nabla u(t)\|^2_{L^2(\Omega)} = \nu \int_\Gamma (\kappa - \alpha) u \cdot u,$$

so

$$\|u(t)\|^2_{L^2(\Omega)} \leq \|u^0\|^2_{L^2(\Omega)} + 2\nu \int_0^t \int_\Gamma (\kappa - \alpha) u \cdot u.$$

But

$$\int_\Gamma (\kappa - \alpha) u \cdot u \leq -\inf_\Gamma \{\alpha - \kappa\} \|u(t)\|^2_{L^2(\Gamma)},$$

so

$$\|u(t)\|^2_{L^2(\Omega)} \leq \|u^0\|^2_{L^2(\Omega)} - 2\nu \inf_\Gamma \{\alpha - \kappa\} \|u\|^2_{L^2([0,t];L^2(\Gamma))}$$

and

$$\|u\|^2_{L^2([0,t];L^2(\Gamma))} \le \|u^0\|^2_{L^2(\Omega)}/(2\nu \inf_\Gamma \{\alpha - \kappa\}).$$

Then (9.2) follows because $\|\gamma\|_{L^\infty(\Gamma)} \inf_\Gamma \{\alpha - \kappa\} \to 1$ as $\|\gamma\|_{L^\infty(\Gamma)} \to 0$. ∎

If we assume enough smoothness of the initial data and of Γ, we can use (9.2) to establish convergence of u to $\widetilde{u}$ as $\|\gamma\|_{L^\infty(\Gamma)} \to 0$.

THEOREM 9.2. *Fix $T > 0$, assume that u^0 is in $V \cap H^3(\Omega)$ with $u^0 = 0$ on Γ, and assume that Γ is C^3. Then for any fixed $\nu > 0$,*

$$u \to \widetilde{u} \text{ in } L^\infty([0,T];L^2(\Omega)) \cap L^2([0,T];\dot{H}^1(\Omega)) \cap L^2([0,T];L^2(\Gamma)) \tag{9.3}$$

as $\gamma \to 0$ in $L^\infty(\Gamma)$. Here, $\dot{H}^1(\Omega)$ is the homogeneous Sobolev space.

Proof. First, u exists and is unique by Theorem 6.1; the existence and uniqueness of $\widetilde{u}$ are classical results. Because u^0 is in $H^3(\Omega)$ and Γ is C^3, $\widetilde{u}$ is in $L^\infty([0,T];H^3(\Omega))$ by the argument on page 205 of [22] following the proof of Theorem 3.6 of [22]. Hence, $\nabla\widetilde{u}$ is in $L^\infty([0,T];H^2(\Omega))$ and so in $L^\infty([0,T];C(\overline{\Omega}))$.

Arguing as in the proof of Theorem 8.4 with $w = u - \widetilde{u}$, we have

$$\int_\Omega \partial_t w \cdot w + \int_\Omega w \cdot (u \cdot \nabla w) + \int_\Omega w \cdot (w \cdot \nabla \widetilde{u}) + \nu \int_\Omega \nabla w \cdot \nabla w - \nu \int_\Gamma (\kappa - \alpha) u \cdot w + \nu \int_\Gamma (\mathbf{n} \cdot \nabla \widetilde{u}) \cdot w = 0.$$

(Even though w is not a valid test function for $\widetilde{u}$, we are working with sufficiently smooth solutions that this integration is still valid. Since w is divergence-free and tangential to the boundary, the pressure term for each equation integrates to zero.) But $\widetilde{u} = 0$ on Γ, so $w = u$ on Γ, and

$$\int_\Omega \partial_t w \cdot w + \int_\Omega w \cdot (w \cdot \nabla \widetilde{u}) + \nu \int_\Omega |\nabla w|^2 + \nu \int_\Gamma (\alpha - \kappa)|u|^2 + \nu \int_\Gamma (\mathbf{n} \cdot \nabla \widetilde{u}) \cdot u = 0.$$

For $\|\gamma\|_{L^\infty(\Gamma)}$ sufficiently small that $\alpha = 1/\gamma > \kappa$ on Γ, integrating over time gives

$$\|w(t)\|^2_{L^2(\Omega)} + \nu \int_0^t \|\nabla w\|^2_{L^2(\Omega)} \le K + 2 \int_0^t \int_\Omega |w|^2 |\nabla \widetilde{u}|, \tag{9.4}$$

where

$$K = -2\nu \int_0^t \int_\Gamma (\mathbf{n} \cdot \nabla \widetilde{u}) \cdot u \le 2\nu \int_0^t \|\nabla \widetilde{u}\|_{L^2(\Gamma)} \|u\|_{L^2(\Gamma)}$$
$$\le C\nu \int_0^t \|\widetilde{u}\|_{H^2(\Omega)} \|u\|_{L^2(\Gamma)} \le C\nu \|\widetilde{u}\|_{L^2([0,T];H^2(\Omega))} \|u\|_{L^2([0,T];L^2(\Gamma))}.$$

By Theorem 3.10 on page 213 of [22], $\|\widetilde{u}\|_{L^2([0,T];H^2(\Omega))}$ is finite (though the bound on it in [22] increases to infinity as ν goes to 0), so by Lemma 9.1,

$$K \le C_1(\nu) \|\gamma\|^{1/2}_{L^\infty(\Gamma)}. \tag{9.5}$$

Because $\nabla \widetilde{u}$ is in $L^\infty([0,T];C(\overline{\Omega}))$,

$$\int_0^t \int_\Omega |w|^2 |\nabla \widetilde{u}| \leq C_2(\nu) \int_0^t \|w(s)\|^2_{L^2(\Omega)}\, ds,$$

where $C_2(\nu) = \|\nabla \widetilde{u}\|_{L^\infty([0,T]\times\Omega)}$, and (9.4) becomes

$$\|w(t)\|^2_{L^2(\Omega)} + \nu \int_0^t \|\nabla w\|^2_{L^2(\Omega)} \leq C_1(\nu) \|\gamma\|^{1/2}_{L^\infty(\Gamma)} + C_2(\nu) \int_0^t \|w(s)\|^2_{L^2(\Omega)}\, ds.$$

By Gronwall's lemma,

$$\|w(t)\|^2_{L^2(\Omega)} \leq C_1(\nu) \|\gamma\|^{1/2}_{L^\infty(\Gamma)} e^{C_2(\nu)t},$$

and the convergence in $L^\infty([0,T];L^2(\Omega))$ and thus also in $L^2([0,T];\dot{H}^1(\Omega))$ follow immediately. Convergence in $L^2([0,T];L^2(\Gamma))$ then follows directly from Lemma 9.1, since $\widetilde{u} = 0$ on Γ. □

10. The boundary layer. In [13], Kato investigates the vanishing viscosity limit of solutions of the Navier–Stokes equations with no-slip boundary conditions to a solution of the Euler equations for a bounded domain in $\mathbb{R}^d$, $d \geq 2$. What Kato shows is the following: Let u be the weak solution to the Navier–Stokes equations with no-slip boundary conditions and with $u(0)$ in H, and let $\overline{u}$ be the solution to the Euler equations, where sufficient smoothness is assumed for $\overline{u}(0)$ that $\nabla \overline{u}$ is bounded on $[0,T]\times\Omega$. Assume also that $u(0) \to \overline{u}(0)$ in H as $\nu \to 0$. Then the following are equivalent:

(i) $u(t) \to \overline{u}(t)$ in $L^2(\Omega)$ as $\nu \to 0$ uniformly over t in $[0,T]$;
(ii) $u(t) \to \overline{u}(t)$ in $L^2(\Omega)$ as $\nu \to 0$ weakly for all t in $[0,T]$;
(iii) $\nu \int_0^T \|\nabla u\|^2_{L^2(\Omega)}\, dt \to 0$ as $\nu \to 0$;
(iii′) $\nu \int_0^T \|\nabla u\|^2_{L^2(\Gamma_{c\nu})}\, dt \to 0$ as $\nu \to 0$.

Here, $\Gamma_{c\nu}$ is the boundary strip of width $c\nu$ with $c > 0$ fixed but arbitrary.

Let us return to the setting of Theorem 8.4 and compare the situation to that of [13]. We now have zero forcing and the same initial conditions for both (NS) and (E), which simplifies the analysis in [13] slightly, but we have different boundary conditions on (NS) and we have insufficient smoothness of u^0 for Kato's conditions to apply. However, we have already proven that condition (i) holds and hence also (ii), and since conditions (iii) and (iii') follow immediately from (7.1), no further work is required to show the equivalence of Kato's four conditions.

(It is also possible to directly adapt Kato's argument to our setting, thereby establishing the vanishing viscosity limit in the spirit of Kato. This requires, however, all of the results developed to prove the vanishing viscosity limit in Theorem 8.4 and considerably more effort besides.)

We can bound the rate at which the convergence in condition (iii′) occurs, giving us some idea of what is happening in the boundary layer.

THEOREM 10.1. *With the assumptions in Theorem* 8.4*,*

$$\nu \int_0^T \|\nabla u\|^2_{L^2(\Gamma_{c\nu})}\, dt \leq C(p,T,\alpha,\kappa,u^0) T (2c)^{1-2/p} \nu^{2-2/p}$$

for all p *in* $(2,\infty)$ *and* t *in* $[0,T]$.

Proof. We have, using (7.1) and Corollary 3.2,

$$\begin{aligned}\|\nabla u\|_{L^2(\Gamma_{\delta/2})} &\le \|z\nabla u\|_{L^2(\Gamma_\delta)} \le \|z\|_{L^{p'}(\Gamma_\delta)} \|\nabla u\|_{L^p(\Omega)} \\ &\le \|z\|_{L^{p'}(\Gamma_\delta)} \left(Cp\,\|\omega\|_{L^p(\Omega)} + C'\,\|u\|_{L^2(\Omega)}\right) \\ &\le \|z\|_{L^{p'}(\Gamma_\delta)} \left(Cp\left(\|\omega^0\|_{L^p} + C_0\right) + C'\,\|u\|_{L^2(\Omega)}\right),\end{aligned}$$

where $1/p' + 1/p = 1/2$. But

$$\|z\|_{L^{p'}} \le C\delta^{1/p'} = C\delta^{1/2-1/p}.$$

Substituting this into the earlier inequality, squaring the result, setting $\delta = 2c\nu$, and integrating over time conclude the proof. ∎

The proof of Theorem 10.1 shows that the square of the gradient of the velocity for a solution to (NS) with Navier boundary conditions vanishes in the L^2-norm nearly linearly with the width of the boundary layer. We could obtain linear convergence for appropriate smoother initial velocities if we could show that $\|\nabla u\|_{L^\infty(\Omega)}$ is bounded uniformly over small ν. It is not at all clear, however, whether such a result is obtainable. In any case, the behavior of the boundary layer for Navier boundary conditions is principally derived from the boundary conditions themselves and is not highly dependent upon the smoothness of the initial velocity.

This is in contrast to no-slip boundary conditions, where for smooth data probably the strongest general statement that can be made was made by Kato in [13] with his equivalent conditions for the vanishing viscosity limit. (See also the incremental improvement in [23] and [27].) For the less regular initial velocities that we assume in Theorem 8.4, it is quite possible that a condition stronger than Kato's condition (iii′) is required to imply convergence in the vanishing viscosity limit. This is because there is no known bound on $\|u\|_{L^2([0,T];L^\infty(\Omega))}$ uniform over small ν, which is required to achieve the vanishing viscosity limit using Osgood's lemma as in section 8. In fact, obtaining such a bound would almost certainly require obtaining a uniform bound on the L^p-norm of the vorticity for some $p > 2$, which is tantamount to establishing the vanishing viscosity limit to begin with, at least for smooth initial data.

Appendix. Compatible sequences. For p in $(1,\infty)$, define the spaces

$$X_0^p = H_0 \cap H^{1,p}(\Omega) \text{ and } X^p = H \cap H^{1,p}(\Omega) = X_0^p \oplus H_c, \tag{A.1}$$

each with the $H^{1,p}(\Omega)$-norm.

LEMMA A.1. *Let p be in $(1,\infty]$. For $p < 2$ let $\widehat{p} = p/(2-p)$, for $p > 2$ let $\widehat{p} = \infty$, and for $p = 2$ let $\widehat{p}$ be any value in $[2,\infty]$. Then for any v in X_0^p,*

$$\|v\|_{L^{\widehat{p}}(\Gamma)} \le C(p)\,\|\omega(v)\|_{L^p(\Omega)}.$$

Proof. For $p < 2$ and any v in X_0^p, we have

$$\begin{aligned}\|v\|_{L^{\widehat{p}}(\Gamma)} &\le C(p)\,\|v\|_{L^p(\Omega)}^{1-\lambda}\,\|\nabla v\|_{L^p(\Omega)}^{\lambda} \le C(p)\,\|\nabla v\|_{L^p(\Omega)} \\ &\le C(p)\,\|\omega(v)\|_{L^p(\Omega)},\end{aligned}$$

where $\lambda = 2(\widehat{p}-p)/(p(\widehat{p}-1)) = 1$ if $p < 2$ and $\lambda = 2/p$ if $p \ge 2$. The first inequality follows from Theorem 3.1 on page 43 of [7], the second follows from (2.3), and the third from Lemma 3.1. ∎

Given a vorticity ω in $L^p(\Omega)$ with p in $(1,\infty)$, the Biot–Savart law gives a vector field v in H whose vorticity is ω. (That v is in $L^2(\Omega)$ follows as in the proof of Lemma A.1, Ω being bounded.) Let $v = u + h$, where u is in H_0 and h is in H_c. Then $\omega(u) = \omega$ as well, so we can define a function $K_\Omega\colon\ L^p(\Omega) \to H_0$ by $\omega \mapsto u$ having the property that $\omega(K_\Omega[\omega]) = \omega$. By (2.3) and Lemma 3.1, u is also in $H^{1,p}(\Omega)$, so, in fact, $K_\Omega\colon\ L^p(\Omega) \to X_0^p$ and is the inverse of the function ω. It is continuous by the same two lemmas. We can write the inequality in Lemma A.1, then, as $\|K_\Omega[\omega]\|_{L^{\widehat{p}}(\Gamma)} \le C(p)\,\|\omega\|_{L^p(\Omega)}$.

THEOREM A.2. *Assume that Γ is C^2 and α is in $L^\infty(\Gamma)$. Let v be in X^p for some p in $(1,\infty)$ and have vorticity ω. Then there exists a sequence (v_n) of compatible vector fields (Definition 7.2) whose vorticities converge strongly to ω in $L^p(\Omega)$. The vector fields (v_n) converge strongly to v in X^p and, if $p \ge 2$, also in V.*

Proof. Our proof is a minor adaptation of that of Lemma 2 of [17], which we first summarize. Let N_n be a tubular neighborhood of Γ of width $2/n$ (for n sufficiently large) and let $U_n = N_n \cap \Omega$. Define $d\colon U_n \to \mathbb{R}^+$ by $d(x) = \operatorname{dist}(x,\Gamma)$ and $r\colon U_n \to \Gamma$ by letting $r(x)$ be the nearest point to x on Γ. Define a cutoff function ζ_n in $C^\infty(\Omega)$ taking values in $[0,1]$ so that $\zeta_n \equiv 0$ on U_{n+1} and $\zeta_n \equiv 1$ on $\Omega \setminus U_n$, and let the sequence (η_k) be an approximation of the identity.

It is shown in [17] that β is a continuous extension operator from $L^{\widehat{p}}(\Gamma)$ into $L^p(\Omega)$, where

$$\beta(G)(x) := \zeta_n(x)(\eta_n * \omega)(x) + (1 - \zeta_n(x))e^{-nd(x)}G(r(x))$$

and where $\widehat{p}$ is defined as in Lemma A.1. In calculating $\eta_n * \omega$, we extend ω by zero to all of $\mathbb{R}^2$. Defining $\Psi\colon L^{\widehat{p}}(\Gamma) \to L^{\widehat{p}}(\Gamma)$ by

$$\Psi(G) = (2\kappa - \alpha)K_\Omega[\beta_n(G)] \cdot \boldsymbol{\tau},$$

it is shown that Ψ is a contraction mapping for sufficiently large n and so has a unique fixed point, G^n. Finally, defining $\omega_n = \beta(G^n)$, the authors show that ω_n converges to ω in $L^p(\Omega)$ (this argument uses Lemma A.1). Key to this last step is demonstrating that $\|G^n\|_{L^{\widehat{p}}(\Gamma)}$ is bounded over n.

Since the authors of [17] are working in a simply connected domain, they can deal exclusively with vorticity. To adapt their proof to multiply connected domains, where we must recover the velocity with the proper harmonic component, requires only one change to their construction. We suppose that $v = u + h$ with $u \in X_0^p$ and h in H_c and define

$$\Psi(G) = (2\kappa - \alpha)(K_\Omega[\beta(G)] + h) \cdot \boldsymbol{\tau}.$$

In forming the difference $\Psi(G_1) - \Psi(G_2)$ the term $(2\kappa - \alpha)h \cdot \boldsymbol{\tau}$ cancels, and the existence of a unique fixed point G^n follows precisely as in [17].

We can now define

$$\omega_n = \beta(G_n),\ v_n = K_\Omega[\omega_n] + h,$$

and observe that *on* Γ,

$$\begin{aligned}\omega(v_n) = \omega_n = \beta(G^n) = G^n = \Psi(G^n) &= (2\kappa - \alpha)(K_\Omega[\beta(G^n)] + h) \cdot \boldsymbol{\tau} \\ &= (2\kappa - \alpha)(K_\Omega[\omega_n] + h) \cdot \boldsymbol{\tau} = (2\kappa - \alpha)v_n \cdot \boldsymbol{\tau},\end{aligned}$$

so v_n satisfies the Navier boundary conditions. (Note that we had to include the harmonic component h of the velocity in the definition of Ψ; we could not simply apply Lemma 2 of [17] to u and add h to the resulting vector field, because such a vector field would not, in general, satisfy the Navier boundary conditions.)

The convergence of ω_n to ω in $L^p(\Omega)$ is argued as in [17], except that now, to show that $\|G^n\|_{L^{\hat{p}}(\Gamma)}$ is bounded over n, we have

$$\begin{aligned}\|G^n\|_{L^{\hat{p}}(\Gamma)} &\le \|2\kappa-\alpha\|_{L^\infty}\,\|K_\Omega[\omega_n]+h\|_{L^{\hat{p}}(\Gamma)}\\ &\le C(\|K_\Omega[\omega_n]\|_{L^{\hat{p}}(\Gamma)}+\|h\|_{L^{\hat{p}}(\Gamma)})\\ &\le C\left(\|\omega\|_{L^p(\Omega)}+\frac{1}{2}\,\|G^n\|_{L^{\hat{p}}(\Gamma)}+\|\nabla h\|_{L^p(\Omega)}\right)\end{aligned}$$

for n sufficiently large. Here, the bound on $\|K_\Omega[\omega_n]\|_{L^{\hat{p}}(\Gamma)}$ is as in [17] and the bound on $\|h\|_{L^{\hat{p}}(\Gamma)}$ follows from Theorem 3.1 on page 43 of [7] and (2.3) as in the proof of Lemma A.1. It follows that $\|G^n\|_{L^{\hat{p}}(\Gamma)} \le C\,\|v\|_{X_p}$ for sufficiently small n, which is what is required to complete the proof of the convergence of ω_n to ω in $L^p(\Omega)$ as in [17].

To prove the convergence of v_n to v in X^p, we observe that

$$\begin{aligned}\|\nabla v-\nabla v_n\|_{L^p(\Omega)} &= \|\nabla u+\nabla h-(\nabla K_\Omega[\omega_n]+\nabla h)\|_{L^p(\Omega)}\\ &= \|\nabla(u-K_\Omega[\omega_n])\|_{L^p(\Omega)} \le Cp\,\|\omega(u-K_\Omega[\omega_n])\|_{L^p(\Omega)}\\ &= Cp\,\|\omega-\omega_n\|_{L^p(\Omega)}\,,\end{aligned}$$

where we used Lemma 3.1. Then by (2.3), v_n converges strongly to v in X^p as well. Convergence in V for $p \ge 2$ follows since Ω is bounded. □

Our only use of Theorem A.2 is in the proofs of Theorem 7.3 and Corollary A.3. In both of these instances we need only the case $p \ge 2$. We include all the cases, however, for the same reason as in [17]: we hope that if the vorticity bound in Lemma 3 of [17] can be extended to p in $(1,2)$, then the convergence in Proposition 1 of [17] can also be extended (for multiply connected Ω).

COROLLARY A.3. *Assume that Γ is C^2, and α is in $L^\infty(\Gamma)$. Then there exists a basis for V lying in $\mathcal{W}$ that is also a basis for H.*

Proof. The space $V = (V \cap H_0) \oplus H_c$ is separable because $V \cap H_0$ is the image under the continuous function K_Ω of the separable space $L^2(\Omega)$ and H_c is finite-dimensional. Let $\{v_i\}_{i=1}^\infty$ be a dense subset of V. Applying Theorem A.2 to each v_i and unioning all the sequences, we obtain a countable subset $\{u_i\}_{i=1}^\infty$ of $\mathcal{W}$ that is dense in V. Selecting a maximal independent set gives us a basis for V and for H as well, since V is dense in H. □

Acknowledgments. The author wishes to thank Josef Málek for suggesting that he explore the vanishing viscosity limit with Navier boundary conditions, and Misha Vishik for many useful discussions.

REFERENCES

[1] D. BAĬNOV AND P. SIMEONOV, *Integral inequalities and applications*, Math. Appl. 57, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

[2] I. BIHARI, *A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations*, Acta Math. Acad. Sci. Hungar., 7 (1956), pp. 81–94.

[3] J.-Y. Chemin, *A remark on the inviscid limit for two-dimensional incompressible fluids*, Comm. Partial Differential Equations, 21 (1996), pp. 1771–1779.
[4] J.-Y. Chemin, *Perfect Incompressible Fluids*, Oxford Lecture Ser. Math. Appl. 14, The Clarendon Press, Oxford University Press, New York, 1998. Translated from the 1995 French original by Isabelle Gallagher and Dragos Iftimie.
[5] T. Clopeau, A. Mikelić, and R. Robert, *On the vanishing viscosity limit for the* $2D$ *incompressible Navier-Stokes equations with the friction type boundary conditions*, Nonlinearity, 11 (1998), pp. 1625–1636.
[6] T. M. Flett, *Differential Analysis*, Cambridge University Press, Cambridge, UK, 1980.
[7] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations, Vol.* I, Springer Tracts Nat. Philos. 38, Springer-Verlag, New York, 1994.
[8] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, 1977.
[9] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman (Advanced Publishing Program), Boston, MA, 1985.
[10] W. Jäger and A. Mikelić, *On the roughness-induced effective boundary conditions for an incompressible viscous flow*, J. Differential Equations, 170 (2001), pp. 96–122.
[11] W. Jäger and A. Mikelić, *Couette flows over a rough boundary and drag reduction*, Comm. Math. Phys., 232 (2003), pp. 429–455.
[12] J. Volker, W. Layton, and N. Sahin, *Derivation and analysis of near wall models for channel and recirculating flows*, Comput. Math. Appl., 48 (2004), pp. 1135–1151.
[13] T. Kato, *Remarks on zero viscosity limit for nonstationary Navier-Stokes flows with boundary*, in Seminar on Nonlinear Partial Differential Equations (Berkeley, CA, 1983), Math. Sci. Res. Inst. Publ. 2, Springer-Verlag, New York, 1984, pp. 85–98.
[14] J. P. Kelliher, *The inviscid limit for two-dimensional incompressible fluids with unbounded vorticity*, Math. Res. Lett., 11 (2004), pp. 519–528.
[15] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.
[16] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. Vol.* 1, Oxford Lecture Ser. Math. Appl. 3, The Clarendon Press, Oxford University Press, New York, 1996.
[17] M. C. Lopes Filho, H. J. Nussenzveig Lopes, and G. Planas, *On the inviscid limit for two-dimensional incompressible flow with Navier friction condition*, SIAM J. Math. Anal., 36 (2005), pp. 1130–1141.
[18] J. C. Maxwell, *On stresses in rarified gases arising from inequalities of temperature*, Phil. Trans. Royal Society, 1879, pp. 704–712.
[19] C. M. L. H. Navier, *Sur les lois de l'equilibre et du mouvement des corps élastiques*, Mem. Acad. R. Sci. Inst. France, 6 (1827), p. 369.
[20] W. F. Osgood, *Beweis der Existenz einer Lösung der Differentialgleichung* $dy/dx = f(x, y)$ *ohne Hinzunahme der Cauchy-Lipschitz'schen Bedingung*, Monatsch. Math. Phys., 9 (1898), pp. 331–345.
[21] M. Tehranchi, *A Succinct Proof of Osgood's Lemma*, Private communication.
[22] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis. Reprint of the* 1984 *edition*, AMS Chelsea, Providence, RI, 2001.
[23] R. Temam and X. Wang, *On the behavior of the solutions of the Navier-Stokes equations at vanishing viscosity*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), pp. 807–828.
[24] V. I. Yudovich, *Some bounds for solutions of elliptic equations*, Mat. Sb. (N.S.), 59 (1962), pp. 229–244 (in Russian); English translation in Amer. Math. Soc. Transl. (2), 56 (1966), pp. 1–18.
[25] V. I. Yudovich, *Non-stationary flows of an ideal incompressible fluid*, Ž. Vyčisl. Mat. i Mat. Fiz., 3 (1963), pp. 1032–1066 (in Russian).
[26] V. I. Yudovich, *Uniqueness theorem for the basic nonstationary problem in the dynamics of an ideal incompressible fluid*, Math. Res. Lett., 2 (1995), pp. 27–38.
[27] X. Wang, *A Kato type theorem on zero viscosity limit of Navier-Stokes flows,* Indiana Univ. Math. J., 50 (Special Issue) (2001), pp. 223–241.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 233–258

UNIQUENESS AND ASYMPTOTICS OF TRAVELING WAVES OF MONOSTABLE DYNAMICS ON LATTICES*

XINFU CHEN†, SHENG-CHEN FU‡, AND JONG-SHENQ GUO§

Abstract. Established here is the uniquenes of solutions for the traveling wave problem $cU'(x) = U(x+1)+U(x-1)-2U(x)+f(U(x))$, $x \in \mathbb{R}$, under the monostable nonlinearity: $f \in C^1([0,1])$, $f(0) = f(1) = 0 < f(s)\ \forall\, s \in (0,1)$. Asymptotic expansions for $U(x)$ as $x \to \pm\infty$, accurate enough to capture the translation differences, are also derived and rigorously verified. These results complement earlier existence and partial uniqueness/stability results in the literature. New tools are also developed to deal with the degenerate case $f'(0)f'(1) = 0$, about which is the main concern of this article.

Key words. traveling wave, monostable, degenerate, lattice dynamics

AMS subject classifications. Primary, 34K05; Secondary, 34E05, 34K25, 34K60

DOI. 10.1137/050627824

1. Introduction.

Consider a system of countably many ordinary differential equations, for $\{u_n(\cdot)\}_{n\in\mathbb{Z}}$,

$$\dot{u}_n(t) = u_{n+1}(t) - 2u_n(t) + u_{n-1}(t) + f(u_n(t)), \qquad n \in \mathbb{Z}, t > 0, \tag{1.1}$$

where f is a nonlinear forcing term satisfying $f(0) = f(1) = 0$. This system can be embedded into a larger one, for an unknown $\{u(x,\cdot)\}_{x\in\mathbb{R}}$,

$$u_t(x,t) = u(x+1,t) - 2u(x,t) + u(x-1,t) + f(u(x,t)), \qquad x \in \mathbb{R},\ t > 0. \tag{1.2}$$

A solution of (1.2) or (1.1) is called a *traveling wave with speed* c if there exists a function U defined on $\mathbb{R}$ such that $u(x,t) = U(x+ct)$ or $u_n(t) = U(n+ct)$. Here U is referred to as the *wave profile*. Of interest are solutions taking values in $[0,1]$, specifically, traveling waves connecting the steady states $\mathbf{0}$ and $\mathbf{1}$, i.e., traveling wave solutions $(c,U) \in \mathbb{R}\times C^1(\mathbb{R})$ of the traveling wave problem

$$\begin{cases} cU'(\cdot) = U(\cdot+1) + U(\cdot-1) - 2U(\cdot) + f(U(\cdot)) & \text{on } \mathbb{R}, \\ U(-\infty) = 0, \qquad U(\infty) = 1, \qquad 0 \leqslant U \leqslant 1 & \text{on } \mathbb{R}. \end{cases} \tag{1.3}$$

Equation (1.1) can be found in many biological models (e.g., [9, 20, 22]). Also, it can be regarded as a spatial-discrete version of the parabolic partial differential equation

$$u_t = u_{xx} + f(u). \tag{1.4}$$

*Received by the editors March 28, 2005; accepted for publication (in revised form) October 12, 2005; published electronically April 21, 2006. This work was partially supported by National Science Council of the Republic of China grants NSC 93-2811-M-003-008, NSC 93-2115-M-004-001, and NSC 93-2115-M-003-011.

http://www.siam.org/journals/sima/38-1/62782.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu). The research of this author was supported by National Science Foundation grant DMS–0203991.

‡Department of Mathematical Sciences, National Chengchi University, Taipei 116, Taiwan (fu@math.nccu.edu.tw).

§Corresponding author. Department of Mathematics, National Taiwan Normal University, Taipei 116, Taiwan (jsguo@math.ntnu.edu.tw).

The existence, uniqueness, and stability of traveling waves of (1.1) have been extensively studied recently under various assumptions on f; see, for example, [1, 5, 6, 7, 10, 12, 24, 25, 26, 27]. The commonly used assumption includes the condition of nondegeneracy $f'(0)f'(1) \neq 0$. For bistable dynamics, i.e., $f'(0) < 0$ and $f'(1) < 0$, the results on traveling waves are quite complete; see, for example, [1, 7, 25, 26] and the references therein. This paper concerns only the monostable dynamics, i.e., f satisfies

$$f \in C^1([0,1]), \quad f(0) = f(1) = 0 < f(s) \quad \forall\, s \in (0,1). \tag{A}$$

Under the nondegeneracy and the condition that $f(s) \leqslant f'(0)s$ for all $s \in [0,1]$, Zinner, Harris, and Hudson established the existence of traveling waves [27]; see also the later developments of Fu, Guo, and Shieh [10] and Chen and Guo [5]. The uniqueness issue was not satisfactorily resolved until a recent paper of Chen and Guo [6]. For easy reference, we quote here the following existence and uniqueness result from [6].

PROPOSITION 1. *Assume* (A).

(i) *There exists* $c_{\min} > 0$ *such that* (1.3) *admits a solution if and only if* $c \geqslant c_{\min}$.

(ii) *Given* $c \geqslant c_{\min}$, *there is a speed* c *wave profile satisfying* $U' > 0$ *on* $\mathbb{R}$.

(iii) *Given* $c > 0$, (1.3) *admits a solution if there is a supersolution of speed* c.

(iv) *When* $f'(0)f'(1) \neq 0$, *wave profiles are unique up to a translation. In addition,*

$$\lim_{x\to-\infty} \frac{U'(x)}{U(x)} = \lambda, \qquad \lim_{x\to\infty} \frac{U'(x)}{U(x)-1} = \mu, \tag{1.5}$$

where λ *is a positive real root of the characteristic equation*

$$c\,\lambda = e^{\lambda} + e^{-\lambda} - 2 + f'(0) \tag{1.6}$$

and μ *is the negative real root of the characteristic equation*

$$c\,\mu = e^{\mu} + e^{-\mu} - 2 + f'(1). \tag{1.7}$$

In addition, when $c > c_{\min}$, λ *is the smaller real root of the characteristic equation* (1.6).

Here by a *supersolution of wave speed* c it means a nonconstant Lipschitz continuous function Φ from $\mathbb{R}$ to $[0,1]$ satisfying

$$c\,\Phi'(x) \geqslant \Phi(x+1) + \Phi(x-1) - 2\Phi(x) + f(\Phi(x)) \quad \text{a.e.} \quad x \in \mathbb{R}.$$

Note that for any real numbers m and k, the function $z \in \mathbb{R} \to e^z + e^{-z} + mz + k$ is strictly convex, so the characteristic equation has at most two real roots. Since $f'(1) \leqslant 0$ and $c > 0$, there is a unique nonpositive real root μ to $c\,\mu = e^{\mu} + e^{-\mu} - 2 + f'(1)$. For the characteristic equation at 0, we define

$$c_* = \min_{z>0} \frac{e^z + e^{-z} - 2 + f'(0)}{z} \begin{cases} > 0 & \text{if } f'(0) > 0, \\ = 0 & \text{if } f'(0) = 0. \end{cases} \tag{1.8}$$

Suppose $f'(0) > 0$. There are two real roots to $c\lambda = e^{\lambda} + e^{-\lambda} - 2 + f'(0)$ when $c > c_*$; both are positive. When $c = c_*$, there is a unique real root, positive and of multiplicity two. When $c < c_*$, there are no real roots, so the assertion of Proposition 1 implicitly implies that $c_{\min} \geqslant c_*$. In addition, suppose $f(s) \leqslant f'(0)s$ for all $s \in [0,1]$.

Then it is easy to verify that $\Phi(x) := \min\{e^{\lambda x}, 1\}$ is a supersolution of speed c if $c\lambda = e^{\lambda} + e^{-\lambda} - 2 + f'(0)$. This implies that $c_{\min} = c_*$. When $f'(0) = 0$, we see that $c_* = 0$ and $\lambda = 0$ is a root to the characteristic equation at 0. Nevertheless, since $c_{\min} > 0$, we see an example that $c_{\min} > c_*$.

It is important to observe that a (monotonic) wave profile $U^{\min}$ of the minimum speed is a supersolution of any wave speed $c > c_{\min}$. Since among all wave profiles of all admissible speeds $U^{\min}$ decays with the largest exponential rate as $x \to -\infty$, it is not always true that near $-\infty$ a supersolution is bigger than a true solution under a certain translation. Thus, Proposition 1(iii) is highly nontrivial; its proof in [6] was based on an original idea of the authors of [27], with a simplification that avoids the use of degree theory.

The purpose of this paper is to remove the nondegeneracy condition $f'(0)f'(1) \neq 0$ made in Proposition 1(iv); that is, we are mainly concerned with the degenerate case $f'(0)f'(1) = 0$. We shall also introduce a number of new techniques. In terms of the differential equation (1.4), existence, uniqueness, and asymptotic stability of traveling waves have been established (cf. [13, 14, 17, 21]). Here we would like to extend the analogous result for (1.4) to (1.1). We summarize our results for the traveling wave problem (1.3) as follows.

THEOREM 1. *Assume* (A). *Wave profiles of a given speed are unique up to a translation.*

THEOREM 2. *Assume* (A). *Any wave profile is monotonic; i.e.,* $U' > 0$ *on* $\mathbb{R}$.

THEOREM 3. *Assume* (A). *Any solution* (c, U) *of* (1.3) *satisfies* (1.5) *and*

$$\lim_{x\to-\infty} \frac{U''(x)}{U'(x)} = \lambda, \qquad \lim_{x\to-\infty} \frac{f(U(x))}{U'(x)} = \begin{cases} c & \text{if } \lambda = 0, \\ f'(0)/\lambda & \text{otherwise,} \end{cases}$$

$$\lim_{x\to\infty} \frac{U''(x)}{U'(x)} = \mu, \qquad \lim_{x\to\infty} \frac{f(U(x))}{U'(x)} = \begin{cases} c & \text{if } \mu = 0, \\ f'(1)/\mu & \text{otherwise,} \end{cases}$$

where λ *is a nonnegative real root of the characteristic equation* (1.6) *and* μ *is the nonpositive real root of* (1.7).

In addition, λ *is the smaller root when* $c > c_{\min}$ *and the larger root when* $c = c_{\min}$.

Note that the root $\mu \leqslant 0$ to (1.7) is unique. In particular, $\mu = 0$ when $f'(1) = 0$. Also, $\lambda = 0$ when $f'(0) = 0$ and $c > c_{\min}$; otherwise, $\lambda > 0$. Note also that when $c_{\min} > c_*$, the characteristic equation (1.6) always has two positive real roots. To our knowledge, it is new in the literature that, as a principle, λ is the larger root of the characteristic equation (1.6) when $c = c_{\min} > c_*$, where c_* is as in (1.8).

In [6], the following general system is considered

$$u_t(x,t) = g(u(x+1,t)) - 2g(u(x,t)) + g(u(x-1,t)) + f(u(x,t)),$$

where $g(\cdot)$ is increasing. Under a variable change $v = [g(u) - g(0)]/[g(1) - g(0)]$, the system can be rewritten as

$$h(v(x,t))v_t(x,t) = v(x+1,t) + v(x-1,t) - 2v(x,t) + \tilde{f}(v(x,t)).$$

Under assumptions that $h \in C^1$ and $h > 0$ on $[0,1]$, all the analysis and results presented in this paper apply to such an extended version.

In one of his celebrated pioneer works in 1982, Weinberger [23] studied the long time (as $n \to \infty$) behavior and the existence of planar traveling waves for fully discrete Fisher's-type models of the form, for $\mathbf{u}^n := \{u_j^n\}_{j\in H}$,

$$\mathbf{u}^{n+1} - \mathbf{u}^n = Q[\mathbf{u}^n], \qquad n = 0, 1, 2, \ldots,$$

where Q is a translation invariant (e.g., autonomous) nonlinear operator and typical examples of H are $H = \mathbb{R}^m$ and $H = \mathbb{Z}^m$ ($m \geqslant 1$). In particular, for each unit vector ξ there exists a constant $c^*(\xi)$ (the minimal wave speed) such that $c^*(\xi)$ is the asymptotic propagation speed for arbitrarily initial disturbance. After deriving a lower and an upper bound for $c^*(\xi)$, the author established the existence of planar traveling wave with speed c for any $c \geq c^*(\xi)$, and nonexistence for $c < c^*(\xi)$. While Weinberger established striking results for an extremely general fully discrete monostable dynamics, here by contrast, we focus our attention only on a one-dimensional semidiscrete (i.e., continuous in time) version (1.1) or (1.2). Our main concerns in this paper are (1) the uniqueness and asymptotic behavior (as $x \to \infty$) of the traveling waves, and (2) the highly nontrivial extension of the current knowledge on nondegenerate monostable dynamics to its degenerate case, i.e., to the case $f'(0)f'(1) = 0$. That is to say, our work extends that of Weingerber's pioneer systematic analysis in two directions: firstly from the fully discrete version to semidiscrete version and, secondly, from nondegenerate steady states to general degenerate and/or nondegenerate steady states.

In the higher space dimensional case, the dynamics

$$u_t(x,t) = \sum_{i,j=1}^{m} a_{ij} \frac{\partial^2 u(x,t)}{\partial x_i \partial x_j} + f(u(x,t)), \qquad x \in \mathbb{R}^m, t > 0,$$

where $(a_{ij})_{m\times m}$ is a positive definite matrix, exhibits a variety of interesting wave phenomena; see, for example, Hamel and Nadirashvili [11], Berestycki and Larrouturou [3], and the references therein. A two-dimensional analogue of (1.1) takes the form

$$\dot{u}_{ij} = a[u_{i+1,j} + u_{i-1,j}] + b[u_{i,j+1} + u_{i,j-1}] + F(u_{ij}), \qquad i, j \in \mathbb{Z},$$

where a, b are positive constants. Here a planar traveling wave refers to a solution of the form $u_{ij}(t) = U(i\cos\theta + j\sin\theta + ct)$ for all $i, j \in \mathbb{Z}$ and $t \in \mathbb{R}$, where $(\cos\theta, \sin\theta)$ is the wave direction and $c = c(\theta)$ is the wave speed. Note that $U \in C^1(\mathbb{R})$ satisfies

$$cU'(\xi) = a[U(\xi + \cos\theta) + U(\xi - \cos\theta)] + b[U(\xi + \sin\theta) + U(\xi - \sin\theta)] + F(U(\xi)).$$

In this direction, we refer the reader to Chen [4], Chow, Mallet-Paret and Shen [7, 8] and Mallet-Paret [15, 16] for the bistable case and Shen [18, 19] for the bistable time almost periodic case. Clearly, our traveling wave problem is only the special case of $|\theta| = \frac{\pi}{4}$. We expect that our results and methods can be extended in a great extent to this new problem.

We remark that limit, as $a \searrow 0$, of the bistable nonlinearity $f(u) = u(1-u)(u-a)$ is the degenerate monostable nonlinearity $f(u) = u^2(1-u)$. The limiting process is continuous in the sense that the unique (modulo the translation invariance) traveling wave for the bistable nonlinearity approaches the unique minimum wave speed traveling wave for the degenerate monostable nonlinearity. The limiting process is not continuous in the sense that for the bistable case there is only one traveling wave, whereas for the monostable case, there are infinitely many traveling waves. We would like to point out that many tools that work for the bistable case do not work here for the monostable case; for example, in general the tools used for the construction of supersolutions in the bistable case do not work for the monostable case. Exaggerating a little bit, one may say that the bistable dynamics and monostable dynamics are different, and so are many of the mathematical tools to study them.

Now we briefly discuss our analysis towards our main results. The proof of uniqueness (Theorem 1) relies on the monotonicity (Theorem 2) and the detailed asymptotic behavior (Theorem 3) of wave profiles. Two new techniques are specifically developed here to study the uniqueness of traveling waves of monostable dynamics. One of them, which we call *magnification* and is originated from [6], is to magnify appropriately the difference between two wave profiles U and V by (for the purpose of demonstration only, considering the case $c > c_{\min}$)

$$W(\xi, x) = \int_{V(x)}^{U(x+\xi)} \frac{ds}{f(s)}.$$

Such a magnification has a special property $\lim_{x\to-\infty} W_x(\xi, x) = 0$ for any $\xi \in \mathbb{R}$ and a general property $\inf_{(\xi,x)\in\mathbb{R}^2} W_\xi(\xi, x) > 0$. From a basic comparison (for monotonic profiles) which says that if $U > V$ on $[a-1, a)\cup(b, b+1]$, then $U > V$ on $[a, b]$, these two properties prohibit W from any oscillations with nonvanishing magnitude as $x \to -\infty$; namely, there exists $\lim_{x\to-\infty} W(\xi, x)$ (which may be infinite). Consequently, any two wave profiles are ordered near $-\infty$; see section 4 for more details. An additional advantage of this magnification is that $\lim_{x\to-\infty} W(\xi, x)$ exists even if V is merely a sub- or a supersolution. This fact will be used in section 5 to find asymptotic expansions of wave profiles.

The other technique, which we call compression, is developed to include the treatment of the degenerate case $f'(1) = 0$. Traditionally near ∞ one uses $\min\{U + \varepsilon, 1\}$ as a supersolution which works for both monostable and bistable dynamics but needs the assumption that $f' \leqslant 0$ on $[1-\delta, 1]$ for some $\delta > 0$. To deal with the general case, we use the following compression to obtain (local) supersolutions:

$$Z(\ell, x) = U([1 + \ell]x), \qquad x \gg 1, \ell \in (0, 1].$$

The asymptotic behavior of wave profiles implies that Z approaches 1 as $x \to \infty$ at a rate faster than any wave profile. With a limiting $\ell \searrow 0$ process, we can show that near ∞, one wave profile is always bigger than a certain translation of any other wave profile.

The asymptotic behavior (1.5) follows from an analysis similar to that in [6]. After a thorough reinvestigation of the method used in [6], we found that the method in [6] can be rephrased into the following quite fundamental theory.

THEOREM 4. *Let $c > 0$ be a constant and $B(\cdot)$ be a continuous function having finite $B(\pm\infty) := \lim_{x\to\pm\infty} B(x)$. Let $z(\cdot)$ be a measurable function satisfying*

$$c\, z(x) = e^{\int_x^{x+1} z(s)ds} + e^{-\int_{x-1}^{x} z(s)ds} + B(x) \quad \forall\, x \in \mathbb{R}. \tag{1.9}$$

Then z is uniformly continuous and bounded. In addition, $\omega^\pm = \lim_{x\to\pm\infty} z(x)$ exist and are real roots of the characteristic equation $c\,\omega = e^\omega + e^{-\omega} + B(\pm\infty)$.

Note that each of $z = U'/U, U'/(U-1)$ and U''/U' satisfies an equation of the form (1.9). This theory provides a powerful tool to study the asymptotic behavior, as $x \to \pm\infty$, of positive solutions of a variety of semilinear finite difference-differential equations. In particular, once the monotonicity $U' > 0$ is shown, $z = U''/U'$ is then well defined and all the limits stated in Theorem 3 follow immediately from the theory.

Now the focus is shifted to show the monotonicity of U. In the nondegenerate case, $\mu < 0 < \lambda$, so that (1.5) and a comparison between $U(x + h)$ and $U(x)$ on a compact interval imply that $U' > 0$ on $\mathbb{R}$. In the degenerate case, $\lambda\mu = 0$, so (1.5) is not sufficient for such an argument. We shall develop a *blow-up* technique, showing

that $U' > 0$ on a sequence of intervals $\{[\xi_i - 1, \xi_i + 1]\}$ of two-unit length, where $\lim_{i\to\pm\infty} \xi_i = \pm\infty$. Then we develop a *modified sliding method* which enables us to compare $U(x + h)$ and $U(x)$ on any finite interval $[\xi_i - 1, \xi_j + 1]$ $(i < j)$ to prove the monotonicity result.

For a solution of (1.2) or (1.4) with initial value $u(x, 0)$, its long time behavior (e.g. approaching a traveling wave) depends on the asymptotic behavior of $u(x, 0)$ as $x \to -\infty$, i.e., tails of which wave profile $U(x)$ that $u(\cdot, 0)$ resembles; see, for example, [2, 5] and the references therein. For this purpose, we shall also provide asymptotic expansions, accurate enough to capture the translation difference of wave profiles near $\pm\infty$. In particular, under the condition that $f(u) = f'(0)u + O(u^{1+\alpha})$ for some $\alpha > 0$ and all small u, we show the following:

(i) If $c = c_{\min}$ and the larger root λ of (1.6) is not a double root, then for some $x_0 \in \mathbb{R}$,

$$\lim_{x\to-\infty} e^{-\lambda x} U(x + x_0) = 1. \tag{1.10}$$

(ii) If $c = c_{\min}$ and λ is a double root, then for some $x_0 \in \mathbb{R}$,

$$\text{either} \quad \lim_{x\to-\infty} \frac{U(x + x_0)}{|x|e^{\lambda x}} = 1 \quad \text{or} \quad \lim_{x\to-\infty} \frac{U(x + x_0)}{e^{\lambda x}} = 1. \tag{1.11}$$

(iii) If $c > c_{\min}$ and $f'(0) > 0$, then (1.10) holds for some $x_0 \in \mathbb{R}$ with λ the smaller root of (1.6).

Note that $\lambda > 0$ in all these cases, so, as we expected from (1.5), $U(x)$ decays to zero exponentially fast as $x \to -\infty$. Earlier results (e.g., [5, 10, 12, 27]) on this matter depend on the construction of global sub- and supersolution pairs that sandwich a wave profile. Such a construction is possible for all large wave speeds for general f and for all nonminimum wave speeds when $f(s) \leqslant f'(0)s$ for all $s \in [0, 1]$. We remark that the stability (which implies uniqueness) result in [5] was established under the assumption (1.10). By proving (1.10), the result in [5] then implies that any solution of (1.2) approaches, as $t \to \infty$, a traveling wave of speed c $(> c_{\min})$ if $u(\cdot, 0)$ takes values on $[0, 1]$ and

$$\lim_{x\to-\infty} e^{-\lambda x} u(x, 0) = 1, \qquad \liminf_{x\to\infty} u(x, 0) > 0.$$

On the other hand, $\lambda = 0$ when $f'(0) = 0$ and $c > c_{\min}$, so from (1.5), an exponential decay is impossible and an algebraic decay is to be expected (cf. [13, 14, 17, 21] for (1.4)). Indeed, under certain additional assumptions (cf. (B1) in section 5) we show the following:

If $c > c_{\min}$ and $f'(0) = 0$, then for some $x_0 \in \mathbb{R}$,

$$\lim_{x\to-\infty} \left\{ \int_{1/2}^{U(x)} \frac{ds}{f(s)[1 + f'(s)/c^2]} - \frac{x + x_0}{c} \right\} = 0. \tag{1.12}$$

For example, when $f(u) = \kappa u^2(1 - u)^p$ $(\kappa > 0,\ p \geq 1)$, the above limit yields

$$U(x) = \frac{c}{\kappa[|x| - x_0 + o(1)] + (pc - 2\kappa/c) \ln |x|} \quad \text{as } x \to -\infty.$$

The asymptotic expansion of $U(x)$ as $x \to \infty$ can be treated similarly. Indeed,

$$\lim_{x\to\infty} \left\{ \int_{1/2}^{U(x)} \frac{ds}{f(s)[1 + f'(s)/c^2]} - \frac{x + x_0}{\nu} \right\} = 0,$$

for some $x_0 \in \mathbb{R}$, where $\nu = c$ if $f'(1) = 0$ and $\nu = f'(1)/\mu$ if $f'(1) < 0$. Since this limiting behavior has nothing to do with the condition needed on the initial data for the long time behavior of solutions of (1.2), we choose to omit the details here.

This paper is organized as follows. In section 2, we derive the asymptotic behavior of wave profiles near $\pm\infty$ and prove Theorem 3. We prove the monotonicity of wave profiles (Theorem 2) in section 3, by using the method of sliding and a new blow-up technique. In section 4, the uniqueness of traveling waves is established. Finally in section 5, we construct suitable local super/subsolutions to verify our asymptotic expansions of wave profiles near $x = \pm\infty$.

2. Asymptotic behavior of wave profiles near $x = \pm\infty$. In the following, the assumption (A) is always assumed.

2.1. The idea in [6]. The most important technique developed in [6] can be presented as follows. Suppose that the following quantities

$$\rho(x) := \frac{U'(x)}{U(x)}, \quad \sigma(x) := \frac{U'(x)}{U(x)-1}, \quad \chi(x) := \frac{U''(x)}{U'(x)}$$

are well defined. This is the case, if $U > 0$, $U < 1$, and $U' > 0$ for ρ, σ, and χ, respectively. Then each of them satisfies an equation of the form (1.9), where $B(\cdot)$ is a continuous function having $\lim_{x\to\pm\infty} B(x) =: B(\pm\infty)$. For any positive constant m, we set

$$v(x) = e^{mx+\int_0^x z(s)ds}.$$

Then

$$c\,v'(x) = [cm + B(x)]v(x) + e^{-m}v(x+1) + e^{m}v(x-1).$$

Assume that $c > 0$. We take a specific $m = \|B(x)\|_{L^\infty(\mathbb{R})}/c$. Then $v'(x) \geq 0$. Consequently,

$$c\,v(x) - c\,v(x-1/2) > \int_{x-1/2}^{x} e^{-m}v(s+1)ds > \frac{1}{2}v(x+1/2)e^{-m}.$$

This implies that $v(x) > v(x+1/2)/(2ce^m) > v(x+1)/(2ce^m)^2$. Therefore,

$$e^{\int_x^{x+1} z(s)ds} = \frac{v(x+1)e^{-m}}{v(x)} \leqslant 4c^2e^m, \qquad e^{-\int_{x-1}^{x} z(s)ds} = \frac{e^m v(x-1)}{v(x)} \leqslant e^m,$$

and so

$$-m < z(x) < m + 4ce^m + e^m/c \quad \forall\, x \in \mathbb{R}, \quad m := \|B\|_{L^\infty(\mathbb{R})}/c. \tag{2.1}$$

The uniform boundedness of z implies that z is uniformly continuous. Hence, for any unbounded sequence $\{x_i\}$, $\{z(x_i+\cdot)\}$ is a bounded and equicontinuous family. Along a subsequence, it converges to a limit r, uniformly in any compact subset of $\mathbb{R}$. In addition, r satisfies the *fundamental equation*

$$c\,r(x) = e^{\int_x^{x+1} r(s)\,ds} + e^{\int_x^{x-1} r(s)\,ds} + b \quad \forall\, x \in \mathbb{R}, \tag{2.2}$$

where $b = B(\infty)$ if $\lim_{i\to\infty} x_i = \infty$ and $b = B(-\infty)$ if $\lim_{i\to\infty} x_i = -\infty$. For the fundamental equation, Chen and Guo established in [6] the following key result.

PROPOSITION 2. *Let $c > 0$, $b \in \mathbb{R}$ and $P(\omega) = c\omega - e^{\omega} - e^{-\omega} - b$. Consider* (2.2).

(i) *When $P(\omega) = 0$ has no real root, there is no solution.*

(ii) *When $P(\omega) = 0$ has only one real root λ, $r \equiv \lambda$ is the only solution.*

(iii) *When $P(\omega) = 0$ has two real roots $\{\lambda, \Lambda\}$ $(\lambda < \Lambda)$, every solution can be written as*

$$r(x) = \frac{u'(x)}{u(x)}, \quad u(x) = \theta e^{\lambda x} + (1-\theta)e^{\Lambda x}, \quad \theta \in [0,1].$$

In particular, any nonconstant solution satisfies $r' > 0$, $r(-\infty) = \lambda$, and $r(\infty) = \Lambda$.

Proof of Theorem 4. We need consider only the case when the characteristic equation has two real roots. For this, let λ and Λ be the roots where $\lambda < \Lambda$. Suppose $\lim_{x\to-\infty} z(x)$ does not exist. Then there exist $\omega \notin \{\lambda, \Lambda\}$ and a sequence $\{x_i\}$ satisfying $\lim_{i\to-\infty} x_i = -\infty$, $z(x_i) = \omega$ and $z'(x_i) \leqslant 0$ for all i. Since $\{z(x_i + \cdot)\}$ is uniformly bounded and equi-continuous, a subsequence converges to a limit r which solves (2.2) with $b = B(-\infty)$. In addition, by the definition of r, we have $r(0) = \omega$ and $r'(0) \leqslant 0$. But from Proposition 2, there are no such kind of solutions. Hence, $\lim_{x\to-\infty} z(x)$ exists and is one of the two roots to the characteristic equation. Similarly, one can show that $\lim_{x\to\infty} z(x)$ exists. □

Remark 1.

(i) By working on the function $\hat{z}(x) := -z(-x)$ the assertion of the theorem remains unchanged when $c < 0$.

(ii) Theorem 4 extends to a more general equation

$$z(x) = a_1(x)e^{\int_x^{x+1} z(s)ds} + a_2(x)e^{-\int_{x-1}^{x} z(s)ds} + B(x),$$

where a_1 and a_2 are continuous positive functions having limits

$$a^{\pm} := \lim_{x\to\pm\infty} a_1(x) = \lim_{x\to\pm\infty} a_2(x) > 0.$$

(iii) Theorem 4 also extends to the case when z is a continuous function defined on $[-1, \infty)$ (or $(-\infty, 1]$) and satisfies (1.9) on $[0, \infty)$ (or $(-\infty, 0]$). The conclusion is that $\lim_{x\to\infty} z(x)$ (or $\lim_{x\to-\infty} z(x)$) exists and is the root of the characteristic equation.

2.2. The asymptotic behavior. Now we establish the limits stated in Theorem 3.

We begin with the limits in (1.5). First we show that $U > 0$. Suppose on the contrary there exists $y \in \mathbb{R}$ such that $U(y) = 0$. Then it is a global minimum so that $U'(y) = 0$ and from the equation in (1.3), $U(y+1) + U(y-1) = 0$ which implies that $U(y \pm 1) = 0$. An induction gives $U(y+k) = 0$ for all $k \in \mathbb{Z}$, contradicting $U(\infty) = 1$. Thus, $U > 0$. Similarly, $U < 1$. Once we know $0 < U < 1$, we can define

$$\rho(x) := \frac{U'(x)}{U(x)} \quad \Rightarrow \quad \int_x^{x+1} \rho(z)dz = \ln\frac{U(x+1)}{U(x)},$$
$$\sigma(x) := \frac{U'(x)}{U(x)-1} \quad \Rightarrow \quad \int_x^{x+1} \sigma(z)dz = \ln\frac{U(x+1)-1}{U(x)-1}.$$

Dividing the ode in (1.3) by U and $U-1$, respectively, we obtain

$$c\rho(x) = e^{\int_x^{x+1}\rho(z)dz} + e^{\int_x^{x-1}\rho(z)dz} - 2 + B_1(x),$$
$$c\sigma(x) = e^{\int_x^{x+1}\sigma(s)\,ds} + e^{\int_x^{x-1}\sigma(s)\,ds} - 2 + B_2(x),$$

where $B_1(x) = f(U(x))/U(x)$ and $B_2(x) = f(U(x))/[U(x)-1]$. Since $U(-\infty) = 0$ and $U(\infty) = 1$, we see that $B_1(-\infty) = f'(0)$, $B_1(\infty) = 0$, $B_2(-\infty) = 0$, and $B_2(\infty) = f'(1)$. The limits in (1.5) thus follow from Theorem 4.

Next, we establish the remaining limits stated in Theorem 3. Here we shall use the fact $U' > 0$, to be proven in the next section. Differentiating the ode in (1.3) with respect to x we have

$$c\,U''(x) = U'(x+1) + U'(x-1) + [f'(U(x)) - 2]U'(x).$$

Define

$$\chi(x) := \frac{U''(x)}{U'(x)} \quad \Rightarrow \quad \int_x^{x+1} \chi(z)dz = \ln \frac{U'(x+1)}{U'(x)}.$$

Then

$$c\,\chi(x) = e^{\int_x^{x+1}\chi(z)dz} + e^{-\int_{x-1}^{x}\chi(z)dz} + f'(U(x)) - 2 \quad \forall\, x \in \mathbb{R}.$$

The stated limits for χ in Theorem 3 thus follow from Theorem 4 and l'Hôpital's rule.

Finally, the limits of $f(U(x))/U'(x)$ as $x \to \pm\infty$ are obtained by using the limits of χ and the identity

$$\begin{aligned}\frac{f(U(x))}{U'(x)} &= c - \frac{[U(x+1) - U(x)] - [U(x) - U(x-1)]}{U'(x)} \\ &= c - \int_0^1 \Big\{ e^{\int_x^{x+z}\chi(s)ds} - e^{-\int_{x-z}^{x}\chi(s)ds} \Big\} dz.\end{aligned}$$

In the next two subsections, we show the additional part of Theorem 3; namely, we show that λ is the smaller real root to the characteristic equation (1.6) when $c > c_{\min}$ and the larger root when $c = c_{\min}$.

2.3. The characteristic values of nonminimum speed waves.

LEMMA 2.1. *If (c, U) is a traveling wave of speed $c > c_{\min}$, then the characteristic equation $c\lambda = e^{\lambda} + e^{-\lambda} - 2 + f'(0)$ has two different real roots and $\lambda := \lim_{x\to-\infty} U'(x)/U(x)$ is the smaller root. In the particular instance when $f'(0) = 0$, $\lim_{x\to-\infty} U'(x)/U(x) = 0$.*

Proof. Recall from Theorem 2 of [6] that $c_{\min} \geqslant c_*$, where

$$c_* := \min_{z>0} \frac{e^z + e^{-z} - 2 + f'(0)}{z}.$$

Hence $c_{\min} z = e^z + e^{-z} - 2 + f'(0)$ always has a root. This implies that $c\,z = e^z + e^{-z} - 2 + f'(0)$ has exactly two roots, which we denote by $\lambda(c)$ and $\Lambda(c)$ with $\lambda(c) < \Lambda(c)$, for $c > c_{\min}$.

Suppose on the contrary that $\lim_{x\to-\infty} U'(x)/U(x) = \Lambda(c)$. Let $\hat{c} \in (c_{\min}, c)$ and $(\hat{c}, \hat{U})$ be a traveling wave of speed $\hat{c}$. By (1.5), $\lim_{x\to-\infty} \hat{U}'(x)/\hat{U}(x) \leqslant \Lambda(\hat{c})$. Then

$$\lim_{x\to-\infty} \frac{d}{dx}\Big(\ln \frac{\hat{U}(x)}{U(x)} \Big) = \lim_{x\to-\infty} \Big\{ \frac{\hat{U}'(x)}{\hat{U}(x)} - \frac{U'(x)}{U(x)} \Big\} \leqslant \Lambda(\hat{c}) - \Lambda(c) < 0$$

by the strictly monotonicity of $\Lambda(c)$ in c. Thus, $\lim_{x\to-\infty}\ln[\hat{U}(x)/U(x)]=\infty$ and there exists $M>0$ such that $\hat{U}(x)>U(x)$ for all $x\leqslant -M$. Similarly,

$$\lim_{x\to\infty}\frac{d}{dx}\left\{\int_{U(x)}^{\hat{U}(x)}\frac{ds}{f(s)}\right\}=\lim_{x\to\infty}\left\{\frac{\hat{U}'(x)}{f(\hat{U}(x))}-\frac{U'(x)}{f(U(x))}\right\}$$
$$=\begin{cases}1/\hat{c}-1/c & \text{if } f'(1)=0,\\ [\mu(\hat{c})-\mu(c)]/f'(1) & \text{if } f'(1)<0.\end{cases}$$

This quantity is positive when $f'(1)=0$; so is the case when $f'(1)<0$ since the negative root $\mu=\mu(c)$ of $c\mu=e^{\mu}+e^{-\mu}-2+f'(1)$ satisfies $\mu(\hat{c})<\mu(c)$. Thus there exists $M_1>0$ such that $\hat{U}(x)>U(x)$ for all $x\geqslant M_1$. In conclusion, $\hat{U}(\cdot+M_1)>U(\cdot-M)$.

Now both $u_1(x,t):=\hat{U}(x+M_1+\hat{c}t)$ and $u_2(x,t):=U(x-M+ct)$ are solutions of (1.2). Since $u_1(\cdot,0)\geqslant u_2(\cdot,0)$, the comparison principle for (1.2) implies $u_1(\cdot,t)\geqslant u_2(\cdot,t)$ for all $t>0$, which is impossible since $c>\hat{c}$. Thus, $\lim_{x\to-\infty}U'(x)/U(x)=\lambda(c)$. □

The asymptotic behavior of U stated in Theorem 3 immediately gives the following corollary.

COROLLARY 2.2. *Suppose (c_1,U_1) and (c_2,U_2) are two traveling waves where $c_1<c_2$. Then there exist $a,b\in\mathbb{R}$ such that*

$$U_1<U_2 \quad in \quad (-\infty,a), \qquad U_1>U_2 \quad in \quad (b,\infty).$$

We remark that in the case of the differential equation $cU'=U''+f(U)$ one can take $a=b$ to conclude that a smaller speed wave profile is steeper than a larger speed wave profile; namely, on the phase plane (U,U'), if one writes $U'=P(c,U)$, then $P(c_1,s)>P(c_2,s)$ for all $s\in(0,1)$ and $c_2>c_1\geqslant c_{\min}$. For (1.3), we believe that this should also be the case.

2.4. The characteristic value of minimum speed waves.

LEMMA 2.3. *If $(c_{\min},U)$ is a wave of minimum speed, then $\Lambda:=\lim_{x\to-\infty}U'(x)/U(x)$ is the larger root (if there are two) of the characteristic equation $c_{\min}z=e^z+e^{-z}-2+f'(0)$.*

Proof. Notice that when $c_{\min}=c_*$ (defined in (1.8)), the characteristic equation has only one real root, so there is nothing to prove in this case. Hence we consider the case when $c_{\min}>c_*$. We denote the smaller real root by λ and the larger root by Λ. We use a contradiction argument by assuming that $\lim_{x\to-\infty}U'(x)/U(x)=\lambda$. As we shall see, this will allow us to construct a supersolution Φ of wave speed c for some $c<c_{\min}$ by joining an exponential function ψ defined on $(-\infty,0]$ and another function ϕ defined on $[0,\infty)$ obtained from the wave profile U of speed $c_{\min}$. We divide this construction into the following steps.

First, set $\omega=(\lambda+\Lambda)/2$ and $\delta:=c_{\min}\,\omega-e^{\omega}-e^{-\omega}+2-f'(0)$. Then $\delta>0$ since the function $P(z):=c_{\min}\,z-e^z-e^{-z}+2-f'(0)$ is concave and vanishes at λ and Λ. Also by translation, we can assume that $U(0)$ is so small that

$$\sup_{0<s\leqslant U(0)e^{\omega}}\left|\frac{f(s)}{s}-f'(0)\right|<\frac{\delta}{2},\qquad \sup_{x\leqslant 1}\frac{U'(x)}{U(x)}<\omega\ .$$

Set $\psi(x)=U(0)e^{\omega x}$. For every $c\in[c_{\min}-\delta/(2\omega),c_{\min}]$,

$$\mathcal{L}\psi(x):=c\,\psi'(x)-\psi(x+1)-\psi(x-1)+2\psi(x)-f(\psi(x))$$
$$=\psi(x)\left\{c\,\omega-e^{\omega}-e^{-\omega}+2-\frac{f(\psi(x))}{\psi(x)}\right\}>0\quad\forall\ x\leq 1.$$

Next, we construct $\phi(c,\cdot)$, to be used as the supersolution defined on $[0,\infty)$. For each $c \in (0, c_{\min}]$, consider the equation $\phi = \mathbf{T}^c\phi$ on $\mathbb{R}$, where

$$\mathbf{T}^c\phi := \begin{cases} e^{-mx/c}\{U(0) + c\int_0^x e^{mz/c} W[m,\phi](z)dz\} & \text{if } x \geqslant 0, \\ U(x) & \text{if } x < 0, \end{cases}$$
$$W[m,\phi](z) := \phi(z+1) + \phi(z-1) + [m-2]\phi(z) + f(\phi(z)).$$

Following [6], a solution can be obtained as follows. Define $\{\phi_n\}_{n=0}^{\infty}$ by

$$\phi_0(c,\cdot) \equiv \mathbf{1}, \qquad \phi_{n+1}(c,\cdot) := \mathbf{T}^c\phi_n(c,\cdot) \quad \forall\, n \in \mathbb{N}.$$

Note that $\mathbf{T}^c$ is a monotonic operator: $\psi_1 \leqslant \psi_2 \Rightarrow \mathbf{T}^c\psi_1 \leqslant \mathbf{T}^c\psi_2$. It follows that $\phi_{n+1} \leqslant \phi_n \leqslant \mathbf{1}$. In addition, since

$$c\,(e^{mx/c}U)' - e^{mx/c}W[m,U] = (c - c_{\min})U'e^{\mu x/c} \leqslant 0,$$

integrating this inequality over $[0,x]$ gives $U \leqslant \mathbf{T}^cU$. This implies that $\phi_n \geqslant U$ for all n. Consequently, $\phi(c,\cdot) := \lim_{n\to\infty}\phi_n$ exists and is a solution to $\phi = \mathbf{T}^c\phi$. It is easy to see that $U \leqslant \phi < \mathbf{1}$ on $[0,\infty)$, $\phi(c,0) = U(0)$, and

$$c\,\phi'(c,x) = \phi(c,x+1) + \phi(c,x-1) - 2\phi(c,x) + f(\phi(c,x)) \quad \forall\, x > 0.$$

This equation implies, for $0 < c_1 < c_2 \leqslant c_{\min}$, that $\phi(c_2,\cdot) \leqslant \mathbf{T}^{c_1}\phi(c_2,\cdot)$, so that $\phi_n(c_1,\cdot) \geqslant \phi(c_2,\cdot)$ for all n and $\phi(c_1,\cdot) > \phi(c_2,\cdot)$ on $(0,\infty)$. Following an idea in [6] or the technique for the uniqueness of U presented in this paper (section 4), one can further show that $\phi(c,\cdot)$ is unique. The uniqueness implies that $\phi(c,\cdot)$ is continuous in c and $\phi(c_{\min},\cdot) \equiv U$. Therefore, $\lim_{c\to c_{\min}}\phi(c,\cdot) = U$ in $C^1([0,\infty))$. This further implies that

$$\lim_{c\to c_{\min}} \frac{\phi'(c,x)}{\phi(c,x)} = \frac{U'(x)}{U(x)} \quad \text{uniformly for } \ x \in [0,1].$$

Finally, let $c \in [c_{\min} - \delta/(2\omega), c_{\min})$ be such that

$$\max_{x\in[0,1]} \frac{\phi'(c,x)}{\phi(c,x)} < \omega.$$

We define

$$\Phi(x) = \begin{cases} \psi(x) & \text{if } \ x \leqslant 0, \\ \phi(c,x) & \text{if } \ x > 0. \end{cases}$$

Since $\psi(0) = U(0) = \phi(c,0)$ and

$$\frac{\psi'(x)}{\psi(x)} = \omega > \frac{\phi'(c,x)}{\phi(c,x)} \quad \forall x \in (-\infty,0)\cup(0,1],$$

$\phi < \psi$ in $(0,1]$ and $\psi < \phi \equiv U$ in $(-\infty,0)$. That is,

$$\Phi = \min\{\phi\ ,\ \psi\} \qquad \text{on } \ (-\infty,1].$$

Consequently, considering separately $x \in (-\infty,0)$, $(0,1]$ and $(1,\infty)$, we see that

$$c\,\Phi'(x) \geqslant \Phi(x+1) + \Phi(x-1) - 2\Phi(x) + f(\Phi(x)) \quad \forall\ x \in (-\infty,0)\cup(0,\infty);$$

that is, Φ is a supersolution of wave speed c.

Thus, by Proposition 1(iii), there is a traveling wave of speed c for some $c < c_{\min}$, contradicting the minimality of $c_{\min}$. This proves the lemma. □

Remark 2. If $f'(\cdot) \leqslant 0$ on $[1-\delta, 1]$ for some $\delta > 0$, then a constructive proof of Lemma 2.3 can be obtained by taking

$$\Phi(x) = [U(0) + \epsilon]e^{\omega x} \quad \forall\, x \leqslant 0, \qquad \Phi(x) = U(x + \varepsilon - \varepsilon e^{-kx}) + \epsilon \quad \forall\, x > 0,$$

where $0 < \epsilon \ll \varepsilon \ll U(0) \ll 1 \ll k$. We leave the verification to the interested reader.

3. Monotonicity of wave profiles. This section is dedicated to the proof of the monotonicity of any wave profile U. We point out here that the limits in (1.5) are established without the knowledge of the monotonicity of U so that we can use them here.

3.1. The method of sliding. This traditional method is to compare $U(\cdot + \tau)$ and $U(\cdot)$ by decreasing τ continuously from a large value down to zero, namely, to show that

$$\inf \{\tau > 0 \mid U(\cdot + \tau) > U(\cdot) \quad \text{on} \mathbb{R}\} = 0. \tag{3.1}$$

This implies $U' \geqslant 0$, and from an integral equation, $U' > 0$ on $\mathbb{R}$. If we know $U' > 0$ near $x = \pm\infty$ (e.g., by (1.5) for the case $\mu < 0 < \lambda$), then (3.1) follows easily from a comparison principle (cf. [6]). When $f'(0) = 0$, it is very difficult to show directly that $U' > 0$ in a vicinity of $x = -\infty$. Similar difficulty occurs near $x = \infty$ when $f'(1) = 0$. To overcome this difficulty, we use a modification of the method, stated in the third part of the following lemma.

LEMMA 3.1.

(i) *If* $[a, b]$ *is an interval on which* $U' \leqslant 0$, *then* $b - a < 1$.

(ii) *If* $U' > 0$ *on* $[\xi, \xi + 1]$, *then* $U(\xi) < U(x)$ *for all* $x > \xi$.

(iii) *If* $U' > 0$ *on* $[\xi - 1, \xi + 1] \cup [\eta - 1, \eta + 1]$, *where* $\xi < \eta$, *then* $U' > 0$ *on* $[\xi, \eta]$.

Proof.

(i) Let $[a, b]$ be an interval on which $U' \leqslant 0$. We want to show that $b - a < 1$. Suppose otherwise $b - a \geqslant 1$. Let $\hat{x} \in [b, \infty)$ be a point such that $U(\hat{x}) \leqslant U(x)$ for all $x \geqslant b$. Then $\hat{x}$ is a global minimum of U restricted on $[a, \infty)$, since $U' \leqslant 0$ on $[a, b]$. This leads to the following contradiction:

$$0 = c\,U'(\hat{x}) = U(\hat{x} + 1) + U(\hat{x} - 1) - 2U(\hat{x}) + f(U(\hat{x})) \geqslant f(U(\hat{x})) > 0.$$

(ii) Assume that $U' > 0$ on $[\xi, \xi + 1]$. Let $\hat{x} \geqslant \xi + 1$ be a point such that $U(\hat{x}) \leqslant U(x)$ for all $x \geqslant \xi + 1$. Then $U(\xi) < U(\hat{x})$ since otherwise $\hat{x} \geqslant \xi + 1$ is a point of global minimum of U on $[\xi, \infty)$ and the same contradiction as above arises. Thus $U(\xi) < U(x)$ for all $x > \xi$.

(iii) Assume that $U' > 0$ on $[\xi - 1, \xi + 1] \cup [\eta - 1, \eta + 1]$, where $\xi < \eta$. By the second assertion, $U(\eta) > U(\xi)$ so that we can define

$$\tau^* := \inf \{\, \tau \in (0, \eta - \xi] \mid U(\cdot) < U(\cdot + \tau) \quad \text{on} [\xi, \eta - \tau]\}.$$

Clearly, $\tau^* \in [0, \eta - \xi)$. We claim that $\tau^* = 0$. Suppose on the contrary that $\tau^* > 0$. Then there exists $\hat{x} \in [\xi, \eta - \tau^*]$ such that

$$U(\hat{x} + \tau^*) - U(\hat{x}) = 0 \leqslant U(x + \tau^*) - U(x) \quad \forall\, x \in [\xi, \eta - \tau^*].$$

For $x \in [\xi-1,\xi]$: (1) if $x+\tau^* \le \xi$, then $U(x+\tau^*)-U(x)>0$ since $U'>0$ on $[\xi-1,\xi]$; (2) if $x+\tau^*>\xi$, by the second assertion, $U(x+\tau^*)>U(\xi)\ge U(x)$. Thus $U(x+\tau^*)>U(x)$ for all $x\in[\xi-1,\xi]$. Similarly, $U(x+\tau^*)>U(x)$ for all $x\in[\eta-\tau^*,\eta-\tau^*+1]$. Hence,

$$U(\hat{x}+\tau^*)-U(\hat{x})=0\leqslant U(x+\tau^*)-U(x)\quad \forall\, x\in[\xi-1,\eta-\tau^*+1].$$

Consequently, $U'(\hat{x}+\tau^*)=U'(\hat{x})$. Using the equation for U, we conclude that

$$U(\hat{x}+\tau^*+1)+U(\hat{x}+\tau^*-1)=U(\hat{x}+1)+U(\hat{x}-1).$$

Since $U(\cdot+\tau^*)\geqslant U(\cdot)$ on $[\xi-1,\eta-\tau^*+1]$, we see that $U(\hat{x}+\tau^*\pm1)=U(\hat{x}\pm1)$. By induction, $U(\hat{x}+\tau^*+k)=U(\hat{x}+k)$ for all integer k satisfying $\hat{x}+k\in[\xi-1,\eta-\tau^*+1]$. But this is impossible since $U(x+\tau^*)>U(x)$ for all $x\in[\xi-1,\xi]$. Thus, $\tau^*=0$.

That $\tau^*=0$ implies $U(\cdot+\tau)>U(\cdot)$ on $[\xi,\eta-\tau]$ along a sequence $\tau\searrow 0$. In particular, $U'(x)\geqslant 0$ on $[\xi,\eta]$. Finally, for $m=\max_{0\leqslant s\leqslant 1}|2-f'(s)|$ and every $x\in[\xi,\eta]$,

$$cU''(x)=U'(x+1)+U'(x-1)+[f'(U)-2]U'(x)\geqslant -mU'(x).$$

It follows that $(U'(x)e^{mx/c})'\geqslant 0$ or $U'(x)e^{mx/c}\geqslant U'(\xi)e^{m\xi/c}>0$ for all $x\in[\xi,\eta]$. □

3.2. A linear equation from blow-up. To show that $U'>0$ on $\mathbb{R}$, we use Lemma 3.1(iii). For this, we need only to find a sequence $\{[\xi_j-1,\xi_j+1]\}$ of intervals on which $U'>0$. To do this, we shall use a blow-up technique for the functions $\rho=U'/U$ and $\sigma=U'/(U-1)$, leading to the following two linear problems:

$$\begin{cases} cR'(x)=R(x+1)+R(x-1)-2R(x) & \forall\, x\le 1,\\ |R|\le 1 \quad \text{on} \quad (-\infty,2], \qquad |R(0)|=1; \end{cases} \tag{3.2}$$

$$\begin{cases} cR'(x)=R(x+1)+R(x-1)-2R(x) & \forall\, x\ge -1,\\ |R|\le 1 \quad \text{on}[-2,\infty), \qquad |R(0)|=1. \end{cases} \tag{3.3}$$

LEMMA 3.2.

(i) *If R solves* (3.2)*, then $|R|>1/2$ on $[A-1,A+1]$ for some $A>0$.*

(ii) *Any solution of* (3.3) *satisfies $|R|>1/2$ on $[A-1,A+1]$ for some $A>0$.*

Proof.

(i) Suppose R solves (3.2). Then $|R'|\leqslant 4/c$ on $(-\infty,1]$. Set $z(x):=R'(x)/[R(x)+2]$. Dividing the ode in (3.2) by $R(x)+2$ we obtain

$$c\,z(x)=e^{\int_x^{x+1}z(t)dt}+e^{-\int_{x-1}^{x}z(t)dt}-2, \qquad |z(x)|\le 4/c \quad \forall\, x\le 1.$$

Following the argument used in the previous section, we conclude that $\lim_{x\to-\infty}z(x)$ exists. Since R is bounded, $\liminf_{x\to-\infty}|R'(x)|=0$. Thus, $\lim_{x\to-\infty}z(x)=0$, which implies that $\lim_{x\to-\infty}R'(x)=0$.

As $R(0)$ is a global extremum of R restricted on $(-\infty,1]$, $R(j)=R(0)$ for all integer $j\le 1$. Upon using $\lim_{x\to-\infty}R'(x)=0$, we derive that $\lim_{x\to-\infty}R(x)=R(0)$. Since $|R(0)|=1$, there exists $A>0$ such that $|R(\cdot)|>1/2$ on $[A-1,A+1]$. This proves the first assertion (i).

(ii) The proof of the second assertion (ii) is analogous to the case (i) and therefore is omitted. □

3.3. The monotonicity of wave profile. That $U' > 0$ follows from Lemma 3.1(iii) and the following lemma.

LEMMA 3.3. *There exists a sequence* $\{\xi_i\}_{i\in\mathbb{Z}}$ *such that* $U' > 0$ *on* $[\xi_i - 1, \xi_i + 1]$ *for each* $i \in \mathbb{Z}$ *and* $\lim_{i\to\pm\infty} \xi_i = \pm\infty$.

Proof. The sequence $\{\xi_i\}_{i\leqslant 0}$: Here we construct the sequence such that $U' > 0$ on $\cup_{i\leqslant 0}[\xi_i - 1, \xi_i + 1]$ and $\lim_{i\to\infty} \xi_i = -\infty$.

When $f'(0) > 0$, $\lim_{x\to-\infty} U'(x)/U(x) = \lambda > 0$ so $U'(x) > 0$ for all $x \ll -1$. Hence, we need consider only the case $f'(0) = 0$ and $\lim_{x\to-\infty} \rho(x) = 0$, where $\rho(x) = U'(x)/U(x)$. Define

$$\varepsilon_j = \max_{x\leq j} |\rho(x)| \ \forall\ j < 0, \qquad \theta = \limsup_{j\to-\infty} \frac{\varepsilon_{j-3}}{\varepsilon_j} \in [0, 1].$$

We claim that $\theta = 1$. Suppose not. Then, for $\hat\theta = (1 + \theta)/2$, there exists $J < 0$ such that $\varepsilon_{j-3} \leq \hat\theta\varepsilon_j$ for all $j \leq J$. Hence, $\varepsilon_{J-3k} \leq \varepsilon_J\hat\theta^k$ for every integer $k \geq 0$. Consequently, $|\rho(x)| \leq \varepsilon_J\hat\theta^{(J-x)/3-1}$ for all $x \leq J$. For $y < J$,

$$\ln\frac{U(J)}{U(y)} = \int_y^J \rho(x)dx \leqslant \int_y^J \varepsilon_J\hat\theta^{(J-x)/3-1}dx \leqslant \frac{3\varepsilon_J}{|\hat\theta\ln\hat\theta|}.$$

Sending $y \to -\infty$ we obtain a contradiction. Hence $\theta = 1$.

Let $\{j_k\}_{k=1}^\infty$ be a sequence such that $\lim_{k\to\infty} j_k = -\infty$ and $\lim_{k\to\infty} \varepsilon_{j_k-3}/\varepsilon_{j_k} = 1$. Let $x_k \leqslant j_k-3$ be a point such that $|\rho(x_k)| = \varepsilon_{j_k-3}$. Define $\rho_k(x) := \rho(x_k+x)/|\rho(x_k)|$. Then $\max_{x\leq 3} |\rho_k(x)| \leq \varepsilon_{j_k}/\varepsilon_{j_k-3}$, $|\rho_k(0)| = 1$, and

$$\begin{aligned} c\,\rho_k'(x) = {}& [\rho_k(x+1) - \rho_k(x)]e^{\rho(x_k)\int_x^{x+1}\rho_k(z)dz} \\ & + [\rho_k(x-1) - \rho_k(x)]e^{-\rho(x_k)\int_{x-1}^{x}\rho_k(z)dz} + \rho_k(x)f_1(U(x_k + x)), \end{aligned}$$

where $f_1(s) = f'(s) - f(s)/s \to 0$ as $s \searrow 0$. This equation implies that $\{\rho_k\}_{k=1}^\infty$ is a family of bounded and equicontinuous functions on $(-\infty, 2]$. Hence, a subsequence which we still denote by $\{\rho_k\}$ converges to a limit R, uniformly in any compact subset of $(-\infty, 2]$. Clearly, R satisfies (3.2).

By Lemma 3.2(i), there exists a constant $A < 0$ such that either $R \geqslant 1/2$ on $[A - 1, A + 1]$ or $R \leqslant -1/2$ on $[A - 1, A + 1]$. As $\lim_{k\to\infty} \rho_k \to R$ on $[A - 1, A + 1]$, there exists an integer $K > 0$ such that for every integer $k \geqslant K$, either $\rho_k > 0$ on $[A - 1, A + 1]$ or $\rho_k < 0$ on $[A - 1, A + 1]$. By Lemma 3.1(i), the latter case is impossible. Thus $\rho_k > 0$ on $[A - 1, A + 1]$, i.e., $U' > 0$ on $[x_k + A - 1, x_k + A + 1]$. Define $\xi_i = A + x_{K+|i|}$ for all integer $i \leq 0$. Then $\lim_{i\to-\infty} \xi_i = -\infty$ and $U' > 0$ on $[\xi_i - 1, \xi_i + 1]$ for every integer $i \leq 0$.

The sequence $\{\xi_i\}_{i\geqslant 1}$: When $f'(1) < 0$, we have $\lim_{x\to\infty} U'(x)/[1 - U(x)] > 0$ so $U'(x) > 0$ for all $x \gg 1$. It remains to consider the case $f'(1) = 0$. Define

$$\sigma(x) = \frac{U'(x)}{U(x) - 1}, \quad \delta_j = \max_{x\in[j,\infty)} |\sigma(x)|, \quad \theta = \limsup_{j\to\infty} \frac{\delta_{j+3}}{\delta_j} \ \in [0, 1].$$

With an analogous argument as before, we can show that $\theta = 1$. Take a sequence $\{j_k\}_{k=1}^\infty$ satisfying $\lim_{k\to\infty} j_k = \infty$ and $\lim_{k\to\infty} \delta_{j_k+3}/\delta_{j_k} = 1$. Let $x_k \geq j_k + 3$ be a point such that $\delta_{j_k+3} = |\sigma(x_k)|$. Set $\sigma_k(x) = \sigma(x+x_k)/|\sigma(x_k)|$. Then $|\sigma_k| \leq \delta_{j_k}/\delta_{j_k+3}$ in $[-3, \infty)$. Same as before, a subsequence of $\{\sigma_k\}_{k=0}^\infty$ converges to a limit R satisfying (3.3). The rest of the proof follows from an analogous argument as before. This completes the proof of Lemma 3.3 and also the proof of Theorems 2 and 3. ∎

4. Uniqueness of traveling waves. In this section we prove Theorem 1. In the following, U and V are two traveling waves with the same speed c. We want to show that $U(\cdot) \equiv V(\cdot - \xi)$ for some $\xi \in \mathbb{R}$.

4.1. A comparison principle. The sliding method applies on compact intervals.

LEMMA 4.1. *If $V \leqslant U$ on $[a-1,a)\cup(b,b+1]$ where $a \leqslant b$, then $V \leqslant U$ on $[a,b]$.*

Proof. Let ξ be the number such that $\min_{[a-1,b+1]}\{U(\cdot) - V(\cdot - \xi)\} = 0$ and let $y \in [a-1, b+1]$ be the maximum value satisfying $U(y) - V(y-\xi) = 0$. Then $y \notin [a,b]$ since, otherwise, $U'(y) = V'(y-\xi)$ and the equations for $U(\cdot)$ and $V(\cdot - \xi)$ evaluated at y would imply $U(y \pm 1) = V(y - \xi \pm 1)$, contradicting the maximality of y. Thus, $y \in [a-1, a) \cup (b, b+1]$, and by the assumption, $V(y) \leqslant U(y) = V(y-\xi)$. Thus $\xi \leqslant 0$. We conclude that $U(\cdot) \geqslant V(\cdot - \xi) \geqslant V(\cdot)$ on $[a-1, b+1]$. ∎

The success of such a simple translation technique relies on (1) the existence of a minimal translation ξ and (2) the existence of a maximum y, both of which attribute to the fact that a continuous function on a compact set attains its global extremes. When the domain of interest is unbounded, neither ξ nor y may exist, and therefore different techniques are needed.

4.2. Comparison near $x = \infty$. We shall compare traveling waves on the unbounded domain $[0, \infty)$. Since simple translation technique does not work, we shall instead construct a family of supersolutions for which translation technique works. If one is willing to make the assumption $f' \leqslant 0$ on $[1-\delta, 1]$ for some $\delta > 0$, then for every $\varepsilon > 0$,

$$\min\{U + \varepsilon, 1\} \quad \text{on} \quad [-1, \infty)$$

is a supersolution on $[0,\infty)$ provided that $U(-1) \geqslant 1 - \delta$. In this manner, no asymptotic behavior of U near $x = \infty$ is needed.

When only the assumption (A) is made, we construct a different family of supersolutions obtained from the detailed asymptotic behavior of wave profiles and compression:

$$Z(\ell, x) := U([1+\ell]x) \quad \forall\, x \in [-1, \infty), \ell \in (0, 1].$$

The idea here is that the rate of Z approaching 1 as $x \to \infty$ is faster than that of any wave profile, and therefore is strictly bigger than any wave profile for sufficiently large x.

Since $\lim_{x\to\infty} U''(x)/U'(x) = \mu \leqslant 0 < c$ and $U'(x+h)/U'(x) = e^{\int_x^{x+h} U''(s)/U'(s)ds}$, by translation, we may assume that

$$\sup_{x \geqslant 0,\ |h| \leqslant 2} \frac{U''(x+h)}{U'(x)} < c. \tag{4.1}$$

For $\ell \in (0,1]$ and $x \geqslant 0$, writing $y = (1+\ell)x$ and $Z(\ell, x) = Z(x)$, we calculate

$$\begin{aligned}
\mathcal{L}Z(x) &:= c\,Z'(x) - Z(x+1) - Z(x-1) + 2Z(x) - f(Z(x)) \\
&= c[1+\ell]\,U'(y) - U(y+1+\ell) - U(y-1-\ell) + 2U(y) - f(U(y)) \\
&= c\,\ell\,U'(y) + U(y+1) + U(y-1) - U(y+1+\ell) - U(y-1-\ell) \\
&= \ell\,U'(y)\left\{c - \int_0^1 \int_{-1-\ell z}^{1+\ell z} \frac{U''(y+h)}{U'(y)} dh dz\right\} > 0.
\end{aligned}$$

This shows that for each $\ell \in (0,1]$, $Z(\ell, \cdot)$ is a (strict) supersolution on $[0,\infty)$.

LEMMA 4.2. *Assume* (4.1). *Suppose* $V \leqslant U$ *on* $[0,1]$. *Then* $V \leqslant U$ *on* $[0,\infty)$.

Proof. Consider the function, for $x \geqslant 0, \xi \in \mathbb{R}$, and $\ell > 0$,

$$\Psi(\xi,\ell,x) := \int_{V(x-\xi)}^{U([1+\ell]x)} \frac{ds}{f(s)}.$$

Note that

$$\lim_{x\to\infty} \frac{\partial\Psi(\xi,\ell,x)}{\partial x} = \lim_{x\to\infty} \left(\frac{(1+\ell)U'}{f(U)} - \frac{V'}{f(V)} \right) > 0 \quad \forall\, \ell > 0, \xi \in \mathbb{R};$$
$$\inf_{x\geq 0,\xi\in\mathbb{R},\ell\in[0,1]} \frac{\partial\Psi}{\partial\xi} = \inf_{y\in\mathbb{R}} \frac{V'(y)}{f(V(y))} > 0.$$

Thus $\lim_{x\to\infty} \Psi(\xi,\ell,x) = \infty$. For each fixed $\ell \in (0,1]$, there exists at least one ξ such that $\Psi(\xi,\ell,\cdot) \geqslant 0$ on $[0,\infty)$. Let $\xi(\ell)$ be the infimum of such numbers.

We claim that $\xi(\ell) \leqslant 0$. Suppose otherwise. Since $\lim_{x\to\infty} \Psi(\xi(\ell),\ell,x) = \infty$, there exists $y \in [0,\infty)$ such that $\Psi(\xi(\ell),\ell,y) = 0$. We must have $y > 1$, since $V(\cdot - \xi(\ell)) < V(\cdot) \leqslant U(\cdot) \leqslant U([1+\ell]\cdot)$ on $[0,1]$. Thus, for $Z(x) = U([1+\ell]x)$,

$$Z(y) = V(y - \xi(\ell)), \qquad V(\cdot - \xi(\ell)) \leqslant Z(\cdot) \quad \text{on} \quad [0,\infty).$$

This implies $V'(y - \xi(\ell)) = Z'(y)$ and a contradiction

$$0 = \mathcal{L}V\,\big|_{y-\xi(\ell)} \geq \mathcal{L}Z\big|_y > 0.$$

This contradiction shows that $\xi(\ell) \leqslant 0$, so that $V(\cdot) \leqslant V(\cdot - \xi(\ell)) \leqslant U([1+\ell]\cdot)$ on $[0,\infty)$. Sending $\ell \searrow 0$, we obtain that $V(\cdot) \leqslant U(\cdot)$ on $[0,\infty)$. □

4.3. Comparison near $x = -\infty$. In general, on the unbounded interval $(-\infty, 0]$, it is very hard to construct a family of supersolutions that can be used for the translation argument such as that in the previous two subsections; this is due to the fact that the constant state **0** is unstable. Hence we compare directly two traveling waves. We shall show that wave profiles are ordered (i.e., one is bigger than the other) near $x = -\infty$, by magnifying differences between any two wave profiles.

For every $\xi \in \mathbb{R}$ and $x \in \mathbb{R}$, we define

$$W(\xi,x) = \begin{cases} \displaystyle\int_{V(x-\xi)}^{U(x)} \frac{ds}{f(s)} & \text{if } \ c > c_{\min}, \\ \ln U(x) - \ln V(x-\xi) & \text{if } \ c = c_{\min}. \end{cases}$$

Note that $W(\xi,x)$ magnifies the differences between U and V. When $c > c_{\min}$,

$$W_x(\xi,x) := \frac{\partial W(\xi,x)}{\partial x} = \frac{U'}{f(U)} - \frac{V'}{f(V)} \longrightarrow 0 \text{ as } \ x \to \pm\infty.$$

This limit shows that the magnified difference between wave profiles changes slowly. The conclusion for $c = c_{\min}$ is analogous.

LEMMA 4.3. *There exist* $\nu > 0$ *and* $A \in [-\infty,\infty]$ *such that*

$$\lim_{x\to-\infty} W(\xi,x) = A + \nu\xi \quad \forall\, \xi \in \mathbb{R}. \tag{4.2}$$

Consequently, near $x = -\infty$, $U < V(\cdot - \xi)$ *if* $A + \nu\xi < 0$ *and* $U > V(\cdot - \xi)$ *if* $A + \nu\xi > 0$.

Proof. First, we consider the case $c > c_{\min}$. Note that

$$\lim_{x\to-\infty}\Big\{W(\xi,x)-W(0,x)\Big\}=\lim_{x\to-\infty}\int_{x-\xi}^{x}\frac{V'(y)dy}{f(V(y))}=\nu\xi,$$

where $\nu = 1/c$ when $f'(0) = 0$ and $\nu = \lambda/f'(0)$ otherwise. Suppose $\lim_{x\to-\infty} W(\xi,x)$ does not exist. Then $A := \limsup_{x\to-\infty} W(\xi,x) > B := \liminf_{x\to-\infty} W(\xi,x)$. Taking an appropriate ξ, we can assume without loss of generality that $A > 0 > B$. Let α,β be finite numbers satisfying $B < \beta < 0 < \alpha < A$. Then there exist sequences $\{x_i\}$ and $\{y_i\}$ satisfying

$$W(\xi,x_i)=\alpha,\quad W(\xi,y_i)=\beta,\quad x_{i+1}<y_i<x_i,\quad \lim_{i\to\infty}x_i=-\infty.$$

Since $\lim_{x\to-\infty} W_x(\xi,x) = 0$, there exists a large integer i such that $W(\xi,\cdot) > 0$ in $[x_{i+1}-1,x_{i+1}]\cup[x_i,x_i+1]$ and $W(\xi,y_i) < 0$. This implies that $V(\cdot-\xi) < U(\cdot)$ on $[x_{i+1}-1,x_{i+1}]\cup[x_i,x_i+1]$ and $V(y_i-\xi) > U(y_i)$ which is impossible by Lemma 4.1. Thus $A = B$.

The case $c = c_{\min}$ is analogous. ∎

4.4. Proof of Theorem 1. Let U and V be two traveling wave profiles with the same speed c. By translation, we can assume that $V(0) = U(0)$ and that U and V satisfy (4.1). By exchanging the roles of U and V if necessary we can use Lemma 4.3 to conclude that (4.2) holds with $A \in [0,\infty]$.

Let $\eta \geqslant 0$ be the unique value such that

$$\min_{x\in[0,1]}\{U(x)-V(x-\eta)\}=0.$$

By Lemma 4.2, $V(\cdot-\eta)\leqslant U(\cdot)$ on $[0,\infty)$. We claim that $V(\cdot-\eta)\leqslant U(\cdot)$ on $(-\infty,0]$. Suppose not. Then $\inf_{x\in\mathbb{R}} W(\eta,x) < 0$. Since $W_\xi > 0$ and $W(\eta,\pm\infty)\geqslant 0$, there is a unique value $\xi > \eta$ such that $\min_{\mathbb{R}} W(\xi,\cdot) = 0$. This implies that there exists $y\in\mathbb{R}$ such that $W(\xi,y) = 0 = \min_{\mathbb{R}} W(\xi,\cdot)$. It further implies that $V(\cdot-\xi)\leqslant U(\cdot)$ and $V(y-\xi) = U(y)$. A comparison principle shows that this is impossible. Hence, $V(\cdot-\eta)\leqslant U(\cdot)$ on $\mathbb{R}$. Since $\min_{[0,1]}\{U(\cdot-\eta)-V(\cdot)\} = 0$, we must have $\eta = 0$ and $U\equiv V$. ∎

5. Asymptotic expansions. Finally, we derive and verify asymptotic expansions for traveling wave profiles near $x = -\infty$, accurate enough to distinguish the translation differences. The idea is to construct, on $(-\infty,1]$, sub/supersolutions having special tails near $x = -\infty$ and slopes on the interval $[0,1]$. The comparison between a wave profile and a sub/super solution near $x = -\infty$ will be made by a result similar to (4.2) in Lemma 4.3. The comparison on $[0,1]$ will be made in a manner similar to that in the last step of the proof of Lemma 2.3.

5.1. Super/subsolutions. In the following, a Lipschitz continuous function defined on $[a-1,b+1]$ is called a super/subsolution (of speed c) on $[a,b]$ if

$$\pm\,\mathcal{L}\,[\phi](x)\geqslant 0\quad \text{a.e. } x\in(a,b),$$

where $\mathcal{L}\,[\phi](x) := c\,\phi'(x)-\phi(x+1)-\phi(x-1)+2\phi(x)-f(\phi(x))$.

LEMMA 5.1. *Suppose ϕ is a subsolution (or supersolution) on $[a,b]$ and $\phi < U$ (or $\phi > U$) on $[a-1,a)\cup(b,b+1]$. Then $\phi < U$ (or $\phi > U$) on $[a,b]$.*

The proof is similar to that for Lemma 4.1 and is omitted.

Our asymptotic expansion for a wave profile is expressed in terms of a constructed function ϕ such that, for some $x_0 \in \mathbb{R}$,

$$U(x+x_0) = \phi(x+o(1)) \ \ \forall\, x \leqslant 0 \quad \text{where} \quad \lim_{x\to-\infty} o(1) = 0. \tag{5.1}$$

For this, we shall use the same idea as that of Lemma 4.3. Consider the case $\lambda \neq 0$. Suppose ϕ is either a subsolution or a supersolution on $(-\infty, 0]$ and

$$\lim_{x\to-\infty} \frac{\phi'(x)}{\phi(x)} = \lim_{x\to-\infty} \frac{U'(x)}{U(x)} = \lambda > 0. \tag{5.2}$$

Consider the function, for $\xi \in \mathbb{R}$ and $x \leqslant 0$,

$$W(\xi, x) = \int_{\phi(x)}^{U(x+\xi)} \frac{ds}{s} = \ln \frac{U(x+\xi)}{\phi(x)}. \tag{5.3}$$

LEMMA 5.2. *Suppose ϕ satisfies* (5.2) *and is either a supersolution or a subsolution on $(-\infty, 0]$. Let W be defined as in* (5.3). *Then* (4.2) *holds for some $A \in [-\infty, \infty]$.*

The proof is similar to that for Lemma 4.3 and therefore is omitted.

Suppose A is shown to be finite. Then for $x_0 := -A/\nu$, every $\varepsilon > 0$, and all $x \ll -1$, $W(x_0 - \varepsilon, x) < 0 < W(x_0 + \varepsilon, x)$; that is, $\phi(x-\varepsilon) < U(x+x_0) < \phi(x+\varepsilon)$ for every $\varepsilon > 0$ and all $x \ll -1$. Hence (5.1) holds. To construct sub/supersolutions and to show that A is finite, we shall assume that

(B) $|f(u) - f'(0)u| \leq Mu^{1+\alpha}$ for all $u \in [0, 1]$ and some positive constants M and α.

In most cases, we shall construct sub/supersolutions via linear combinations of exponential functions. Note that for $\phi = ae^{\omega x}$, $\mathcal{L}\phi = P(\omega)\phi + [f'(0)\phi - f(\phi)]$, where

$$P(\omega) := c\,\omega - e^{\omega} - e^{-\omega} + 2 - f'(0).$$

Observe that $P(\cdot)$ is concave, positive between its two roots, and negative outside of these two roots. Denote by λ and Λ, where $0 \leqslant \lambda \leqslant \Lambda$, the two roots of $P(\cdot) = 0$. Among all possibilities, we divide them into four cases:

(i) $c = c_{\min}$ and (1.6) has two real roots;
(ii) $c = c_{\min}$ and (1.6) has only one real root;
(iii) $c > c_{\min}$ and $f'(0) > 0$;
(iv) $c > c_{\min}$ and $f'(0) = 0$.

Note that $\lim_{x\to-\infty}\{U'(x)/U(x)\} > 0$ in the cases (i)–(iii). For the last case (iv), $\lambda = 0$ so that sub/supersolutions have to be constructed by nonexponential functions. For this, we need extra assumptions on f.

5.2. The case $c = c_{\min}$ and (1.6) has two real roots. Assume that $c = c_{\min}$ is the minimum wave speed and that the characteristic equation $c_{\min} z = e^z + e^{-z} - 2 + f'(0)$ has two real roots. Let λ be the smaller real root and Λ be the large real root. Then $\lambda < \Lambda$ and

$$\lim_{x\to-\infty} \frac{U'(x)}{U(x)} = \Lambda > 0 \implies \frac{U(x)}{U(0)} = e^{\int_0^x U'/U} = e^{\Lambda x + o(x)}.$$

Choose ω_1 and ω_2 satisfying

$$\lambda < \omega_1 < \Lambda < \omega_2, \qquad \omega_2 < (1+\alpha)\Lambda.$$

Then $P(\omega_1) > 0 = P(\Lambda) > P(\omega_2)$. Consider, for $\varepsilon \in [0,1]$ and small $\delta > 0$,

$$\phi^{\pm}(\varepsilon,\delta,x) := \delta\Big\{e^{\Lambda x} \pm \varepsilon(e^{\omega_1 x} - e^{\Lambda x}) \pm \delta^{\alpha/2}(e^{\Lambda x} - e^{\omega_2 x})\Big\}.$$

Note that when $\varepsilon > 0$ and $x \ll -1$, $\phi^+ \gg U$ and $\phi^- < 0$. Also, for all $x \leqslant 0$,

$$\mathcal{L}\,[\phi^+] = \delta\Big\{\varepsilon P(\omega_1)e^{\omega_1 x} - P(\omega_2)\delta^{\alpha/2}e^{\omega_2 x} + O(1)\delta^{\alpha}\left[\varepsilon^{1+\alpha}e^{(1+\alpha)\omega_1 x} + e^{(1+\alpha)\Lambda x}\right]\Big\} > 0$$

if $\varepsilon \in [0,1]$ and $\delta \in (0,\delta_0]$ for some $\delta_0 > 0$. Similarly, for every $\varepsilon \in [0,1]$ and $\delta \in (0,\delta_0]$, $\max\{0,\phi^-(\varepsilon,\delta,\cdot)\}$ is a subsolution on $(-\infty,0]$. Taking δ_0 small enough we can assume that $\phi_x^{\pm} > 0$ for all $x \in [0,1]$, $\varepsilon \in [0,1]$ and $\delta \in [0,\delta_0]$.

Take ξ negatively large such that $\delta := U(\xi) < \delta_0$. Comparing $U(\cdot + \xi - 1)$ with $\phi^+(\varepsilon,\delta,\cdot)$ on $(-\infty,0]$ for every $\varepsilon \in (0,1]$, we see that $U(x+\xi-1) \leq \phi^+(\varepsilon,\delta,x)$ for all $x \leq 0$. Here the positivity of ε guarantees that $\phi^+ > U$ near $x = -\infty$. Now sending $\varepsilon \searrow 0$ we conclude that $U(x+\xi-1) \leq \delta[1+\delta^{\alpha/2}]e^{\Lambda x}$ for all $x \leqslant 0$. Similarly, $U(x+\xi+1) > \delta[1-\delta^{\alpha/2}]e^{\Lambda x}$ for all $x \leqslant 0$.

Now applying Lemma 5.2 to $\phi = \phi^+(0,\delta_0,x)$, we see that there is the limit

$$A = \lim_{x\to-\infty}\Big\{\ln U(x) - \ln\phi^+(0,\delta_0,x)\Big\} = \lim_{x\to-\infty}\Big\{\ln U(x) - \Lambda x\Big\} - \ln\left[\delta_0\left(1+\delta_0^{\alpha/2}\right)\right].$$

From the estimate in the previous paragraph, A must be finite. Hence we proved the following theorem.

THEOREM 5.1. *Assume* (A) *and* (B). *Let* $(c_{\min}, U)$ *be a traveling wave of the minimum speed where the characteristic equation has two roots* λ, Λ, $\lambda < \Lambda$. *Then, for some* $x_0 \in \mathbb{R}$,

$$U(x) = e^{\Lambda[x+x_0+o(1)]} \quad \forall x \leqslant -1, \quad \textit{where} \quad \lim_{x\to-\infty} o(1) = 0.$$

5.3. The case $c = c_{\min}$ and (1.6) has only one real root. Let $P(z) = c_{\min}z - [e^z + e^{-z} - 2 + f'(0)]$ be the characteristic function at 0. That $P(\cdot) = 0$ has only one real root, denoted by λ, implies that $P(\lambda) = P'(\lambda) = 0$; that is,

$$c_{\min} = e^{\lambda} - e^{-\lambda}, \qquad f'(0) = \lambda(e^{\lambda} - e^{-\lambda}) + (2 - e^{\lambda} - e^{-\lambda}). \tag{5.4}$$

Take $\omega \in (\lambda, [1+\alpha]\lambda)$ and consider the function, for small $\delta > 0$,

$$\phi^*(\delta,x) = \delta[-xe^{\lambda x} - \delta^{\alpha/2}(e^{\lambda x} - e^{\omega x})]. \tag{5.5}$$

Note that $\phi^* > 0$ in $(-\infty,0)$ and $\phi^* < 0$ in $(0,\infty)$. Since $P(\omega) < 0$, for $x \leq 0$,

$$\mathcal{L}\phi^* = \delta\Big\{\delta^{\alpha/2}P(\omega)e^{\omega x} + O(1)\delta^{\alpha}[|x|+1]^{1+\alpha}e^{(1+\alpha)\lambda x}\Big\} < 0.$$

It follows that $\phi^- := \max\{\phi^*, 0\}$ is a subsolution for every $\delta \in (0,\delta_0]$, where $\delta_0 > 0$.

From Lemma 5.2, there exists the limit

$$A = \lim_{x\to-\infty}\Big\{\ln U(x) - \lambda\, x - \ln|x|\Big\}. \tag{5.6}$$

We claim that $A < \infty$. Suppose $A = \infty$. Then for each fixed $\xi \in \mathbb{R}$, $U(x+\xi) > \phi^-(\delta,x)$ for all $x \ll -1$. Since $\phi^- = 0$ on $[0,\infty)$ and ϕ^- is a subsolution, a comparison

gives $U(x+\xi) > \phi^-(\delta, x)$ for all $x \in \mathbb{R}$. This is impossible for every $\xi \in \mathbb{R}$. Thus $A < \infty$.

We now consider the lower bound of A. Since $P(\cdot)$ is a concave function, that λ is a double root to $P(\cdot) = 0$ implies that $P(\omega) < 0$ for every $\omega \neq \lambda$. It is then very hard to construct supersolutions. As the existence of a supersolution implies the existence of a traveling wave, the construction of a supersolution is equivalent to find $c_{\min}$ which is not totally determined by the local behavior of $f(s)$ near $s = 0$. That $c_{\min}$ is the solution of (5.4) which is uniquely determined by $f'(0)$ requires special properties on the nonlinearity on f. The whole nonlinear structure of f on $[0, 1]$ determines whether A is bounded from below. As will be seen in a moment, the answer to whether A is bounded is all we need to determine uniquely the asymptotic behavior of U as $x \to -\infty$, i.e., the alternatives in (1.11).

Case 1. $A > -\infty$. Then A is finite, so from (5.6), the first alternative in (1.11) holds.

Case 2. $A = -\infty$. Fix $\omega \in (\lambda, (1+\alpha)\lambda)$. Consider, for $\varepsilon \in [0, 1]$ and small $\delta > 0$,

$$\phi^+(\varepsilon, \delta, x) = \delta \left\{ [1 - \varepsilon x] e^{\lambda x} - \delta^{\alpha/2} e^{\omega x} \right\}.$$

Direct calculation shows that ϕ^+ is a supersolution on $(-\infty, 0]$ for every $\varepsilon \in [0, 1]$ and $\delta \in (0, \delta_0]$. Fix a translation such that $U(1) \leq \delta_0/2$. For every $\varepsilon \in (0, 1]$ we compare $U(\cdot)$ and $\phi^+(\varepsilon, \delta_0, \cdot)$ on $(-\infty, 0]$. When $x \in [0, 1]$, $U(x) \leq U(1) < \delta_0/2 < \phi(\varepsilon, \delta_0, x)$. Since $A = -\infty$, we see that $U < \phi$ for all $x \ll -1$. It then follows that $U(\cdot) < \phi(\varepsilon, \delta_0, \cdot)$ on $(-\infty, 1]$. Sending $\varepsilon \searrow 0$ we obtain $U(x) \leq \delta_0 e^{\lambda x}$ for all $x \in (-\infty, 0]$.

Also, by Lemma 5.2, there exists the limit

$$\tilde{A} := \lim_{x \to -\infty} \left\{ \ln U(x) - \ln \phi^+(0, \delta_0, x) \right\} = \lim_{x \to -\infty} \left\{ \ln U(x) - \lambda x \right\} - \ln \delta_0.$$

In addition, since $U(x) \leq \delta_0 e^{\lambda x}$ for all $x \in (-\infty, 0]$, $\tilde{A} \leq 0$.

Next we show that $\tilde{A} > -\infty$. To do this, for every $\omega_1 \in [\lambda, \omega]$, consider the function $\phi^-(\omega_1, \delta, x) := \delta[e^{\omega_1 x} + e^{\omega x}]$. It is easy to show that ϕ^- is a subsolution on $(-\infty, 0]$ for every $\omega_1 \in [\lambda, \omega]$ and every $\delta \in (0, \delta_0]$.

Fix a translation such that $U(-1) > 2\delta_0$. For every $\omega_1 \in (\lambda, \omega]$, by comparing U and $\phi^-(\omega_1, \delta_0, x)$, we see that $U > \phi^-(\omega_1, \delta_0, x)$, since $\omega_1 > \lambda$ implies $U > \phi^-$ for all $x \ll -1$. Now sending $\omega_1 \searrow \lambda$ we see that $U(x) \geq \delta_0 e^{\lambda x}$ for all $x \leq 0$. Thus $\tilde{A}$ is finite; namely, the second alternative in (1.11) holds.

Finally, we provide two examples showing that both alternatives in (1.11) can happen.

Example 1. This example provides the second alternative in (1.11). We define

$$U(x) = \frac{e^x}{1 + e^x}, \quad \lambda = 1, \quad c = e - \frac{1}{e},$$
$$f(u) = \frac{u(1-u)(e-1)[2(1-u)^2 + 2eu^2 + (e^2+1)(e+1)u(1-u)/e]}{e(1-u)^2 + eu^2 + u(1-u)(e^2+1)}.$$

Using $e^x = U(x)/[1 - U(x)]$, one can verify that (c, U) is a traveling wave. Since $f'(0) = 2 - 2/e$, $\lambda = 1$ is a double root of the characteristic equation $c\omega = e^\omega + e^{-\omega} - 2 + f'(0)$. Consequently, $c_{\min} = e - 1/e$.

Example 2. We show that the first alternative in (1.11) holds if

$$f \in C^{1+\alpha}([0, 1]), \quad f(0) = f(1) = 0 < f(u) \leqslant f'(0)u \quad \forall\, u \in (0, 1). \tag{5.7}$$

First of all, defining $(c_{\min}, \lambda)$ as in (5.4), one can show that $\min\{1, e^{\lambda x}\}$ is a supersolution with $c = c_{\min}$ so that there is a traveling wave of speed $c_{\min}$. Consequently, the minimum wave speed is given by the solution of (5.4); see, for example, [5, 6, 27].

Also, there is a supersolution given by

$$\phi^+(x) = [1 - \tfrac{\lambda}{1+\lambda}\, x]e^{\lambda x} \quad \forall\, x < 0, \qquad \phi^+(x) = 1 \quad \text{for} \quad x \geqslant 0.$$

Note that, for a large constant M, $\phi^+(x + M) > \phi^*(\delta_0, x)$ on $\mathbb{R}$, where ϕ^* is as in (5.5). Following the existence proof of [5], $(\max\{\phi^*, 0\}, \phi^+)$ sandwiches a solution which satisfies the first alternative in (1.11).

We conclude the following theorem.

THEOREM 5.2. *Assume* (A) *and* (B). *Suppose* $c = c_{\min}$ *and the characteristic equation has a root* λ *of multiplicity 2, i.e.,* (5.4) *holds. Then there is the alternative* (1.11). *In addition, under* (5.7)*, only the first alternative in* (1.11) *holds.*

5.4. The case $c > c_{\min}$ and $f'(0) > 0$. Let λ and Λ, $\lambda < \Lambda$, be two roots of the characteristic equation $P(\cdot) = 0$, where $P(z) = c\, z - [e^z + e^{-z} - 2 + f'(0)]$. Pick ω such that $\lambda < \omega < \min\{\Lambda, (1 + \alpha)\lambda\}$. Then $P(\omega) > 0$. For each $\varepsilon \in (0, e^{-\omega}]$ and small δ, consider functions

$$\phi^{\pm}(\varepsilon, \delta, x) := \delta\left([1 \mp \varepsilon]e^{\lambda x} \pm \varepsilon e^{\omega x}\right), \qquad x \leqslant 1.$$

Note that

$$\min_{0 \leqslant x \leqslant 1} \frac{\phi^+_x(\varepsilon, \delta, x)}{\phi^+(\varepsilon, \delta, x)} = \lambda + \varepsilon(\omega - \lambda), \qquad \max_{0 \leqslant x \leqslant 1} \frac{\phi^-_x(\varepsilon, \delta, x)}{\phi^-(\varepsilon, \delta, x)} = \lambda - \varepsilon(\omega - \lambda).$$

In addition, for all $x \leqslant 0$, $\varepsilon \in (0, 1]$, and $\delta \in (0, 1]$, using $|f(u) - f'(0)u| \leq Mu^{1+\alpha}$ and $0 < \phi^{\pm} \leq 2\delta e^{\lambda x}$ we obtain

$$\begin{aligned} \pm\,\mathcal{L}\,[\phi^{\pm}\delta] &= \delta\varepsilon P(\omega)e^{\omega x} \pm [f(\phi^{\pm}\delta) - f'(0)\phi^{\pm}\delta] \\ &\geqslant \delta e^{\omega x}\left\{\varepsilon\, P(\omega) - 2^{1+\alpha} M\delta^{\alpha} e^{[(1+\alpha)\lambda - \omega]x}\right\}. \end{aligned}$$

Hence, we have the following:

(i) For every $\varepsilon \in (0, e^{-\omega}]$, there exists $x_\varepsilon \leqslant 0$ such that $\phi^{\pm}(\varepsilon, 1, \cdot)$ is a super/subsolution on $(-\infty, x_\varepsilon]$.

(ii) For every $\varepsilon \in (0, e^{-\omega}]$, there exists $\delta_\varepsilon > 0$ such that for every $\delta \in (0, \delta_\varepsilon]$, $\phi^{\pm}(\varepsilon, \delta, \cdot)$ is a super/subsolution on $(-\infty, 0]$.

Indeed, we need only take

$$x_\varepsilon := \min\left\{0, \frac{\ln[\varepsilon P(\omega)] - \ln[2^{1+\alpha}M]}{(1+\alpha)\lambda - \omega}\right\}, \qquad \delta_\varepsilon = \min\left\{1, \left(\frac{\varepsilon P(\omega)}{2^{1+\alpha}M}\right)^{1/\alpha}\right\}.$$

THEOREM 5.3. *Assume* (A), (B), *and* $f'(0) > 0$. *Let* (c, U) *be a traveling wave with speed* $c > c_{\min}$. *Then* $U(x) = e^{\lambda(x + x_0 + o(1))}$ *for some* $x_0 \in \mathbb{R}$, *where* $\lim_{x \to -\infty} o(1) = 0$.

Proof. First of all, note that (4.2) holds for W defined as in (5.3) with $\phi = \phi^+(\varepsilon, 1, x)$.

We show that $A > -\infty$. Suppose $A = -\infty$. Fix $\varepsilon = e^{-\omega}$. Since

$$\lim_{x \to \infty} U'(x)/U(x) = \lambda,$$

there exists $\xi < 0$ such that $U'(x)/U(x) < \lambda + \varepsilon(\omega - \lambda)$ for all $x < \xi + 2$. Now we compare $U(\cdot + \xi)$ with $\phi := \phi^+(\varepsilon, U(\xi), \cdot)$ on $(-\infty, 0]$. By taking negatively large ξ, we may assume that $U(\xi) < \delta_\varepsilon$ so that ϕ is a supersolution on $(-\infty, 0]$.

Note that $\phi(0) = U(0 + \xi)$ and

$$\frac{\phi'(x)}{\phi(x)} > \lambda + \varepsilon(\omega - \lambda) > \frac{U'(x+\xi)}{U(x+\xi)} \quad \forall\, x \in [0, 1]$$

so that $U(\cdot + \xi) < \phi(\cdot)$ on $(0, 1]$. Also, $\lim_{x\to-\infty}[\ln \phi(x) - \ln U(x + \xi)] = \infty$. It follows by comparison that $\phi(\cdot) > U(\cdot + \xi)$ on $(-\infty, 0]$, contradicting $\phi(0) = U(0 + \xi)$. Thus $A > -\infty$.

Similarly, by using the subsolution ϕ^-, one can show that $A < \infty$. Thus $A = \lim_{x\to-\infty}\{\ln U(x) - \lambda x\}$ exists and is finite. This completes the proof. □

5.5. The case $c > c_{\min}$ and $f'(0) = 0$. When $c > c_{\min}$, $\lambda := \lim_{x\to-\infty} U'(x)/U(x)$ is the smaller root to the characteristic equation $cz = e^z + e^{-z} - 2 + f'(0)$. When $f'(0) = 0$, we have $\lambda = 0$. Thus as $x \to -\infty$, $U(x)$ does not decay to 0 exponentially fast. To find the precise rate of decay, we shall assume the following:

(B1) $0 \leqslant ff'' \leqslant Mf'^2$ on $(0, \epsilon]$ for some $\epsilon > 0$ and $M > 0$; $\int_0^\epsilon f'^2(s)/f(s)ds < \infty$.

Simple examples of such functions are

$$f(u) = \kappa u^{1+q}(1-u)^p, \qquad f(u) = \kappa e^{-1/u}(1-u)^p \qquad (\kappa > 0, q > 0, p \geqslant 1).$$

THEOREM 5.4. *Assume* (A), (B1), *and* $f'(0) = 0$. *Let* (c, U) *be a traveling wave with nonminimum speed* c. *Then* (1.12) *holds for some* $x_0 \in \mathbb{R}$.

Proof.

The idea. The proof is based on the following formal calculation. When $f'(0) = 0$ and $c > c_{\min}$, it follows from Theorem 3 that $cU' \approx f(U)$. Then at least formally we should have $c^2U'' \approx cf'(U)U' \approx f(U)f'(U)$. Since by the mean value theorem $U(x+1) + U(x-1) - 2U(x) = U''(y) \approx U''(x)$, we obtain

$$cU' \approx U'' + f(U) \approx f(U)f'(U)/c^2 + f(U) = f(U)[1 + f'(U)/c^2].$$

This suggests that sub/super solutions can be obtained from solutions of ODEs of the form $c\,\phi' = f(\phi)[1 + f'(\phi)/c^2] \pm o(1)$, where $o(1)$ is a small positive term large enough to offset the error of the approximation $U(x+1)+U(x-1)-2U(x) = U''(y) \approx U''(x)$.

Construction of super/subsolutions. Let δ_0 be a small enough constant and be fixed. For every $\delta \in (0, \delta_0]$ and $K \in [1, 1/(4f'^2(\delta))]$, let ϕ be the solution of

$$c\,\phi' = f(\phi)\,\{\,1 + f'(\phi)/c^2 \pm Kf'^2(\phi)\,\} \quad \text{on} \quad (-\infty, 1], \qquad \phi(0) = \delta. \tag{5.8}$$

The solution is given implicitly by

$$\int_\delta^{\phi(x)} \frac{ds}{f(s)[1 + f'(s)/c^2 \pm Kf'^2(s)]} = \frac{x}{c} \quad \forall\, x \leqslant 1.$$

When δ_0 is small, we have $\phi \leq \delta[1 + o(1)]$ and $c\phi' = f(\phi)[1 + o(1)]$ on $(-\infty, 1]$. In the following, $O(1)$ is a quantity bounded by a constant independent of K and δ.

Write (5.8) as $c\,\phi' = (1 + g(\phi))f(\phi)$, where $g := f'/c^2 \pm Kf'^2$. In the following, the arguments of f, f', f'', and g are evaluated at $\phi(x)$, if not specified. Since $f'' \geqslant 0$ and $ff'' = O(1)f'^2$ on the interval of interest, we see that

$$|g| + |g'f/f'| = O(f') + O(f'^2)K.$$

Consequently,

$$c^2\phi''(x) = \{(1+g)f' + fg'\}(1+g)f = ff'\{1 + O(f') + O(f'^2)K\}.$$

Also by the mean value theorem,

$$\phi(x+1) + \phi(x-1) - 2\phi(x) = \phi''(y) \quad \text{for some } \ y \in [x-1, x+1],$$

$$\frac{f'(\phi(y))}{f'(\phi(x))} = \exp\left(\int_x^y \frac{(1+g)ff''}{cf'}\right) = \exp\left(\int_x^y O(f'(\phi(z))dz\right).$$

This implies that

$$f'(\phi(y)) = [1 + O(f'(\phi(x)))]f'(\phi(x)).$$

Similarly,

$$f(\phi(y)) = [1 + O(f'(\phi(x)))]f(\phi(x)).$$

This follows that

$$c^2\phi''(y) = f'f\{1 + O(f') + O(f'^2)K\}\Big|_{\phi(x)}.$$

Hence, for all $x \leqslant 1$,

$$\begin{aligned}\mathcal{L}[\phi](x) &= c\phi' - f - f'f\Big\{c^{-2} + O(f') + O(f'^2)K\Big\} \\ &= ff'^2\Big\{\pm K + O(1) + O(f')K\Big\}.\end{aligned}$$

Thus we have the following lemma.

LEMMA 5.3. *There exist a small positive constant δ_0 and a large constant K_0 such that for every $\delta \in (0, \delta_0]$ and every $K \in [K_0, 1/(4f'^2(\delta))]$, the solution $\phi^{\pm}(\delta, x) := \phi(x)$ of* (5.8) *is a super/subsolution on $(-\infty, 0]$.*

The comparison. Consider the function

$$W^{\pm}(\xi, x) = \int_{\phi^{\pm}(\delta,x)}^{U(x+\xi)} \frac{ds}{f(s)[1 + f'(s)/c^2]} \qquad x \leq 1, \xi \in \mathbb{R}.$$

Following a proof similar to that for Lemma 4.3, we can show that (4.2) holds with $W = W^{\pm}$, $A = A^{\pm} \in [-\infty, \infty]$ and $\nu = 1/c$. Note that

$$\begin{aligned}W^+ - W^- = &\int_{\phi^+}^{\delta} \left\{\frac{1}{f[1 + f'/c^2]} - \frac{1}{f[1 + f'/c^2 + Kf'^2]}\right\} ds \\ &- \int_{\phi^-}^{\delta} \left\{\frac{1}{f[1 + f'/c^2]} - \frac{1}{f[1 + f'/c^2 - Kf'^2]}\right\} ds,\end{aligned}$$

since the two integrals involving K cancel each other. Sending $x \to -\infty$ and using $\phi^{\pm}(-\infty) = 0$ and $\int_0^{\epsilon} f'^2(s)/f(s)ds < \infty$, we then obtain

$$\lim_{x\to-\infty}\{W^+(\xi, x) - W^-(\xi, x)\} = \int_0^{\delta} \frac{2Kf'^2}{f\{[1 + f'/c^2]^2 - [Kf'^2]^2\}}\, ds < \infty.$$

We now show that $A^+ > -\infty$. Suppose on the contrary that $A^+ = -\infty$. For each $\delta \in (0, \delta_0]$, taking $K = 1/(4f'(\delta)^2)$ we see that

$$\frac{\phi^{+\prime}(x)}{f(\phi^+(x))} = \frac{1}{c} - \frac{f'(\phi^+)}{c^3} + \frac{f'^2(\phi^+)}{4cf'^2(\delta)} \geqslant \frac{1}{c} + \frac{1}{8c} \quad \forall\, x \in [0,1]$$

if δ_0 is small enough. As we know that $\lim_{x\to-\infty} U'/f(U) = 1/c$, there exits $\xi < 0$ such that $U'/f(U) < 1/c + 1/(8c)$ for all $x \leq \xi + 1$. Now set $\delta = U(\xi)$ and compare $U(\xi + \cdot)$ and $\phi^+(\delta, \cdot)$ on $(-\infty, 0]$.

As $\phi^{+\prime}/f(\phi^+) > U'/f(U)$ on $[0,1]$ and $\phi(0) = U(\xi + 0)$, we have $\phi^+(\cdot) > U(\xi + \cdot)$ on $(0,1]$. Also, $A^+ = -\infty$ implies that $\phi^+(x) > U(\xi + x)$ for all $x \ll -1$. By comparison, $\phi^+ > U$ on $(-\infty, 0]$, contradicting $\phi^+(0) = U(\xi + 0)$. Thus $A^+ > -\infty$. Similarly, using ϕ^-, we can show that $A^- < \infty$. Hence $A^\pm$ are finite.

Finally, we observe that

$$\lim_{x\to-\infty} W^+(0,x) = \lim_{x\to-\infty} \left\{ \int_\delta^{U(x)} \frac{ds}{f(s)[1 + f'(s)/c^2]} - \frac{x}{c} \right\} - \int_0^\delta \left\{ \frac{1}{1 + f'(s)/c^2} - \frac{1}{1 + f'(s)/c^2 + Kf'^2(s)} \right\} \frac{ds}{f(s)},$$

the assertion of the theorem, i.e., (1.12) thus follows. □

As an illustration, we consider the case when

$$f(u) = \kappa u^2(1-u)^p \qquad (\kappa > 0,\ p \geqslant 1).$$

Then for some integral constant a

$$\int_{1/2}^u \frac{ds}{f(s)[1 + f'(s)/c^2]} = -\frac{1}{\kappa u} + \left(\frac{p}{\kappa} - \frac{2}{c^2}\right) \ln u + a + O(u) \quad \text{as } u \to 0.$$

After translation, we see that, as $x \to -\infty$,

$$-\frac{1}{\kappa U(x)} + \left(\frac{p}{\kappa} - \frac{2}{c^2}\right) \ln U(x) = \frac{x}{c} + o(1).$$

This implies that, as $x \to -\infty$,

$$\frac{1}{U(x)} = \frac{\kappa|x|}{c} + O(\ln|x|) = \frac{\kappa|x|}{c}\Big(1 + o(1)\Big), \qquad \ln U(x) = \ln \frac{c}{\kappa|x|} + o(1).$$

Thus, after another translation,

$$\begin{aligned} U(x) &= \frac{c}{\kappa[|x| - x_0 + o(1)] + (p\,c - 2\kappa/c)\ln|x|} \\ &= \frac{c}{\kappa|x|} - \frac{(pc^2 - 2\kappa)\ln|x|}{\kappa^2 x^2} - \frac{cx_0 + o(1)}{\kappa x^2} \quad \text{as } \ x \to -\infty. \end{aligned}$$

Note that the translation is distinguished by the third term in the Taylor's expansion.

Finally, observe that

$$\int_{1/2}^u \frac{ds}{f(s)[1 + f'(s)/c^2]} = \int_{1/2}^u \frac{ds}{f(s)} - \frac{\ln f(u)}{c^2} + a + o(1) \quad \text{as } u \to 0.$$

In particular, if $f(u) = \kappa u^{1+q}[1 + o(1)]$ for some $q > 0$, then $U \propto |x|^{-1/q}$ so that $\ln f(U) \approx -b \ln |x| + B + o(1)$ for some $b > 0$ and $B \in \mathbb{R}$. Therefore, it is generic that for some constants $b > 0$ and $x_0 \in \mathbb{R}$,

$$\int_{1/2}^{U(x)} \frac{ds}{f(s)} = \frac{c[x + x_0 + o(1)] - b \ln |x|}{c^2}.$$

In a similar manner, we can establish an asymptotic expansion near ∞. We omit the details.

Acknowledgments. We are grateful to the anonymous referees for many helpful comments.

REFERENCES

[1] P. W. Bates, X. Chen, and A. Chmaj, *Traveling waves of bistable dynamics on a lattice* SIAM J. Math. Anal., 35 (2003), pp. 520–546.
[2] M. Bramson, *Convergence of solutions of the Kolmogorov equation to traveling waves*, Mem. Amer. Math. Soc. 44 (1983).
[3] H. Berestycki and B. Larrouturou, *Planar travelling front solutions of reaction-diffusion problems*, preprint.
[4] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.
[5] X. Chen and J.-S. Guo, *Existence and asymptotic stability of traveling waves of a discrete monostable equation*, J. Differential Equations, 184 (2002), pp. 1137–1149.
[6] X. Chen and J.-S. Guo, *Uniqueness and existence of traveling waves for discrete quasilinear monostable dynamics*, Math. Ann., 326 (2003), pp. 123–146.
[7] S.-N. Chow, J. Mallet-Paret, and W. Shen, *Traveling waves in lattice dynamical systems,* J. Differential Equations, 149 (1998), pp. 248–291.
[8] S.-N. Chow, J. Mallet-Paret, and W. Shen, *Stability and bifurcation of traveling wave solution in coupled map lattices*, Dynam. Systems Appl., 4 (1995), pp. 1–26.
[9] P. C. Fife, *Mathematical Aspect of Reacting and Diffusing Systems*, Lect. Notes in Biomath. 28, Springer, Berlin, 1979.
[10] S.-C. Fu, J.-S. Guo, and S.-Y Shieh, *Traveling wave solutions for some discrete quasilinear parabolic equations*, Nonlinear Anal., 48 (2002), pp. 1137–1149.
[11] F. Hamel and N. Nadirashvili, *Travelling fronts and entire solutions of the Fisher-KPP equation in* $\mathbb{R}^N$, Arch. Ration. Mech. Anal., 157 (2001), pp. 91–163.
[12] W. Hudson and B. Zinner, *Existence of traveling waves for a generalized discrete Fisher's equation*, Comm. Appl. Nonlinear Anal., 1 (1994), pp. 23–46.
[13] A. L. Kay, J. A. Sherratt, and J. B. McLeod, *Comparison theorems and variable speed waves for a scalar reaction-diffusion equation*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1133–1161.
[14] S. J. A. Malham and M. Oliver, *Accelerating fronts in autocatalysis*, Proc. R. Soc. Lond. Ser. A, 456 (2000), pp. 1609–1624.
[15] J. Mallet-Paret, *The Fredholm alternative for functional-differential equations of mixed type*, J. Dynam. Differential Equations, 11 (1999), pp. 1–47.
[16] J. Mallet-Paret, *The global structure of traveling waves in spatially discrete dynamical systems*, J. Dynam. Differential Equations, 11 (1999), pp. 49–127.
[17] D. J. Needham and A. N. Barnes, *Reaction-diffusion and phase waves occurring in a class of scalar reaction-diffusion equations*, Nonlinearity, 12 (1999), pp. 41–58.
[18] W. Shen, *Travelling waves in time almost periodic structures governed by bistable nonlinearities: I. Stability and uniqueness*, J. Differential Equations, 159 (1999), pp. 1–54.
[19] W. Shen, *Travelling waves in time almost periodic structures governed by bistable nonlinearities:* II. *Existence*, J. Differential Equations, 159 (1999), pp. 55–101.
[20] N. Shigesada and K. Kawasaki, *Biological Invasions: Theory and Practice*, Oxford Ser. in Ecology and Evolution, Oxford University Press, New York, 1997.
[21] J. A. Sherratt and B. P. Marchant, *Algebraic decay and variable speeds in wavefront solutions of a scalar reaction-diffusion equation*, IMA J. Appl. Math., 56 (1996), pp. 289–302.

[22] B. Shorrocks and I. R. Swingland, *Living in a Patch Environment*, Oxford University Press, New York, 1990.
[23] H. F. Weinberger, *Long-time behavior of a class of biological models*, SIAM J. Math. Anal., 13 (1982), pp. 353–396.
[24] J. Wu and X. Zou, *Asymptotic and periodic boundary values problems of mixed PDEs and wave solutions of lattice differential equations*, J. Differential Equations, 135 (1997), pp. 315–357.
[25] B. Zinner, *Stability of traveling wavefronts for the discrete Nagumo equations,* SIAM J. Math. Anal., 22 (1991), pp. 1016–1020.
[26] B. Zinner, *Existence of traveling wavefronts for the discrete Nagumo equations*, J. Differential Equations, 96 (1992), pp. 1–27.
[27] B. Zinner, G. Harris, and W. Hudson, *Traveling wavefronts for the discrete Fisher's equation*, J. Differential Equations, 105 (1993), pp. 46–62.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 259–278

DEAD CORES AND BURSTS FOR QUASILINEAR SINGULAR ELLIPTIC EQUATIONS*

PATRIZIA PUCCI† AND JAMES SERRIN‡

Abstract. We consider divergence structure quasilinear singular elliptic partial differential equations on domains of $\mathbb{R}^n$ and show that there exist solutions with dead cores and, furthermore, solutions which involve both a dead core and *bursts* within the core. The results are obtained under appropriate monotonicity conditions on both the nonlinearity and the elliptic operator. Important special cases treated here are the p-Laplace and the mean curvature operators.

We also study related problems for p-Laplace equations with weights, which include the Matukuma equation as a prototype.

While it is usually thought that dead cores arise due to loss of smoothness of the underlying equation, we show by examples that they can occur equally for analytic p-Laplace equations.

Key words. quasilinear singular elliptic equations, dead cores

AMS subject classifications. Primary, 35J15; Secondary, 35J70

DOI. 10.1137/050630027

1. Introduction.

We consider quasilinear elliptic partial differential equations having the canonical divergence structure

$$\operatorname{div}\{A(|Du|)Du\} = f(u) \quad \text{in } \Omega. \tag{1.1}$$

Here Ω is a domain (connected open set) in $\mathbb{R}^n$, $n \geq 1$, and Du denotes the vector gradient of the given function $u = u(x)$, $x \in \Omega$. *Unless otherwise stated explicitly, we assume throughout the paper the following conditions on the operator $A = A(\varrho)$ and the nonlinearity $f = f(u)$:*

(A1) $A \in C(\mathbb{R}^+)$, $\mathbb{R}^+ := (0, \infty)$;

(A2) $\varrho \mapsto \varrho A(\varrho)$ is strictly increasing in $\mathbb{R}^+$ and $\varrho A(\varrho) \to 0$ as $\varrho \to 0$;

(F1) $f \in C(\mathbb{R})$; and

(F2) $f(0) = 0$, f is nondecreasing on $\mathbb{R}$, and $f(u) > 0$ for $u > 0$.

Condition (A2) is a minimal requirement for ellipticity of (1.1). Furthermore, it allows both singular and degenerate behavior of the operator A at $\varrho = 0$, that is, at critical points of u. We emphasize that no assumptions of differentiability are made on either A or f.

We also study the related elliptic equation

$$\operatorname{div}(g(|x|)|Du|^{p-2}Du) = h(|x|)f(u) \quad \text{in } \Omega, \quad p > 1, \tag{1.2}$$

where $g,\ h : \mathbb{R}^+ \to \mathbb{R}^+$ are radial functions of class $C^1(\mathbb{R}^+)$ and where the general elliptic operator A is replaced by the p-Laplacian function $A(\varrho) = \varrho^{p-2}$. The celebrated Matukuma equation is a prototype for (1.2).

*Received by the editors April 26, 2005; accepted for publication December 2, 2005; published electronically April 21, 2006. This research was supported by the Project Metodi Variazionali ed Equazioni Differenziali Non Lineari.

http://www.siam.org/journals/sima/38-1/63002.html.

†Dipartimento di Matematica e Informatica, Università degli Studi di Perugia, Via Vanvitelli 1, 06123 Perugia, Italy (pucci@dipmat.unipg.it).

‡ Department of Mathematics, University of Minnesota, Minneapolis, MN 55455 (serrin@math.umn.edu).

By a *classical solution* (or a classical distribution solution) of (1.1) or (1.2) in Ω we mean a function $u \in C^1(\Omega)$ which satisfies (1.1) or (1.2) in the distribution sense.

With the notation $\Phi(\varrho) = \varrho A(\varrho)$ when $\varrho > 0$, and $\Phi(0) = 0$, we introduce the function

$$H(\varrho) = \varrho\Phi(\varrho) - \int_0^\varrho \Phi(s)ds, \qquad \varrho \geq 0.$$

This function is easily seen to be strictly increasing, as follows from the inequality

$$\varrho_1\Phi(\varrho_1) - \varrho_0\Phi(\varrho_0) > (\varrho_1 - \varrho_0)\Phi(\varrho_1) > \int_{\varrho_0}^{\varrho_1} \Phi(s)ds$$

when $\varrho_1 > \varrho_0 \geq 0$. Alternatively, monotonicity follows from the representation

$$H(\varrho) = \int_0^{\Phi(\varrho)} \Phi^{-1}(s)ds, \qquad \varrho \geq 0,$$

this being a consequence of the Stieltjes formula $H(\varrho) = \int_0^\varrho s\, d\Phi(s)$.

For the Laplace operator, that is, when (1.1) takes the classical form

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

we have $A(\varrho) \equiv 1$ and $H(\varrho) = \frac{1}{2}\varrho^2$. Similarly, for the degenerate p-Laplace operator, here denoted by Δ_p, $p > 1$, we have $A(\varrho) = \varrho^{p-2}$ and $H(\varrho) = \varrho^p/p'$, while for the mean curvature operator, one has $A(\varrho) = 1/\sqrt{1+\varrho^2}$ and $H(\varrho) = 1 - 1/\sqrt{1+\varrho^2}$. In the last example, note the anomalous behavior $\Phi(\infty) = H(\infty) = 1$, a possibility which requires extra care and will be treated and discussed separately.

It is also worth observing that (1.1) is precisely the Euler–Lagrange equation for the variational integral

$$I[u] = \int_\Omega \{\mathcal{G}(|Du|) + F(u)\}dx, \qquad F(u) = \int_0^u f(s)ds,$$

where $\mathcal{G}$ and A are related by $A(\varrho) = \mathcal{G}'(\varrho)/\varrho$, $\varrho > 0$. In this case $H(\varrho) = \varrho\mathcal{G}'(\varrho) - \mathcal{G}(\varrho)$, the pre-Legendre transform of $\mathcal{G}$. Similarly, the variational integral for (1.2) is given by

$$I[u] = \int_\Omega \left\{ g(|x|)\frac{|Du|^p}{p} + h(|x|)F(u) \right\} dx,$$

where now $H(\varrho) = \varrho^p/p'$ and p' is the Hőlder conjugate of p.

An elliptic equation or inequality is said to have a *dead core solution* u in some domain $\Omega \subset \mathbb{R}^n$ provided that there exists an open subset Ω_1 with compact closure in Ω, called the *dead core* of u, such that

$$u \equiv 0 \quad \text{in } \Omega_1, \qquad u > 0 \quad \text{in } \Omega \setminus \overline{\Omega_1}.$$

The condition $u > 0$ could be replaced by $u \neq 0$, but for definiteness (and physical reality) we prefer the condition as stated. By the strong maximum principle (see Theorem 1.1 of [7, 8]) a nonnegative solution of (1.1) or (1.2) can have a dead core only if

$$\int_{0+} \frac{ds}{H^{-1}(F(s))} < \infty, \tag{1.3}$$

where H^{-1} is the inverse of H. *We assume that condition* (1.3) *holds throughout the sequel, except for Theorems* 1.2, 1.3, 6.2, *and* 6.3.

The equation $\Delta u = u^q$, for example, allows dead cores only if $0 < q < 1$. Actually condition (1.3) is not only necessary but also sufficient for the existence of solutions with dead cores. More specifically, we have the following main result for (1.1).

THEOREM 1.1. *Suppose* $\Phi(\infty) = H(\infty) = \infty$. *Assume the dead core condition* (1.3) *holds and let* u *be a solution of* (1.1), *with* $0 \le u(x) \le m$ *on* $\partial\Omega$ *for some positive constant* m. *Then the following properties are valid:*

(a) $0 \le u < m$ *in* Ω.

(b) *Assume that*

$$C = \int_0^\infty \frac{ds}{H^{-1}(F(s)/n)} < \infty, \tag{1.4}$$

and let B_R *be a ball with radius* $R \ge C$, *compactly contained in* Ω. *Then* u *has a dead core in* Ω *for all* $m > 0$.

(c) *If* B *is any ball compactly contained in* Ω, *then* $u \equiv 0$ *in* B *provided that* $m > 0$ *is suitably small.*

Equation (1.2) allows a corresponding dead core result, which however we defer until section 6.

A more refined version of Theorem 1.1 can be obtained when $\Omega = B_R$, where B_R is any open ball in $\mathbb{R}^n$, $n \ge 1$, of radius $R > 0$. Until explicitly noted in section 9, we continue to assume that $\Phi(\infty) = H(\infty) = \infty$.

THEOREM 1.2. *The problem*

$$\begin{cases} \operatorname{div}\{A(|Du|)Du\} = f(u) & \text{in } B_R, \\ u = m > 0 \quad \text{on } \partial B_R, & u \in C(\overline{B_R}), \end{cases} \tag{1.5}$$

admits a unique classical (distribution) solution u, *necessarily radial. Moreover* $u = u(r) = u(r, m)$ *is of class* $C^1[0, R]$ *and satisfies* $u \ge 0$, $u' \ge 0$ *in* $[0, R]$ *and* $u'(0) = 0$, *where* $' = d/dr$.

Finally, at any $r > 0$ *where* $u(r, m) > 0$ *we have also* $u'(r, m) > 0$.

It follows from Theorem 1.2 that the solution $u(\cdot\,, m)$ must be one of the following three types:

1. $u > 0$ in B_R.
2. $u(0, m) = 0$ and $u'(r, m) > 0$ when $r > 0$.
3. There exists $S \in (0, R)$ such that $u \equiv 0$ in B_S and $u'(r, m) > 0$ when $r > S$.

That is, in case 3 the solution u of (1.5) has a dead core B_S. The solution $u = u(\cdot\,, m)$ of (1.5) has further properties of interest, given in the next result.

THEOREM 1.3. *The function* $u = u(\cdot\,, m)$ *is continuous and nondecreasing in the variable* $m\,(> 0)$, *and* $u < m$ *in* B_R.

The following theorem gives an important relation between the value m and dead cores solutions of (1.5).

THEOREM 1.4. *Let* $u(\cdot\,, m)$ *be the unique solution of* (1.5). *Then either* $u(\cdot\,, m)$ *has a dead core for all* $m > 0$, *or there is a unique (finite) number*

$$m = m_0 = m_0(R) > 0$$

for which a solution $u_0 = u_0(r) = u_0(r\,, m_0)$ *of* (1.5) *in* B_R *exists, with the properties that*

(i) $u_0(0) = 0$;

(ii) $u(0\,,m) > 0$ *for every* $m > m_0$*; and*
(iii) $u(\cdot\,,m)$ *has a dead core for every* $0 < m < m_0$.

For convenience we define $m_0 = m_0(R)$ to be ∞ when $u(0\,,m) = 0$ for all $m > 0$. The examples

$$\Delta u = (\operatorname{sign} u)\sqrt{|u|}, \tag{1.6}$$

$$\Delta_4 u = u \tag{1.7}$$

are particularly interesting as illustrations of the main theorems above. Indeed, both of these are included in the canonical case

$$\Delta_p u = u|u|^{q-1}, \qquad p > 1, \qquad q > 0,$$

for which $F(u) = |u|^{q+1}/(q+1)$. Here the dead core condition (1.3) reduces exactly to

$$0 < q < p - 1.$$

For these special cases, we search for u_0 in the form $c\,r^k$, c, $k > 0$. Then from (1.5) one finds

$$k = \frac{p}{p-1-q}, \qquad c = k^{-k/p'}(n + kq)^{-k/p}, \qquad m_0 = c\,R^k. \tag{1.8}$$

For the case (1.6) we have $p = 2$, $q = 1/2$, $k = 4$, so that

$$m_0 = \frac{1}{(n+2)^2}\left(\frac{R}{2}\right)^4,$$

while $p = 4$, $q = 1$, $k = 2$ for (1.7), and, in turn,

$$m_0 = \frac{R^2}{2\sqrt{2(n+2)}},$$

which reduces exactly to $m_0 = R^4/400$ and $m_0 = R^2/2\sqrt{10}$ when $n = 3$. In particular for the unit radius $R = 1$ we obtain, respectively, the unexpectedly small numbers $m_0 = 0.00125$ and $m_0 \cong 0.158$.

Equation (1.7), when written in full for $n = 2$, has the form

$$|Du|^2\Delta u + 2u_x^2 u_{xx} + 4u_x u_y u_{xy} + 2u_y^2 u_{yy} = u,$$

which is analytic in all its variables. Thus dead core behavior is not due simply to a lack of smoothness in the basic equation. In fact (1.7) is an analytic partial differential equation, elliptic except at the singular point $Du = 0$, which has a nonanalytic solution.

As a final example, consider the equation

$$\Delta u = (\operatorname{sign} u)\sqrt{|u|} + u|u|^2.$$

Here $A(\varrho) \equiv 1$, $H^{-1}(\varrho) = \sqrt{2\varrho}$, and $F(u) = \frac{2}{3}|u|^{3/2} + \frac{1}{4}|u|^4$. Then

$$C = \sqrt{\frac{n}{2}}\int_0^\infty \frac{ds}{\sqrt{(2/3)s^{3/2} + s^4/4}} < \infty.$$

By numerical calculation $C \cong 6.4334$ if $n = 2$. Therefore, by the results of section 4 we have $m_0 = \infty$ whenever $R \geq 7$. In particular, for the problem

$$\begin{cases} \Delta u = (\operatorname{sign} u)\sqrt{|u|} + u|u|^2 & \text{in } B_7 \subset \mathbb{R}^2, \\ u = m > 0 & \text{on } \partial B_7, \end{cases}$$

a dead core occurs for all $m > 0$. This result also follows without recourse to numerical calculation, since one can write, when $n = 2$,

$$C = \left(\frac{9}{2}\right)^{1/5} \int_0^\infty \frac{dt}{\sqrt{t^{3/2} + t^4}} < \left(\frac{9}{2}\right)^{1/5} \left\{ \int_0^1 \frac{dt}{\sqrt{t^{3/2}}} + \int_1^\infty \frac{dt}{\sqrt{t^4}} \right\} = 5(4.5)^{1/5} \cong 6.75.$$

The case $n = 3$ can be treated in the same way, with $C \cong 7.879$, but here the radius $R = 7$ should be replaced by $R = 8$.

In a related paper [1] Bandle and Vernier-Piro also studied the dead core problem for the weighted equation (1.2). Because of the different assumptions on the weights g and h made there it is hard to compare the two papers. A further dead core theorem concerning p-regular equations (see section 11 of [7, 8]) was given by Diaz and Veron [3]. Again the assumptions are different enough to make it difficult to compare the results.

Sperb [10] considers similar dead core problems for the particular special case of the Laplace equation without weights, that is, $A \equiv 1$, $g \equiv 1$, and $h \equiv 1$. He estimates the critical value m_0 for more general domains than balls, but only for the homogeneous case $f(u) = \text{Const}\, u|u|^{q-1}$, $0 < q < 1$. For balls B_R his estimate is weaker than our exact result (1.8). Similarly his estimates for the size of dead cores apply to more general domains than balls, but again are weaker than ours in the latter case.

Theorems 1.4 for the general equation (1.1) and 7.2 for the weighted p-Laplace equation (1.2) seem to capture and extend many of the ideas of these earlier papers.

2. Proof of Theorem 1.2. *Existence of a radial solution u of* (1.5)*, with $u \geq 0$, $u' \geq 0$, and $u'(0) = 0$.* For the purpose of this proof only, we shall redefine f so that $f(v) = f(m)$ for all $v \geq m$, and $f(v) = 0$ when $v \leq 0$. This will not affect the conclusion of the theorem, since clearly any ultimate solution u of (1.5), with $u \geq 0$, $u' \geq 0$ in $[0, R]$, satisfies $0 \leq u \leq m$.

We shall make use of the Leray–Schauder fixed point theorem, as in Proposition 4.1 of [7, 8]. Denote by X the Banach space $X = C[0, R]$, endowed with the usual norm $\|\cdot\|_\infty$, and let $\mathcal{T}$ be the mapping from X to X defined pointwise for all $w \in X$ by

$$\mathcal{T}[w](r) = m - \int_r^R \Phi^{-1}\left(s^{1-n} \int_0^s t^{n-1} f(w(t))dt\right) ds, \qquad r \in [0, R]. \tag{2.1}$$

Clearly $\mathcal{T}[w](R) = m$. Also

$$\mathcal{T}[w]'(r) = \Phi^{-1}\left(r^{1-n} \int_0^r t^{n-1} f(w(t))dt\right), \qquad r \in (0, R]. \tag{2.2}$$

Obviously $\mathcal{T}[w]'$ is continuous and nonnegative in $(0, R]$, since $0 \leq f(w) \leq f(m)$ for all $w \in X$. Moreover $r^{1-n} \int_0^r t^{n-1} f(w(t))dt$ tends to zero as $r \to 0^+$. Therefore $\mathcal{T}[w]'(r)$ approaches 0 as $r \to 0^+$, since $\Phi(0) = 0$, and in turn $\mathcal{T}[w] \in C^1[0, R]$ with $\mathcal{T}[w]'(0) = 0$.

We claim that if w is a fixed point of $\mathcal{T}$ in X, then $w(0) \geq 0$. Otherwise $w(0) < 0$ and $w(R) = m > 0$. Thus there exists a first point $r_0 \in (0, R)$ such that $w(r) < 0$ in $[0, r_0)$ and $w(r_0) = 0$. Consequently $f(w(r)) = 0$ in $[0, r_0]$ and so $w' \equiv 0$ for $r \in [0, r_0]$ by (2.2). Hence $w(r_0) = w(0) < 0$, which is impossible, proving the claim.

Define the homotopy $\mathcal{H} : X \times [0, 1] \to X$ by

$$\mathcal{H}[w, \sigma](r) = \sigma m - \int_r^R \Phi^{-1}\left(\sigma s^{1-n} \int_0^s t^{n-1} f(w(t))dt\right) ds. \tag{2.3}$$

By the above argument, any fixed point $w_\sigma = \mathcal{H}[w_\sigma, \sigma]$ is of class $C^1[0, R]$ and has the properties $w_\sigma \geq 0$, $w'_\sigma \geq 0$ in $[0, R]$, and $w_\sigma(R) = \sigma m$. Additionally, by (2.2) we find that $\Phi(w'_\sigma) \in C^1[0, R]$, and then from (2.1) that w_σ is a classical (distribution) solution of the problem

$$\begin{cases} [r^{n-1}\Phi(w'_\sigma(r))]' - \sigma r^{n-1} f(w_\sigma(r)) = 0 \quad \text{in } (0, R], \\ w'_\sigma(0) = 0, \quad w_\sigma(R) = \sigma m. \end{cases} \tag{2.4}$$

In turn, it is evident that any function w_1 which is a fixed point of $\mathcal{H}[w, 1]$ (that is $w_1 = \mathcal{H}[w_1, 1]$) is a nonnegative radial distribution solution of problem (1.5), in $B_R \setminus \{0\}$, with $w'(0) = 0$ and $w' \geq 0$ in $[0, R]$.

Since $f > 0$ for $u > 0$ it follows equally from (2.4) that the final statement of the theorem is valid.

We assert that such a fixed point $w = w_1$ exists. We shall use Browder's version of the Leray–Schauder theorem for this purpose (see Theorem 11.6 of [5]).

To begin with, obviously $\mathcal{H}[w, 0] \equiv 0$ for all $w \in X$; that is, $\mathcal{H}[w, 0]$ maps X into the single point $w_0 = 0$ in X. (This is the first hypothesis required in the application of the Leray–Schauder theorem.)

We show next that $\mathcal{H}$ is compact from $X \times [0, 1]$ into X. First, $\mathcal{H}$ is continuous on $X \times [0, 1]$. Indeed, let $w_j \to w$, $\sigma_j \to \sigma$, $(w_j, \sigma_j) \in X \times [0, 1]$. Then in (2.3) clearly $\sigma_j f(w_j) \to \sigma f(w)$, since the modified function f is continuous on $\mathbb{R}$. Hence $\mathcal{H}[w_j, \sigma_j] \to \mathcal{H}[w, \sigma]$, as required.

Next let $(w_k, \sigma_k)_k$ be a bounded sequence in $X \times [0, 1]$. It is clear from (2.2) that

$$\|\mathcal{H}[w_k, \sigma_k]'\|_\infty \leq \Phi^{-1}\left(Rf(m)/n\right). \tag{2.5}$$

As an immediate consequence of the Ascoli–Arzelà theorem $\mathcal{H}$ then maps bounded sequences into relatively compact sequences in X, so $\mathcal{H}$ is compact.

To apply the Leray–Schauder theorem it is now enough to show that there is a constant $M > 0$ such that

$$\|w\|_\infty \leq M \quad \text{for all } (w, \sigma) \in X \times [0, 1], \quad \text{with } \mathcal{H}[w, \sigma] = w. \tag{2.6}$$

Let (w, σ) be a pair of type (2.6). But, as observed above, one has $w \geq 0$, $w' \geq 0$, so that $\|w\|_\infty = w(R) \leq \sigma m \leq m$. Thus we can take $M = m$ in (2.6).

The Leray–Schauder theorem therefore implies that the mapping $\mathcal{T}[w] = \mathcal{H}[w, 1]$ has a fixed point $w \in X$ as asserted, which is the required solution of (1.5) in $B_R \setminus \{0\}$. *The fixed point $u = w$ is a C^1 distribution solution of* (1.5) *in* B_R. The proof is standard. Let $\varphi \in C_c^1(B_R)$. We have to show that

$$\int_{B_R} A(|Du|)Du \cdot D\varphi \, dx = -\int_{B_R} f(u)\varphi \, dx.$$

To this end let $\psi = \varphi k_\varepsilon$, $0 < 2\varepsilon < R$, where

$$k_\varepsilon(x) = \begin{cases} 0 & \text{for} \quad |x| \le \varepsilon, \\ 1 & \text{for} \quad |x| \ge 2\varepsilon, \end{cases}$$

and such that $k_\varepsilon \in C^1(\mathbb{R}^n)$, $0 \le k_\varepsilon \le 1$ in $\mathbb{R}^n$, $|Dk_\varepsilon(x)| \le 2/\varepsilon$ for all x with $\varepsilon \le |x| \le 2\varepsilon$. Consequently, using ψ as a test function in $B_R \setminus \{0\}$, we get

$$\begin{aligned} &\int_{B_R \setminus B_{2\varepsilon}} A(|Du|)Du \cdot D\varphi \, dx + \int_{B_{2\varepsilon} \setminus B_\varepsilon} A(|Du|)Du \cdot (k_\varepsilon D\varphi + \varphi Dk_\varepsilon) dx \\ &\quad = -\int_{B_R \setminus B_{2\varepsilon}} f(u)\varphi dx - \int_{B_{2\varepsilon} \setminus B_\varepsilon} f(u)k_\varepsilon \varphi dx. \end{aligned}$$

Now

$$\begin{aligned} \left| \int_{B_{2\varepsilon} \setminus B_\varepsilon} A(|Du|)Du \cdot (k_\varepsilon D\varphi + \varphi Dk_\varepsilon) dx \right| &\le \sup_{B_{2\varepsilon}} \left\{ \Phi(|Du|) \cdot \left[|D\varphi| + \frac{2}{\varepsilon} |\varphi| \right] \right\} \cdot |B_{2\varepsilon}| \\ &= o(\varepsilon^{n-1}) \end{aligned}$$

since $Du(0) = \mathbf{0}$, and Φ is continuous at $\varrho = 0$ by (A2). Moreover

$$\left| \int_{B_{2\varepsilon} \setminus B_\varepsilon} f(u)k_\varepsilon \varphi \, dx \right| \le \text{ Const } \varepsilon^n.$$

Letting $\varepsilon \to 0$ we get the required conclusion.

Uniqueness of C^1 distribution solutions of (1.5). This is an immediate consequence of the weak comparison principle given in Theorem 5.4 of [7, 8].

THEOREM 2.1 (weak comparison principle [7, 8]). *Assume* (A1), (A2), (F1), *and* (F2) *are satisfied. Let u and v be, respectively, classical solutions of*

$$\operatorname{div}\{A(|Du|)Du\} - f(u) \ge 0 \quad \textit{and} \quad \operatorname{div}\{A(|Dv|)Dv\} - f(v) \le 0$$

in a bounded domain Ω; that is, u is a subsolution and v is a supersolution of (1.1). *Suppose also that u and v are continuous in $\overline{\Omega}$, with $u \le v$ on $\partial\Omega$. Then $u \le v$ in Ω.*

3. Proof of Theorem 1.3. *Monotonicity.* This follows from the weak comparison principle, as above.

Continuity. Let $0 < m_1 < m_2$ and write $u_1(r) = u(r, m_1)$ and $u_2(r) = u(r, m_2)$. We show that

$$0 \le u_2(r) - u_1(r) \le m_2 - m_1, \qquad r \in [0, R]. \tag{3.1}$$

By (2.1), for all $r \in [0, R]$,

$$\begin{aligned} u_2(r) &= m_2 - \int_r^R \Phi^{-1}\left(s^{1-n} \int_0^s t^{n-1} f(u_2(t)) dt \right) ds, \\ u_1(r) &= m_1 - \int_r^R \Phi^{-1}\left(s^{1-n} \int_0^s t^{n-1} f(u_1(t)) dt \right) ds. \end{aligned}$$

Then by subtraction

$$\begin{aligned} u_2(r) - u_1(r) &= m_2 - m_1 \\ &\quad - \int_r^R \left[\Phi^{-1}\left(s^{1-n} \int_0^s t^{n-1} f(u_2(t)) dt \right) - \Phi^{-1}\left(s^{1-n} \int_0^s t^{n-1} f(u_1(t)) dt \right) \right] ds. \end{aligned}$$

The function Φ^{-1} is strictly increasing by (A2) and f is nondecreasing by (F2). Therefore, since $u_1 \le u_2$ in $[0,R]$ by monotonicity, one sees that the quantity in square brackets above is nonnegative, and (3.1) is proved.

Proof that $u < m$ in B_R. By (2.1) it is enough to show that

$$I = \int_r^R \Phi^{-1}\left(s^{1-n}\int_0^s t^{n-1}f(u(t))dt\right)ds > 0 \quad \text{for } r \in [0,R).$$

Clearly $u > 0$ in some interval $(r_0, R]$ with $r_0 \ge 0$, and in turn $f(u(s)) > 0$ in $(r_0, R]$ by (F2). Therefore

$$I \ge \int_{\max\{r_0,\, r\}}^R \Phi^{-1}\left(s^{1-n}\int_{r_0}^s t^{n-1}f(u(t))dt\right)ds > 0,$$

as required. □

4. The critical value m_0. We begin with a preliminary result, of interest in itself.

THEOREM 4.1. *If $u_1 = u(\cdot, m_1)$ has a dead core B_{S_1}, then $u_2 = u(\cdot, m_2)$, where $m_2 < m_1$, has a dead core B_{S_2}, with $S_2 > S_1$. Similarly, if either $u_1(0) > 0$ or $u_1(0) = 0$ and $u_1(r) > 0$ for $r \in (0,R]$, then $u_2 > u_1$ in B_R when $m_2 > m_1$.*

Proof. To prove the first part of the lemma, assume for contradiction that $m_2 < m_1$, and either $u_2(r) > 0$ in $(0,R]$, or $0 < S_2 \le S_1$. In the first of these cases the solutions u_1 and u_2 must cross at some point $r_0 \in (S_1, R)$. Then applying Theorem 2.1 we find that $u_1 \equiv u_2$ in $[0, r_0]$, which is an obvious contradiction since $u_2(r) > 0$ on $(0, r_0]$, while $u_2 \equiv u_1 \equiv 0$ in $[0, S_1]$. The next case $0 < S_2 < S_1$ leads to a contradiction in the same way.

The remaining case, when $S = S_2 = S_1 > 0$, needs more care. For $\varepsilon \in (0,R)$ define

$$u_\varepsilon(r) = \begin{cases} 0, & r \in [0,\varepsilon], \\ u_1(r-\varepsilon), & r \in (\varepsilon, R]. \end{cases}$$

If $\varepsilon > 0$ is suitably small then one has $m_1 > u_\varepsilon(R) > m_2 = u_2(R)$, while at the same time

$$u_2(S+\varepsilon) > 0 = u_1(S) = u_\varepsilon(S+\varepsilon). \tag{4.1}$$

Thus there is a point $r_0 \in (S+\varepsilon, R)$ where u_ε and u_2 cross.

We assert that u_ε is a supersolution of (1.1) in the annulus $B_R \setminus \overline{B_\varepsilon}$. Indeed in this set we have

$$\begin{aligned} \operatorname{div}\{A(|Du_\varepsilon|)Du_\varepsilon\} - f(u_\varepsilon) &= \{A(|u_\varepsilon'|)u_\varepsilon'\}' + \frac{n-1}{r}A(|u_\varepsilon'|)u_\varepsilon' - f(u_\varepsilon) \\ &= \left(\frac{n-1}{r} - \frac{n-1}{r-\varepsilon}\right)\Phi(u_1'(r-\varepsilon)) \\ &= -\varepsilon\,\frac{n-1}{r(r-\varepsilon)}\Phi(u_1'(r-\varepsilon)) \le 0. \end{aligned} \tag{4.2}$$

Observing that $u_2(0) = u_\varepsilon(0) = 0$, we can then apply the comparison principle, Theorem 2.1, in B_{r_0}. Therefore $u_2 \le u_\varepsilon$ in $[0, r_0]$, which contradicts (4.1) at the point $r = S + \varepsilon$ and completes the first part of the proof.

To obtain the second part of the theorem, assume for contradiction that $u_2(0) = u_1(0)$ when $m_2 > m_1$. Of course by Theorem 1.2 we have $u_1'(r) > 0$ for $r \in (0, R]$. Define

$$\tilde{u}_\varepsilon(r) = \begin{cases} u_2(0), & r \in [0, \varepsilon], \\ u_2(r - \varepsilon), & r \in (\varepsilon, R], \end{cases}$$

where ε is chosen so small that $m_2 > \tilde{u}_\varepsilon(R) > m_1 = u_1(R)$. On the other hand $u_1(\varepsilon) > \tilde{u}_\varepsilon(\varepsilon) = 0$. Hence there is a crossing point $r_0 \in (\varepsilon, R)$ where $u_1(r_0) = \tilde{u}_\varepsilon(r_0)$. As before u_ε is a supersolution of (1.1) in B_R so that $u_1 \le \tilde{u}_\varepsilon$ in B_{r_0} by Theorem 2.1. Therefore, $u_1 \equiv \tilde{u}_\varepsilon \equiv 0$ in $[0, \varepsilon]$, which is impossible, since $u_1(r) > 0$ for all $r \in (0, R]$.

That $u_2 > u_1$ in all B_R now follows at once, since otherwise u_2 and u_1 would cross at some value $r = r_0$ in which case comparison would lead to the absurd result $u_2 \equiv u_1$ in B_{r_0}. □

Proof of Theorem 1.4. For the purpose of this proof, we suppose that there is some $m > 0$ for which $u(0\,, m) > 0$.

Existence of u_0. Define

$$m_0 = \inf\{m > 0 \ : \ u(0, m) > 0\}. \tag{4.3}$$

We assert first that $m_0 > 0$. Let $\bar{m} > 0$ be so small that

$$C_{\bar{m}} = \int_0^{\bar{m}} \frac{ds}{H^{-1}(F(s)/n)} < R, \tag{4.4}$$

which of course is possible by assumption (1.3); see Lemma 3.2 of [7, 8]. Define $v(r) = w(r - S)$, $r \in [S, R]$, $S = R - C_{\bar{m}}$, where w is the function constructed in the dead core Lemma 7.1 of [7, 8], with $\sigma = 1/n$ and $C = C_{\bar{m}}$. We assert that v is a supersolution of (1.1) in the set $B_R \setminus \overline{B}_S$. In fact

$$\operatorname{div}\{A(|Dv|)Dv\} = [\Phi(v')]' + \frac{n-1}{r}\Phi(v') \le \left\{1 + \frac{n-1}{r}(r - S)\right\} \sigma f(v)$$

by (iii) and (iv) of Lemma 7.1 of [7, 8]. Thus

$$\operatorname{div}\{A(|Dv|)Dv\} \le \left\{1 - \frac{n-1}{nr}S\right\} f(v) \le f(v),$$

as required. Then, since $v(S) = v'(S) = 0$ and $v(R) = \bar{m}$, by defining v to be zero on B_S, the extended function v is a supersolution of (1.1) in B_R. By the comparison principle, Theorem 2.1, we find that $u(\cdot\,, \bar{m}) \equiv 0$ in B_S. Therefore $m_0 \ge \bar{m} > 0$ by (4.3) and the claim is proved.

Next, if (i) would be false, then $u_0(0) > 0$ and by Theorem 1.3 also $u(0\,, m) > 0$ for all $m > 0$ sufficiently near m_0, which would contradict (4.3). Property (ii) is again a direct consequence of the definition (4.3) of m_0 and Theorem 1.3. Finally if there is $m \in (0, m_0)$ such that the corresponding solution $u(\cdot\,, m)$ of (1.5) has no dead core, then $u(0\,, m) \ge 0$ and $u(r\,, m) > 0$ for $r \in (0, R]$. Thus by Theorem 4.1, with $m_1 = m$ and $m_2 = m_0$, we get $u_0(0) > u(0, m) \ge 0$, contradicting (i) and proving (iii).

Uniqueness of u_0. Suppose both m_0 and $\bar{m}_0$ have the properties (i)–(iii) of the theorem. Then $u_0(0) = u_0(0\,, m_0) = 0$ by (i), while $u(0\,, m) > 0$ when $m > \bar{m}_0$ by (ii). Hence $m_0 \le \bar{m}_0$. Similarly $\bar{m}_0 \le m_0$. Therefore $\bar{m}_0 = m_0$, as desired.

The case $m_0 = \infty$. If every solution of (1.5) is such that $u(0\,,m) = 0$ for all $m > 0$, then $u(\cdot\,,m)$ has a dead core for all $m > 0$. Otherwise there would be $\bar m > 0$ for which $u(0\,,\bar m) = 0$ and $u(r\,,\bar m) > 0$ for $r \in (0, R]$. Hence $u(0\,,m) > 0$ for $m > \bar m$ by Theorem 4.1, contradicting the assumption. This also justifies the earlier agreement that $m_0 = \infty$ in this case. □

Remark. In summary, if m_0 is finite and $m > m_0$, then the solution $u = u(\cdot\,,m)$ of (1.5) is positive; namely, $u(r\,,m) > 0$ for all $r \in [0, R]$. On the other hand, if $m < m_0 \le \infty$ then the solution $u = u(\cdot\,,m)$ of (1.5) has a dead core $B_S \subset B_R$, $0 < S < R$.

5. The size of a dead core and proof of Theorem 1.1. Recall the assumption that $\Phi(\infty) = H(\infty) = \infty$, and let

$$C = \int_0^\infty \frac{ds}{H^{-1}(F(s)/n)}. \tag{5.1}$$

Clearly $0 < C \le \infty$ since the integral is convergent at 0 by (1.3) and by Lemma 3.2 of [7, 8] with $\sigma = 1/n$. Of course the integral can possibly diverge at ∞.

THEOREM 5.1. *We have*

$$m_0 = \infty \qquad \textit{if} \quad C < \infty \quad \textit{and} \quad R \ge C, \tag{5.2}$$

while

$$m_0 \ge \bar m \qquad \textit{if} \quad R < C, \tag{5.3}$$

where $\bar m$ *is defined by the relation*

$$R = \int_0^{\bar m} \frac{ds}{H^{-1}(F(s)/n)}.$$

Proof. The proof of (5.2) is essentially the same as the proof of the first part of Theorem 1.4, the only exception being that $C_{\bar m}$ given by (4.4) is replaced by C.

To obtain (5.3), we define $v(r) = w(r)$ as in the proof of Theorem 1.4, with $S = 0$, $\sigma = 1/n$, and $C = R$. Then by Lemma 7.1 of [7, 8] we have $v(0) = v'(0) = 0$, $v(R) = w(R) = \bar m$, while v is a supersolution of (1.1) in B_R. It follows by comparison that $0 \le u(r, \bar m) \le v(r)$. Hence $u(0\,,\bar m) = v(0) = 0$, and in turn from the definition (4.3) of m_0 we get $m_0 \ge \bar m$, as required in (5.3). □

THEOREM 5.2. *Let* $m < m_0$, *so that a dead core exists by Theorem* 1.4, (iii). *In particular the solution* $u = u(\cdot\,,m)$ *satisfies*

$$u \equiv 0 \qquad \textit{in } B_S \subset B_R,$$

where

$$R - \int_0^m \frac{ds}{H^{-1}(F(s)/n)} < S < R.$$

If $R \ge C$, *then for all* $m > 0$ *one has*

$$R - C < S < R.$$

Proof. The proof is the same as the first part of the proof of Theorem 1.4. □

Remark. For any $\varepsilon > 0$, if m is suitably small (depending on ε) we have $R - \varepsilon < S < R$.

Proof of Theorem 1.1. (a) That $u \geq 0$ follows from Theorem 2.1 by comparing the given solution u with the trivial solution 0.

The constant function m is a supersolution of (1.1), so that again by Theorem 2.1 we have $u \leq m$ in Ω. In fact $u < m$ in Ω. To see this, let y be any point of Ω and B a ball in Ω centered at y. Let $v(\cdot\,, m)$ be the radial solution of (1.1) in B constructed in Theorem 1.2, with $v(|x-y|\,, m) = m$ for $x \in \partial B$. Therefore $u(x) \leq m = v(|x-y|\,, m)$ for $x \in \partial B$, and in turn $u(x) \leq v(|x-y|\,, m) < m$ for $x \in B$ by the final part of Theorem 1.3.

(b) This is a direct consequence of Theorem 5.1.

(c) Clearly there exists $R > 0$ such that $\overline{B} \subset B_R \subset\subset \Omega$, with B and B_R centered at the same point of Ω. By (a) we know that $u < m$ on ∂B_R. Denote by $R - \varepsilon$ the radius of B; then by comparison, together with the remark after Theorem 5.2, we have $u \equiv 0$ in B when $m > 0$ is suitably small. □

6. The equation $\mathbf{div}\{g(|x|)|Du|^{p-2}Du\} = h(|x|)f(u)$. Consider the quasilinear singular elliptic equation

$$\mathrm{div}(g(|x|)|Du|^{p-2}Du) = h(|x|)f(u) \qquad \text{in } \Omega, \qquad p > 1, \tag{6.1}$$

where $g,\ h : \mathbb{R}^+ \to \mathbb{R}^+$ are radial functions of class $C^1(\mathbb{R}^+)$, and Ω is a domain of $\mathbb{R}^n$, $n \geq 1$, containing the origin. Prototypes of (6.1), with nontrivial functions g, h, are given, for example, by equations of Matukuma type and equations of Batt–Faltenbacher–Horst type. More precisely, the Matukuma-type equation is given by

$$\Delta_p u = \frac{f(u)}{1 + r^\sigma}, \qquad r = |x|, \qquad \sigma > 0, \tag{6.2}$$

where Δ_p denotes the p-Laplace operator, $p > 1$, and where also $g(|x|) \equiv 1$, $h(|x|) = 1/(1 + r^\sigma)$. A second example is the equation

$$\Delta_p u = \frac{r^\sigma}{(1 + r^{p'})^{\sigma/p'}} \cdot \frac{f(u)}{r^{p'}}, \qquad \sigma > 0, \tag{6.3}$$

where now $g(|x|) \equiv 1$, $h(|x|) = r^{\sigma - p'}/(1 + r^{p'})^{\sigma/p'}$.

All these equations are discussed in detail in section 4 of [6], as special cases of the main example[1]

$$\mathrm{div}(r^k|Du|^{p-2}Du) = r^\ell \left(\frac{r^s}{1 + r^s}\right)^{\sigma/s} f(u), \tag{6.4}$$
$$k \in \mathbb{R}, \qquad \ell \in \mathbb{R}, \qquad s > 0, \qquad \sigma > 0.$$

In particular, in [6] conditions on the exponents were found so that, under appropriate behavior of the nonlinearity f, radial ground states for (6.2)–(6.4) are unique.

We shall also be interested in the radial version of (6.1), when Ω is a ball B_R *centered at* 0 with radius $R > 0$, namely,

$$[a(r)|u'|^{p-2}u']' = b(r)f(u) \qquad \text{in } (0, R), \tag{6.5}$$

[1] In [6] the equation (6.1) was written in the form

$$\mathrm{div}(g(|x|)|Du|^{p-2}Du) + h(|x|)f(u) = 0.$$

The two versions are reconciled by replacing f by $-f$.

where, with obvious notation,

$$a(r) = r^{n-1}g(r), \quad b(r) = r^{n-1}h(r). \tag{6.6}$$

Motivated by the case $a(r) = b(r) = r^{n-1}$ in [6], the functions a and b are assumed to be such that (6.5) *can be transformed* by the change of variables $r \mapsto t(r)$,

$$t(r) = \int_0^r [b(s)/a(s)]^{1/p}\, ds, \qquad r \geq 0, \tag{6.7}$$

$t : \mathbb{R}_0^+ \to \mathbb{R}_0^+$, $t(0) = 0$, to the form

$$[q(t)|v_t|^{p-2}v_t]_t = q(t)f(v), \tag{6.8}$$

where

$$q(t) = [a(r(t))]^{1/p}[b(r(t))]^{1/p'}, \qquad t > 0. \tag{6.9}$$

This requires, in particular, the following conditions on the coefficients:

$$a,\ b > 0, \qquad a,\ b \in C^1(\mathbb{R}^+), \qquad (b/a)^{1/p} \in L^1[0, R]. \tag{A3}$$

We shall ask as well that the transformed equation (6.8) be compatible with the structure:

$$q \in C^1(\mathbb{R}^+), \quad q > 0 \quad \text{in} \ \ \mathbb{R}^+; \tag{Q1}$$

$$\text{there is} \quad \delta > 0 \quad \text{such that} \quad q_t > 0 \quad \text{in} \ \ (0, \delta). \tag{Q2}$$

By (Q2) the weight $q(t)$ has a finite limit as $t \to 0^+$ and in turn also $\bar{q}(r) = q(t(r))$ is bounded as $r \to 0^+$ by (6.7). Hence $b = (b/a)^{1/p}\bar{q} = O((b/a)^{1/p})$ as $r \to 0^+$ by (6.9), and so by (A3)

$$b \in L^1[0, R] \quad \text{and} \quad h \in L^1_{\text{loc}}(B_R). \tag{6.10}$$

Definition. A classical solution of (6.1) is a nonnegative function u of class $C(\Omega) \cap C^1(\Omega \setminus \{0\})$, which is a distribution solution of (6.1) in Ω, of course also with

$$g|Du|^{p-1} \in L^1_{\text{loc}}(\Omega). \tag{6.11}$$

THEOREM 6.1 (weak comparison principle). *Let u and $\tilde{u}$ be, respectively, classical super- and subsolutions of* (6.1) *in a bounded domain Ω. Suppose also that u and $\tilde{u}$ are continuous in $\overline{\Omega}$, with $u \geq \tilde{u}$ on $\partial\Omega$. Then $u \geq \tilde{u}$ in Ω.*

Proof. We follow the proof of Theorem 5.4 of [7, 8].

Let $w = u - \tilde{u}$ in $\overline{\Omega}$. If the conclusion fails, then there exists a point $x_1 \in \Omega$ such that $w(x_1) < 0$. Fix $\varepsilon > 0$ so small that $w(x_1) + \varepsilon < 0$. Consequently, since $w \geq 0$ on $\partial\Omega$ it follows that the function $w_\varepsilon = \min\{w + \varepsilon,\ 0\}$ is nonpositive and has compact support in Ω. By the distribution meaning of solutions, taking the Lipschitz continuous function w_ε as a test function, we get

$$\int_\Omega g(|x|)\{|Du|^{p-2}Du - |D\tilde{u}|^{p-2}D\tilde{u}\}Dw_\varepsilon \leq \int_\Omega h(|x|)\{f(\tilde{u}) - f(u)\}w_\varepsilon. \tag{6.12}$$

The left-hand side of (6.12) is positive since $p > 1$ and $g(|x|) > 0$ for $x \in \Omega \setminus \{0\}$ and $Dw_\varepsilon \equiv Dw = Du - D\tilde{u} \not\equiv \mathbf{0}$ when $w + \varepsilon < 0$, while otherwise $Dw_\varepsilon = \mathbf{0}$ (a.e.).

Moreover, when $w + \varepsilon < 0$ there holds $0 \le u < \tilde{u} - \varepsilon$; hence $f(\tilde{u}) - f(u) \ge 0$ since f is nondecreasing by (F2). Thus the right-hand side of (6.12) is nonpositive, since $h(|x|) > 0$ for $x \in \Omega \setminus \{0\}$, which is a contradiction. □

In the rest of this section we continue to denote by B_R the open ball *centered at* 0 with radius $R > 0$, that is, $B_R = B(0, R)$. Our first result is the analogue of Theorem 1.2, where now we use $A(\varrho) = \varrho^{p-2}$, $\Phi(\varrho) = \varrho^{p-1}$, consistent with section 1. In this case of course $\Phi(\infty) = H(\infty) = \infty$ and $\Phi^{-1}(\tau) = \tau^{1/(p-1)}$.

THEOREM 6.2. *The problem*

$$\begin{cases} \operatorname{div}\{g(|x|)|Du|^{p-2}Du\} = h(|x|)f(u) & \text{in } B_R, \\ u = m > 0 \quad \text{on } \partial B_R, \quad u \in C(\overline{B_R}), \end{cases} \tag{6.13}$$

admits a unique classical solution u, necessarily radial.

Moreover $u = u(r) = u(r, m)$ is of class $C^1[0, R]$ and satisfies $u \ge 0$, $u' \ge 0$ in $[0, R]$ and $u'(0) = 0$, where $' = d/dr$. Finally, at any $r > 0$ where $u(r, m) > 0$ we have also $u'(r, m) > 0$.

It follows from Theorem 6.2 that the solution $u(\cdot, m)$ must be one of the following three types:

1. $u > 0$ in B_R.
2. $u(0, m) = 0$ and $u'(r, m) > 0$ when $r > 0$.
3. There exists $S \in (0, R)$ such that $u \equiv 0$ in B_S and $u'(r, m) > 0$ when $r > S$.

That is, in case 3 the solution u of (1.5) has a dead core B_S.

Proof. Existence of a radial solution u of (6.13)*, with $u \ge 0$, $u' \ge 0$, and $u'(0) = 0$.* This will be accomplished by first showing that (6.8) has a solution $v = v(t)$ on $[0, T]$, $T = t(R)$, with $v \ge 0$, $v_t \ge 0$, and $v_t(0) = 0$, where $t = t(r)$ is given by (6.7).

Indeed, by following the proof of Theorem 1.2 almost word for word, including the redefinition of f, but with the exception that r^{n-1} is replaced by $q(t)$, one obtains[2] an appropriate fixed point $w = w(t) \in X = C[0, T]$ satisfying

$$w(t) = m - \int_t^T \left(\frac{1}{q(s)} \int_0^s q(\tau) f(w(\tau)) d\tau \right)^{1/(p-1)} ds, \qquad t \in [0, T]. \tag{6.14}$$

Clearly $w(T) = m$ and

$$w_t(t) = \left(\frac{1}{q(t)} \int_0^t q(\tau) f(w(\tau)) d\tau \right)^{1/(p-1)}, \qquad t \in (0, T]. \tag{6.15}$$

Obviously w_t is continuous and nonnegative in $(0, T]$, since $0 \le f(w) \le f(m)$ for all $w \in X$. Moreover by (Q2)

$$0 < \frac{1}{q(t)} \int_0^t q(\tau) f(w(\tau)) d\tau \le t f(m) \to 0 \qquad \text{as} \quad t \to 0^+.$$

[2] In view of (Q2), the bound $Rf(m)/n$ in (2.5), holding for the standard weight r^{n-1}, should be replaced by

$$L = \begin{cases} Rf(m) & \text{if } 0 < R \le \delta, \\ R \dfrac{\max_{t \in [\delta, R]} q(t)}{\min_{t \in [\delta, R]} q(t)} f(m) & \text{if } \delta < R. \end{cases}$$

Therefore $w_t(t)$ approaches 0 as $t \to 0^+$ by (6.15). In turn we have $w \in C^1[0,T]$ with $w_t(0) = 0$.

We claim moreover that $w(0) \geq 0$. Otherwise $w(0) < 0$ and $w(T) = m > 0$. Thus there exists a first point $t_0 \in (0,T)$ such that $w(t) < 0$ in $[0,t_0)$ and $w(t_0) = 0$. Consequently $f(w(t)) = 0$ in $[0,t_0]$ and so $w_t \equiv 0$ for $t \in [0,t_0]$ by (6.15). Hence $w(t_0) = w(0) < 0$, which is impossible, proving the claim.

The function $v = w$ is the desired solution of (6.8) in $[0,T]$, and in turn $u(r) = v(t(r))$ is a C^1 solution of (6.5) on $(0,R]$.

The function $u(|x|) = v(t(|x|))$ is a C^1 distribution solution of (6.13) *in B_R.* First we show (6.11) for $\Omega = B_R$. Recall that $v \in C^1[0,T]$ and $v_t(0) = 0$, while also, as shown above, q is bounded as $t \to 0^+$. Then using the change of variables (6.7) we get

$$a(r)|u'(r)|^{p-1} = q(t)|v_t(t)|^{p-1} \to 0 \qquad \text{as} \quad r,\ t \to 0^+.$$

Therefore, recalling (6.6), it follows that (6.11) holds in B_R for $u(|x|) = v(t(|x|))$. The rest of the proof is standard, as in the last part of the proof of Theorem 1.2. Thus u is a C^1 distribution solution of (6.13) in B_R.

Uniqueness of C^1 distribution solutions of (6.13). This is an immediate consequence of the weak comparison principle given in Theorem 6.1. □

The solution $u = u(\cdot\,, m)$ of (6.13) has further properties of interest, given in the next result.

THEOREM 6.3. *The function $u = u(\cdot\,, m)$ is continuous and nondecreasing in the variable m, and $u < m$ in B_R.*

Proof. Monotonicity. This follows from the weak comparison principle Theorem 6.1.

Continuity. Using the notation of the proof of Theorem 1.3, we show that if $0 < m_1 < m_2$, then

$$0 \leq v_2(t) - v_1(t) \leq m_2 - m_1 \qquad \text{on } [0,T]. \tag{6.16}$$

By (6.14) we have for all $t \in [0,T]$

$$v_2(t) - v_1(t) = m_2 - m_1 - \int_t^T \left[\left(\frac{1}{q(s)} \int_0^s q(\tau) f(v_2(\tau)) d\tau \right)^{1/(p-1)} - \left(\frac{1}{q(s)} \int_0^s q(\tau) f(v_1(\tau)) d\tau \right)^{1/(p-1)} \right] ds,$$

and (6.16) now follows exactly as in the proof of Theorem 1.3. Consequently

$$0 \leq u_2(r) - u_1(r) = v_2(t(r)) - v_1(t(r)) \leq m_2 - m_1 \qquad \text{for } r \in [0,R],$$

as required.

Proof that $u < m$ in B_R. It is equivalent to show the same property for the corresponding fixed point solution v of (6.9). By (6.14) it is enough to prove that

$$I = \int_t^T \left(\frac{1}{q(s)} \int_0^s q(\tau) f(v(\tau)) d\tau \right)^{1/(p-1)} ds > 0 \quad \text{for } t \in [0,T).$$

Clearly $v > 0$ in some interval $(t_0, T]$ with $t_0 \geq 0$, and in turn $f(v(t)) > 0$ in $(t_0, T]$ by (F2). Therefore

$$I \geq \int_{\max\{t_0,\, t\}}^T \left(\frac{1}{q(s)} \int_{t_0}^s q(\tau) f(v(\tau)) d\tau \right)^{1/(p-1)} ds > 0,$$

as required. □

7. The critical value m_0 for (6.1). We now turn to the dead core theorem for (6.1), analogous to Theorem 1.4 for (1.1). In the present case we have $H^{-1}(\tau) = (p'\tau)^{1/p}$ and the dead core condition (1.3) becomes

$$\int_{0+} \frac{ds}{[F(s)]^{1/p}} < \infty. \tag{7.1}$$

In what follows we assume also that

$$\frac{q_t}{q} \quad \text{is nonincreasing on } \mathbb{R}^+; \tag{Q3}$$

$$\limsup_{t\to 0^+} \frac{tq_t(t)}{q(t)} < \infty. \tag{Q4}$$

THEOREM 7.1. *If $u_1 = u(\cdot\,, m_1)$ has a dead core B_{S_1}, then $u_2 = u(\cdot\,, m_2)$, $m_2 < m_1$, has a dead core B_{S_2}, with $S_2 > S_1$. On the other hand, if either $u_1(0) > 0$ or $u_1(0) = 0$ and $u_1(r) > 0$ for $r \in (0, R]$, then $u_2 > u_1$ in B_R when $m_2 > m_1$.*

Proof. It is equivalent to show the same result with u_1 and u_2 replaced by $v_1 = v_1(t) = u_1(r(t))$ and $v_2 = v_2(t) = u_2(r(t))$, $t \in [0, T]$, respectively. To do this we can repeat the proof of Theorem 4.1 almost word for word, of course with u_1 and u_2 replaced by v_1 and v_2, and also with the following necessary changes: (i) replace (4.2) by

$$\begin{aligned}\frac{1}{q(t)}\{[q(t)|v_{\varepsilon,t}|^{p-2}v_{\varepsilon,t}]_t - q(t)f(v_\varepsilon)\} &= [|v_{\varepsilon,t}|^{p-2}v_{\varepsilon,t}]_t + \frac{q_t(t)}{q(t)}|v_{\varepsilon,t}|^{p-2}v_{\varepsilon,t} - f(v_\varepsilon)\\ &= \left[\frac{q_t(t)}{q(t)} - \frac{q_t(t-\varepsilon)}{q(t-\varepsilon)}\right][v_{1,t}(t-\varepsilon)]^{p-1},\end{aligned}$$

which is nonpositive in view of (Q3) and the fact that $v_{1,t} \geq 0$, that is,

$$[q(t)|v_{\varepsilon,t}|^{p-2}v_{\varepsilon,t}]_t - q(t)f(v_\varepsilon) \leq 0;$$

and (ii) replace the use of Theorem 2.1 in B_{r_0} by the use of the radial analogue of Theorem 6.1 in $[0, r_0)$. □

The following theorem gives an important relation between the value m and dead core solutions of (6.13).

THEOREM 7.2. *Let $u(\cdot\,, m)$ be the unique solution of (6.13). Then either $u(\cdot\,, m)$ has a dead core for all $m > 0$, or there is a unique (finite) number*

$$m = m_0 = m_0(R) > 0$$

for which a solution $u_0 = u_0(r) = u_0(r\,, m_0)$ of (6.13) in B_R exists, with the properties that

(i) *$u_0(0) = 0$;*
(ii) *$u(0\,, m) > 0$ for every $m > m_0$;*
(iii) *$u(\cdot\,, m)$ has a dead core for every $0 < m < m_0$.*

For convenience we define $m_0 = m_0(R)$ to be ∞ when all solutions of (6.13) are such that $u(0\,, m) = 0$ for all $m > 0$.

Before proving Theorem 7.2, it is useful to give a preliminary lemma. For this and later purposes we introduce the constant $M \geq 1$ by

$$\sup_{0<t<T} \frac{tq_t(t)}{q(t)} = M - 1, \tag{7.2}$$

which is well defined by (Q1), (Q2), and (Q4).

LEMMA 7.3. *Any nonnegative solution u of (6.1) in B_R which is suitably small on ∂B_R has a dead core.*

Proof. Let $\bar{m} > 0$ be so small that

$$D_{\bar{m}} = \int_0^{\bar{m}} \frac{ds}{[F(s)/M]^{1/p}} = M^{1/p} \int_0^{\bar{m}} \frac{ds}{[F(s)]^{1/p}} < T, \qquad T = t(R), \tag{7.3}$$

which of course is possible by assumption (7.1), where $T = t(R)$ and t is given in (6.7). Define $\tilde{v}(t) = w(t - \tilde{S})$, $t \in [\tilde{S}, T]$, $\tilde{S} = T - D_{\bar{m}}$, where w is the function constructed in the dead core Lemma 7.1 of [7, 8], with $\sigma = 1/M$. By analogy with (7.4) of [7, 8] the function $\tilde{v}$ is a supersolution of (6.8) in the interval $[\tilde{S}, T]$; that is, for $t \in (\tilde{S}, T)$

$$\begin{aligned} \{(\tilde{v}_t(t))^{p-1}\}_t + \frac{q_t(t)}{q(t)}(\tilde{v}_t(t))^{p-1} - f(\tilde{v}(t)) &\le \left\{ \sigma \left[1 + (M-1)\frac{t - \tilde{S}}{t} \right] - 1 \right\} f(\tilde{v}(t)) \\ &\le (\sigma M - 1) f(\tilde{v}(t)) = 0, \end{aligned} \tag{7.4}$$

where we have used (iii) and (iv) of the dead core Lemma 7.1 of [7, 8]. Of course also $\tilde{v}(\tilde{S}) = \tilde{v}_t(\tilde{S}) = 0$ and $\tilde{v}(T) = \bar{m}$. Let $\tilde{u}(x) = \tilde{v}(t(|x|))$, $x \in B_R \setminus \overline{B_S}$, where $\tilde{S} = t(S)$. Then

$$\operatorname{div}(g(|x|)|D\tilde{u}|^{p-2} D\tilde{u}) - h(|x|) f(\tilde{u}) \le 0 \quad \text{in } B_R \setminus \overline{B_S}.$$

Moreover defining $\tilde{u}$ to be zero on B_S, the extended function $\tilde{u}$ is a supersolution in B_R of (6.1). Let the given solution u of (6.1) in B_R have nonnegative boundary values less than or equal to m. By the weak comparison Theorem 6.1 we get $u \le \tilde{u}$ whenever $m \le \bar{m}$. Therefore $u \equiv 0$ in B_S. This completes the proof. □

Proof of Theorem 7.2. Existence of u_0. Suppose that there is some $m > 0$ for which $u(0\,, m) > 0$ and define as before

$$m_0 = \inf\{m > 0 \ : \ u(0, m) > 0\}, \tag{7.5}$$

so that $m_0 > 0$ by Lemma 7.3.

The rest of the proof is exactly the same as the proof of Theorem 1.4, with the exception that Theorems 1.3 and 4.1 are replaced by the corresponding Theorems 6.3 and 7.1.

Uniqueness of u_0. The proof is exactly as for the corresponding result in Theorem 1.4.

The case $m_0 = \infty$. If every solution of (6.1) in B_R is such that $u(0\,, m) = 0$ for all $m > 0$, then $u(\cdot\,, m)$ has a dead core for all $m > 0$. Otherwise there would be $\bar{m} > 0$ for which $u(0\,, \bar{m}) = 0$ and $u(r\,, \bar{m}) > 0$ for $r \in (0, R]$. Hence $u(0\,, m) > 0$ for $m > \bar{m}$ by Theorem 7.1, contradicting the assumption. This also justifies the earlier agreement that $m_0 = \infty$ in this case. □

Remark. In summary, if m_0 is finite and $m > m_0$, then the solution $u = u(\cdot\,, m)$ of (6.13) is positive; namely, $u(r\,, m) > 0$ for all $r \in [0, R]$. On the other hand, if $m < m_0 \le \infty$ then the solution $u = u(\cdot\,, m)$ of (6.13) has a dead core $B_S \subset B_R$, $0 < S < R$.

8. The size of a dead core and the case of general domains for (6.1). Recall the assumption that $\Phi(\infty) = H(\infty) = \infty$, and let

$$D = \int_0^{\infty} \frac{ds}{[F(s)/M]^{1/p}}, \tag{8.1}$$

where $M \geq 1$ is the number given in (7.2). Clearly $0 < D \leq \infty$ since the integral is convergent at 0 by (7.1) and by Lemma 3.2 of [7, 8] with $\sigma = 1/M \leq 1$, but possibly divergent at ∞. The next two results are the analogues of Theorems 5.1 and 5.2.

THEOREM 8.1. *We have*

$$m_0 = \infty \quad \text{if} \quad D < \infty \quad \text{and} \quad R \geq r(D), \tag{8.2}$$

where r is the inverse function of change of variable t given in (6.7)*, while*

$$m_0 \geq \bar{m} \qquad \text{if } R < r(D), \tag{8.3}$$

where $\bar{m}$ is given by the relation $R = r(D_{\bar{m}})$.

Proof. The proof of (8.2) is essentially the same as that of Lemma 7.3, the only exception being that $D_{\bar{m}}$ is replaced by D, and R by $T = t(R)$.

To obtain (8.3) one again follows the outline of the proof of Lemma 7.3. Now, however, in addition to replacing $D_{\bar{m}}$ by D, we take $\tilde{S} = S = 0$. One thus obtains (7.4) for $\tilde{v}$, with $\tilde{v}(0) = \tilde{v}_t(0) = 0$ and $\tilde{v}(T) = \bar{m}$. Hence $\tilde{u}(0) = \tilde{u}'(0) = 0$ and $\tilde{u}(R) = \bar{m}$, where $\tilde{u}(r) = \tilde{v}(t(r))$. The conclusion is that $0 \leq \tilde{u}(r\,, \bar{m}) \leq u(r\,, \bar{m})$, by virtue of the weak comparison Theorem 6.1. Hence $\tilde{u}(0\,, \bar{m}) = u(0\,, \bar{m}) = 0$, and in turn from the definition (7.5) of m_0 we get $m_0 \geq \bar{m}$, as required in (8.3). ∎

Remark. It is interesting to calculate the function $t \mapsto r(t)$ for the classical Matukuma equation, in which $g \equiv 1$ and $h(r) = 1/(1 + r^2)$, $p = 2$. One finds from (6.7) that $t(r) = \operatorname{arcsinh} r$, so that $r(t) = \sinh t$. Thus in this case (8.2) becomes

$$m_0 = \infty \quad \text{if } D < \infty \quad \text{and} \quad R \geq \sinh D,$$

with a similar relation for (8.3).

THEOREM 8.2. *Let $m < m_0$, so that a dead core exists in B_R by Theorem* 7.2*. In particular*

$$u \equiv 0 \qquad \text{in } B_S \subset B_R,$$

where $T = t(R)$, $\tilde{S} = t(S)$, and

$$T - \int_0^m \frac{ds}{[F(s)/M]^{1/p}} < \tilde{S} < T. \tag{8.4}$$

If $R \geq r(D)$, then for all $m > 0$ one has

$$R - r(D) < S < R.$$

Proof. This is the same as the proof of (8.2). ∎

Remark. Let $\varepsilon > 0$. Then if m is suitably small (depending on ε) we have $T - \varepsilon < \tilde{S} < T$ by (8.4).

The next result is the analogue for (6.1) of Theorem 1.1 for (1.1) and for general domains Ω. Let $0 \in \Omega$. As before we denote by B_R the open ball *centered at* 0 with radius $R > 0$, and we let

$$\overline{R} = \sup\{R > 0 \ : \ B_R \subset \Omega\}.$$

THEOREM 8.3. *Assume the dead core condition* (7.1) *holds and let u be a solution of* (6.1)*, with $0 \leq u(x) \leq m$ on $\partial\Omega$ for some positive constant m. Then the following properties are valid:*

(a) $0 \le u \le m$ *in* Ω *and* $0 \le u < m$ *in* $B_{\overline{R}}$.
(b) *If* $\overline{R} \ge r(D)$, *with* D *given in* (8.1), *then* u *has a dead core in* Ω *for all* $m > 0$.
(c) *Let* B_R *be compactly contained in* Ω. *Then* $u \equiv 0$ *in* B_R *provided that* $m > 0$ *is suitably small.*

Proof of Theorem 8.3. (a) That $u \ge 0$ follows by comparison (Theorem 6.1) of the given solution u with the trivial solution 0. Also the constant function m is a supersolution of (6.1), so that again by comparison $u \le m$ in Ω.

Moreover $u < m$ in $B_{\overline{R}}$. To see this, let $\overline{u}(\cdot\,, m)$ be the radial solution of (6.1) in $B_{\overline{R}}$ constructed in Theorem 6.2. Clearly $u(x) \le m = \overline{u}(|x|, m)$ for $x \in \partial B_{\overline{R}}$, and in turn $u(x) \le \overline{u}(|x|, m) < m$ for $x \in B_{\overline{R}}$ by the final part of Theorem 6.2.

(b) This is a direct consequence of Theorem 8.1.

(c) Let $\varepsilon > 0$ be so small that $B_R \subset B_{R+\varepsilon} \subset \Omega$. In analogy with part (a), let $\overline{u}(\cdot\,, m)$ be the solution of (6.13) with B_R replaced by $B_{R+\varepsilon}$. Then $u(x) \le \overline{u}(|x|\,, m) = m$ for $x \in \partial B_{R+\varepsilon}$. By the remark after Theorem 8.2 we have $\overline{u} \equiv 0$ in B_R when $m > 0$ is suitably small, and in turn $u \equiv 0$ in B_R by the weak comparison Theorem 6.1. □

9. The case $\Phi(\infty) < \infty$. This is the case, for example, for the mean curvature operator noted in the introduction, in which $\Phi(\infty) = H(\infty) = 1$. Here the proof of the critical Theorem 1.2 requires the modification that the parameter m in (1.5) should be restricted so that

$$f(m) + \Phi(m) < \left(\frac{R}{R+1}\right)^{n-1} \Phi(\infty) \qquad \text{and} \qquad F(m) \le nH(\infty); \tag{9.1}$$

see Proposition 4.1 of [7, 8]. Denote by m_∞ the supremum of all $m > 0$ satisfying (9.1).

Then the main results stated in the introduction remain true provided that the condition $m < m_\infty$ is assumed in all the statements. For instance we have the following analogue of Theorem 1.1.

THEOREM 9.1. *Assume the dead core condition* (1.3) *holds and let* u *be a solution of* (1.1), *with* $0 \le u(x) \le m$ *on* $\partial\Omega$ *for some positive constant* $m < m_\infty$. *Then the following properties are valid:*

(a) $0 \le u < m$ *in* Ω.
(b) *Assume that* $C = \int_0^{m_\infty} \frac{ds}{H^{-1}(F(s)/n)} < \infty$, *and let* B_R *be a ball with radius* $R \ge C$, *compactly contained in* Ω. *Then* u *has a dead core in* Ω *for all* $m \in (0, m_\infty)$.
(c) *If* B *is any ball compactly contained in* Ω, *then* $u \equiv 0$ *in* B *provided that* $m > 0$ *is suitably small.*

It is not hard to show that if $\Phi(\infty) = \infty$ then necessarily $H(\infty) = \infty$, but it is possible to have $\Phi(\infty) < \infty$ and $H(\infty) = \infty$, as shown by $A(\varrho) = 1/(1 + \sqrt{1+\varrho^2})$, with corresponding

$$H(\varrho) = \frac{1}{2}\left[\frac{\varrho^2}{1+\sqrt{1+\varrho^2}} - \log\frac{1+\sqrt{1+\varrho^2}}{2}\right].$$

In this example $\Phi(\infty) = 1$, while $H(\infty) = \infty$.

The case $H(\infty) < \infty$ for unrestricted $m > 0$ was treated by Siegel in [9].

10. Bursts. Here we assume (F1) and apply the previous theory to problems with nonlinearities f with (F2) replaced by the following:

(F3) f is nondecreasing in $(0,\delta)$, $F(u)>0$ on $(0,\beta)$, with $F(\beta)=0$, and there is $\gamma\in(\beta,\infty]$ such that the following conditions hold:
- (i) $f(u)<0$ on $[\beta,\gamma)$, $\quad f(\gamma)=0$ if $\gamma<\infty$;
- (ii) $\max_{[0,\beta]} F(u)+|F(\gamma)|<H(\infty)$ if $H(\infty)<\infty$;
- (iii) $\liminf_{u\to\infty}\frac{H^{-1}(|F(u)|)}{u}=0$ if $H(\infty)=|F(\gamma)|=\infty$ and $\Phi(\infty)<\infty$;
- (iv) $\liminf_{u\to\infty}\frac{H^{-1}(|F(u)|)}{u}<\infty$ if $H(\infty)=\Phi(\infty)=|F(\gamma)|=\infty$.

In (F3) for the case $\gamma=\infty$ we define $F(\gamma)=\lim_{u\to\infty}F(u)$, which certainly exists though possibly being $-\infty$. Clearly $F(\gamma)=-\infty$ can occur only if $\gamma=\infty$. As noted above, $H(\infty)=\infty$ whenever $\Phi(\infty)=\infty$.

For such functions f Franchi, Lanconelli, and Serrin proved in Theorem A of [4] that (1.1) *admits a nonnegative radial ground state $u=u(r)$, with central value $u(0)=\alpha\in(\beta,\gamma]$ if $\gamma<\infty$, or $u(0)=\alpha\in(\beta,\infty)$ if $\gamma=\infty$; moreover $u'(r)\le 0$ for all $r\ge 0$. Furthermore, since the dead core condition* (1.3) *is assumed to hold, then the solution is compactly supported.*

Remarks. Clearly, if $F(\gamma)>-\infty$ and $H(\infty)=\infty$ then only condition (i) is needed. For the Laplacian operator and the p-Laplacian degenerate operator $H(\infty)=\infty$ and $\Phi(\infty)=\infty$, so conditions (ii), (iii), and (iv) are not needed if $\gamma<\infty$, though (iv) must be used when $|F(\gamma)|=\infty$. On the other hand, for the mean curvature operator $H(\infty)=1$, and so in consequence of (ii), the function F must be quite restrictive to apply Theorem A of [4] in order to get existence of radially symmetric ground states for (1.1).

Consider next the Dirichlet problem

$$\begin{cases}\operatorname{div}\{A(|Du|)Du\}=f(u) & \text{in } B_R\subset\mathbb{R}^n,\quad n\ge 2,\\ u=\delta & \text{on } \partial B_R,\end{cases}\tag{10.1}$$

under the dead core condition (1.3) for functions f satisfying (F1) and (F3). If $H(\infty)=\infty$, define

$$C'=\int_0^{\delta}\frac{ds}{H^{-1}(F(s)/n)}<\infty,$$

by (1.3) and (F1). Then there exists, according to Theorems 1.2 and 5.2, a solution of (10.1) which has a dead core B_S, with $R-S<C'$. In other words the main equation (1.1) has two nonnegative solutions—the compact support *burst* shown above by Theorem A of [4], and the dead core solution just derived by Theorem 1.2.

These two solutions can clearly be superposed, even if problem (10.1) is nonlinear, to obtain another combined solution of (10.1), with a dead core $B_S\setminus B_T$, $T<S<R$, and a symmetrical *burst* at its center.

In particular, the Dirichlet problem (10.1), with the *loop* nonlinearity f verifying (F1) and (F3), neither has a *unique* solution nor obeys the ordinary maximum principle.

Since (1.1) in problem (10.1) is translation invariant, it is evident that the *burst* need not be centered in the ball B_R. Even more, if R is made many times larger, one can place multiple bursts into original dead cores.

The main open problem arising in the study of dead cores is the question of stability, particularly in dynamic situations for parabolic time dependent problems of the type, e.g.,

$$u_t=\operatorname{div}\{A(|Du|)Du\}-f(u),\qquad u=u(t,x),\qquad (t,x)\in\mathbb{R}^+\times\Omega.$$

Acknowledgments. The paper is based on a minicourse given by P. Pucci in September 2002, in Grado, Italy, at the workshop Stationary and Evolution Problems, supported by the *GNAMPA* of the Istituto Nazionale di Alta Matematica "F. Severi," and on the lecture of J. Serrin given in June 2003 in the series Lezioni Leonardesche, organized by the Mathematics departments of the two universities in Milan and of the Politechnic of Milan.

REFERENCES

[1] C. Bandle and S. Vernier-Piro, *Estimates for solutions of quasilinear problems with dead cores*, Z. Angew. Math. Phys., 54 (2003), pp. 815–821.

[2] J. Batt, W. Faltenbacher, and E. Horst, *Stationary spherically symmetric models in stellar dynamics*, Arch. Ration. Mech. Anal., 93 (1986), pp. 159–183.

[3] J.I. Diaz and L. Veron, *Local vanishing properties of solutions of elliptic and parabolic quasilinear equations*, Trans. Amer. Math. Soc., 290 (1985), pp. 787–814.

[4] B. Franchi, E. Lanconelli, and J. Serrin, *Existence and uniqueness of nonnegative solutions of quasilinear equations in* $\mathbb{R}^n$, Adv. Math., 118 (1996), pp. 177–243.

[5] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.

[6] P. Pucci, M. García-Huidobro, R. Manásevich, and J. Serrin, *Qualitative proprieties of ground states for singular elliptic equations with weights*, Ann. Mat. Pura Appl. (4), 185 (2006), pp. 205–243.

[7] P. Pucci and J. Serrin, *The strong maximum principle revisited*, J. Differential Equations, 196 (2004), pp. 1–66.

[8] P. Pucci and J. Serrin, *Erratum: "The strong maximum principle revisited,"* J. Differential Equations, 207 (2004), pp. 226–227.

[9] D. Siegel, *Height estimates for capillarity surfaces*, Pacific J. Math., 88 (1980), pp. 471–515.

[10] R. Sperb, *Some complementary estimates in the dead core problem*, in Nonlinear Problems in Applied Mathematics. In Honor of Ivar Stakgold on His 70th Birthday, T.S. Angell, et al., eds., Philadelphia, 1996, PA, pp. 217–224.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 279–308

ON SYMMETRIC FUNCTIONALS OF THE GRADIENT HAVING SYMMETRIC EQUIDISTRIBUTED MINIMIZERS*

ANDREA CIANCHI† AND ADELE FERONE‡

Abstract. Within a general family of functionals, not necessarily of integral type, depending on the modulus of the gradient, we characterize those possessing only spherically symmetric minimizers in classes of Sobolev functions with level sets of prescribed Lebesgue measures.

Key words. symmetrizations, constrained minimum problems, Sobolev spaces

AMS subject classifications. 46E35, 46E30

DOI. 10.1137/050625011

1. Introduction and main results. In a fundamental paper in the theory of rearrangements [BZ], Brothers and Ziemer elucidated minimal conditions ensuring that functionals that depend on the modulus of the gradient admit only spherically symmetric minimizers in classes of Sobolev functions whose level sets have prescribed Lebesgue measures. The functionals considered in [BZ] have the form

$$J_A(|\nabla u|), \tag{1.1}$$

where J_A is defined at a real-valued measurable function f in $\mathbb{R}^n$ as

$$J_A(f) = \int_{\mathbb{R}^n} A(|f(x)|)\,dx, \tag{1.2}$$

and A is a Young function, namely a convex function from $[0,\infty)$ into $[0,\infty]$ vanishing at 0. It has been long known under the name of the Pólya–Szegö principle that, if a function u belongs to $W^1_+(\mathbb{R}^n)$, the space of nonnegative functions from $W^{1,1}(\mathbb{R}^n)$ satisfying $\mathcal{L}^n(\{u>0\}) < +\infty$, then also its *spherically symmetric rearrangement* $u^\bigstar$ is in $W^1_+(\mathbb{R}^n)$, and

$$J_A(|\nabla u^\bigstar|) \le J_A(|\nabla u|). \tag{1.3}$$

Here, $\mathcal{L}^n$ denotes the Lebesgue measure. The contribution of [BZ] amounts to showing that if A satisfies a slightly stronger condition than just convexity, then any function $u \in W^1_+(\mathbb{R}^n)$, attaining equality in (1.3), necessarily equals $u^\bigstar$ $\mathcal{L}^n$-a.e. (up to translations), provided that

$$\mathcal{L}^n(\{\nabla u = 0\} \cap \{0 < u < \operatorname{ess\,sup} u\}) = 0. \tag{1.4}$$

An exhaustive discussion of the necessity of assumption (1.4) can be found in [BZ]. The condition on A appearing in that paper has a somewhat technical nature, but

*Received by the editors February 22, 2005; accepted for publication September 7, 2005; published electronically May 5, 2006.
http://www.siam.org/journals/sima/38-1/62501.html
†Dipartimento di Matematica e Applicazioni per l'Architettura, Università di Firenze, Piazza Ghiberti 27, 50122 Firenze, Italy (cianchi@unifi.it).
‡Dipartimento di Matematica, Seconda Università di Napoli, Via Vivaldi 43, 81100 Caserta, Italy (adele.ferone@unina2.it).

was later realized that strict convexity of A is sufficient (and necessary) for the conclusion to be true (see, e.g., [B], [CF1]; see also [FV2], where a more elementary proof of Brothers and Ziemer's theorem is offered, and [CF2], dealing with asymmetry estimates for u in the case where condition (1.4) and the strict convexity of A are dropped).

Apart from its own interest, this symmetry result is relevant in view of applications to the uniqueness of minimizers of variational problems. For instance, as recently observed in [CNV] and in [M], it can be used to prove that the only extremals in the Sobolev inequality with best constant are those exhibited in [Ta]. On the other hand, Brothers and Ziemer's theorem does not apply to certain functionals $\Phi(|\nabla u|)$, not in integral form, on which, nevertheless, Schwarz symmetrization is known to act monotonically. Functionals enjoying this property are, for example, when $\Phi(\,\cdot\,) = \|\,\cdot\,\|_A$, the Luxemburg norm, or $\Phi(\,\cdot\,) = \|\,\cdot\,\|_{(A)}$, the Orlicz norm, in the Orlicz space $L^A(\mathbb{R}^n)$ associated with a Young function A. Recall that these (equivalent) norms are defined at a measurable function f in $\mathbb{R}^n$ as

$$\|f\|_A = \inf\left\{\lambda > 0\colon \int_{\mathbb{R}^n} A\left(\frac{|f(x)|}{\lambda}\right) dx \le 1\right\} \tag{1.5}$$

and

$$\|f\|_{(A)} = \sup\left\{\int_{\mathbb{R}^n} |f(x)g(x)|\, dx\colon \int_{\mathbb{R}^n} \widetilde{A}(|g(x)|)dx \le 1\right\}, \tag{1.6}$$

respectively, where $\widetilde{A}(s) = \sup\{rs - A(r) : r \ge 0\}$, the Young conjugate of A.

Other functionals of the gradient not included in [BZ], but yet satisfying a Pólya–Szegö principle, arise from the definition of the Lorentz spaces $L^{p,q}(\mathbb{R}^n)$ (see, e.g., [A]). Given $p, q \in (0, +\infty]$, the relevant functionals are defined at a measurable function f in $\mathbb{R}^n$ by

$$\|f\|_{p,q} = \|s^{\frac{1}{p}-\frac{1}{q}} f^*(s)\|_{L^q(0,\infty)}, \tag{1.7}$$

where f^* is the *decreasing rearrangement* of f, and

$$\|f\|_{(p,q)} = \|s^{\frac{1}{p}-\frac{1}{q}} f^{**}(s)\|_{L^q(0,\infty)}, \tag{1.8}$$

where

$$f^{**}(s) = \frac{1}{s}\int_0^s f^*(r)dr \qquad \text{for } s > 0.$$

Notice that, if $1 < p \le \infty$ and $0 < q \le \infty$, the quantities $\|\ \|_{p,q}$ and $\|\ \|_{(p,q)}$ are equivalent (up to multiplicative constants)—see, e.g., [BS, Chapter 4, Lemma 4.5].

The fact that, when evaluated at $|\nabla u|$, functionals (1.5), (1.6), (1.8), and also (1.7) for suitable values of p and q do not increase under Schwarz symmetrization of u is a consequence of an extended version of the Pólya–Szegö principle stated in Theorem 1.2 below. This theorem deals with general functionals of the form $\Phi(|\nabla u|)$, with Φ defined in

$$L^1_+(\mathbb{R}^n) = \{f \in L^1(\mathbb{R}^n)\colon f \ge 0 \text{ and } \mathcal{L}^n(\{f > 0\}) < +\infty\}.$$

The key property of Φ in order for the functional $\Phi(|\nabla u|)$ to support a Pólya–Szegö principle is the $**$-monotonicity defined as follows.

DEFINITION 1.1. *A functional* $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ *is called* $**$-increasing *if*

$$\Phi(f) \le \Phi(g) \qquad \textit{whenever} \qquad f^{**}(s) \le g^{**}(s) \quad \textit{for } s > 0.$$

Besides those considered above, other examples of $**$-increasing functionals are provided by all rearrangement invariant norms in the sense of Luxemburg, namely norms in Banach function spaces depending only on the decreasing rearrangement—see Proposition 5.2, section 5. However, rearrangement invariant norms do not exhaust the class of $**$-increasing functionals, as the case where $\Phi(\cdot) = \|\cdot\|_{(p,q)}$, with $0 < q < 1$, demonstrates. Relations between $**$-monotonicity and other notions of monotonicity, convexity, and semicontinuity will be examined in sections 2 and 5.

THEOREM 1.2. *Let* $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ *be a* $**$*-increasing functional. Then*

$$\Phi(|\nabla u^\bigstar|) \le \Phi(|\nabla u|) \tag{1.9}$$

for every $u \in W^1_+(\mathbb{R}^n)$.

Theorem 1.2 is essentially known in literature as a consequence of the fact that every $u \in W^1_+(\mathbb{R}^n)$ satisfies

$$|\nabla u^\bigstar|^{**}(s) \le |\nabla u|^{**}(s) \qquad \text{for } s > 0 \tag{1.10}$$

(see, e.g., [CP], [F], [K]). Here, we shall present a proof of Theorem 1.2 where an intermediate step is enucleated, showing that a third quantity, involving $|\nabla u|$, always lies between the left-hand side and the right-hand side of (1.9) (see also [RT] and [FV1] for contributions in this direction). This is crucial in preparation for our main results concerning the equality case in (1.9).

Indeed, the purpose of the present paper is to complement the result of [BZ] and to characterize the class of those $**$-increasing functionals Φ for which the equation

$$\Phi(|\nabla u^\bigstar|) = \Phi(|\nabla u|) < +\infty, \tag{1.11}$$

coupled with condition (1.4), necessarily implies $u = u^\bigstar$ (up to translations). According to the terminology of [BZ], we shall call any $u \in W^1_+(\mathbb{R}^n)$ satisfying (1.11) a *minimal rearrangement* of Φ.

In fact, we give a more detailed picture of the situation at hand. In order to grasp the spirit of our discussion, recall that the proof of the symmetry result for extremals in the standard inequality (1.3), where $\Phi = J_A$, consists of two steps. First, (1.11) is used to deduce that

$$\text{the level sets } \{u > t\} \text{ are (equivalent to) balls} \quad \text{for } \mathcal{L}^1\text{-a.e. } t > 0, \tag{1.12}$$

and that

$$|\nabla u| \text{ is constant } \mathcal{H}^{n-1}\text{-a.e. on } \{u = t\} \quad \text{for } \mathcal{L}^1\text{-a.e. } t > 0, \tag{1.13}$$

where $\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure. Combining (1.12) and (1.13) with (1.4) then entails that the balls $\{u > t\}$ must be concentric, and hence that u is symmetric. In particular, we stress that (1.11) plays a role in the derivation only of (1.12) and (1.13). In view of these facts, the problem arises of characterizing all $**$-increasing functionals having the property that every minimal rearrangement u necessarily fulfills (1.12) or (1.13). Let us emphasize that the issue is not immaterial, since not every $**$-increasing functional enjoys this property. A typical example in

this connection is provided by the L^∞-norm, since functions u can be easily exhibited which fulfill (1.11) with $\Phi(\,\cdot\,) = \|\cdot\|_{L^\infty}$, but not (1.12) (nor, a *fortiori*, (1.13)—see the comment following Theorem 1.6 below). Our first main result tells us that failure of (1.12) can occur only for minimal rearrangements of $**$-increasing functionals, which, like $\|\ \|_{L^\infty}$, are not strictly monotone in a sense specified in the next definition.

DEFINITION 1.3. *A $**$-increasing functional $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ is said to be* $*$-strictly increasing *if*

$$f^*(s) \le g^*(s) \text{ for } s \ge 0 \quad \text{and} \quad \Phi(f) = \Phi(g) < +\infty \text{ imply that } f^* \equiv g^*.$$

Actually, the following theorem holds.

THEOREM 1.4. *Let $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ be a $**$-increasing functional. Then every minimal rearrangement u of Φ necessarily fulfills* (1.12) *if and only if Φ is $*$-strictly increasing.*

Theorem 1.4 settles the question concerning the validity of (1.12) for minimal rearrangements u. The parallel issue for (1.13) turns out to be related to a stronger assumption on Φ, which, loosely speaking, amounts to a proper weak notion of strict convexity. The necessity of an assumption of this kind is suggested by another simple instance corresponding to the choice $\Phi(\,\cdot\,) = \|\cdot\|_{L^1}$. Since such a functional is clearly $*$-strictly increasing, by Theorem 1.4 every minimal rearrangement u satisfies (1.12); however, such a u does not necessarily fulfill (1.13), since the balls $\{u > t\}$ at different levels t need not be concentric. This assertion can be verified via the coarea formula, and remains true even for any other functional given by (1.2) with a nonstrictly convex A.

A sharp assumption on Φ guaranteeing property (1.13) for every minimal rearrangement u is introduced in the next definition.

DEFINITION 1.5. *A $**$-increasing functional $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ is said to be* $**$-strictly increasing *if*

$$f^{**}(s) \le g^{**}(s) \text{ for } s > 0 \quad \text{and} \quad \Phi(f) = \Phi(g) < +\infty \text{ imply that } f^* \equiv g^*.$$

Connections among $**$-strict monotonicity, classical strict convexity, and other related notions will be discussed in Proposition 5.3 of section 5.

We are now in position to state our second main result.

THEOREM 1.6. *Let $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ be a $**$-increasing functional. Then every minimal rearrangement u of Φ necessarily fulfills* (1.13) *if and only if Φ is $**$-strictly increasing.*

Notice that, since every $**$-strictly increasing functional is also $*$-strictly increasing, Theorems 1.4 and 1.6 ensure, in particular, that (1.12) is fulfilled by every minimal rearrangement u of Φ whenever (1.13) is.

Thanks to Theorems 1.4 and 1.6, any minimal rearrangement u of a $**$-strictly increasing functional Φ has to satisfy (1.12)–(1.13), and hence, if also (1.4) is in force, it must be spherically symmetric. Thus, the following general version of Brothers and Ziemer's theorem holds.

THEOREM 1.7. *Let $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ be a $**$-strictly increasing functional. Let u be any minimal rearrangement u of Φ satisfying* (1.4). *Then $u \equiv u^\bigstar$ $\mathcal{L}^n$-a.e in $\mathbb{R}^n$ (up to translations).*

Let us mention that an extension of the theorem of [BZ] in the special case where $\Phi(\,\cdot\,) = \|\cdot\|_{(p,q)}$ can also be found in [Ra].

Theorems 1.2, 1.4, and 1.6 are established in sections 2, 3, and 4, respectively. Applications of these theorems to functionals (1.1) and (1.5)–(1.8) are presented in

section 5; let us just anticipate here that some of the conclusions are somewhat surprising (to us at least). For instance, as a consequence of Theorem 1.4 and Proposition 5.7, strictly convex Young functions A exist, like $A(t) = e^{t^\alpha} - 1$ with $\alpha > 1$, such that

$$\| \, |\nabla u^\bigstar| \, \|_A = \| \, |\nabla u| \, \|_A$$

for some $u \in W^1_+(\mathbb{R}^n)$ fulfilling (1.4), which is not symmetric and does not even satisfy (1.12).

2. The Pólya–Szegö inequality for $**$-increasing functionals. In this section we provide a proof of Theorem 1.2. After recalling a few basic definitions about rearrangements, our first concern will be, however, to clarify the connections between the notion of $**$-monotonicity of functionals, upon which Theorem 1.2 relies, and other definitions of monotonicity which play a role in what follows.

Given any function $f \in L^1_+(\mathbb{R}^n)$, the function $\mu_f : [0, +\infty) \to [0, +\infty)$, defined by

$$\mu_f(t) = \mathcal{L}^n(\{x \in \mathbb{R}^n : f(x) > t\}) \qquad \text{for } t \geq 0,$$

is called the *distribution function* of f. The *decreasing rearrangement* of f is the function $f^* : [0, +\infty) \to [0, +\infty)$ obeying

$$f^*(s) = \sup\{t \geq 0 : \mu_f(t) > s\} \qquad \text{for } s \geq 0.$$

The *spherically symmetric rearrangement* $f^\bigstar : \mathbb{R}^n \to [0, +\infty)$ of f is defined as

$$f^\bigstar(x) = f^*(\omega_n |x|^n) \qquad \text{for } x \in \mathbb{R}^n, \tag{2.1}$$

where ω_n denotes the measure of the unit ball in $\mathbb{R}^n$. Functions having the same distribution function will be called *equidistributed* or *equimeasurable*. Clearly, f, f^*, and $f^\bigstar$ are equidistributed.

A basic property of rearrangements tells us that, if f is as above, then

$$\sup\left\{\int_E f(x)dx : \mathcal{L}^n(E) = s\right\} = \int_0^s f^*(r)dr \qquad \text{for every } s > 0 \tag{2.2}$$

and

$$\int_{\mathbb{R}^n} f(x)dx = \int_0^{+\infty} f^*(r)dr.$$

In what follows, given f and g in $L^1_+(\mathbb{R}^n)$, we write $f \prec g$ to denote that

$$f^{**}(s) \leq g^{**}(s) \quad \text{ for } s > 0 \quad \text{ and } \qquad \int_{\mathbb{R}^n} f(x)dx = \int_{\mathbb{R}^n} g(x)dx.$$

DEFINITION 2.1. *Let Φ be a functional from $L^1_+(\mathbb{R}^n)$ into $[0, \infty]$.*

(i) *Φ is called* pointwise increasing *if*

$$\Phi(f) \leq \Phi(g) \quad \textit{whenever} \quad f(x) \leq g(x) \quad \textit{for } \mathcal{L}^n\textit{-a.e. } x \in \mathbb{R}^n;$$

Φ is called strictly pointwise increasing *if, in addition,*

$$f(x) \leq g(x) \textit{ for } \mathcal{L}^n\textit{ -a.e. } x \in \mathbb{R}^n \quad \textit{and} \quad \Phi(f) = \Phi(g) < +\infty \textit{ imply } f = g \textit{ for } \mathcal{L}^n\textit{ -a.e. } x \in \mathbb{R}^n.$$

(ii) Φ *is called* $*$-increasing *if*

$$\Phi(f) \le \Phi(g) \quad \textit{whenever} \quad f^*(s) \le g^*(s) \quad \textit{for} \ \ s > 0;$$

Φ *is called* $*$-strictly increasing *if, in addition,*

$$f^*(s) \le g^*(s) \quad \textit{for } \mathcal{L}^1\textit{-a.e. } s > 0 \ \ \textit{and} \ \ \Phi(f) = \Phi(g) < +\infty \ \textit{imply } f^* \equiv g^*.$$

(iii) Φ *is called* $\prec$-increasing *if*

$$\Phi(f) \le \Phi(g) \quad \textit{whenever} \quad f \prec g;$$

Φ *is called* $\prec$-strictly increasing *if, in addition,*

$$f \prec g \quad \textit{and} \quad \Phi(f) = \Phi(g) < +\infty \ \ \textit{imply} \ \ f^* \equiv g^*.$$

We give one more definition involving functionals on $L^1_+(\mathbb{R}^n)$.

DEFINITION 2.2. *A functional* $\Phi : L^1_+(\mathbb{R}^n) \to [0, \infty]$ *is called* rearrangement invariant *if*

$$\Phi(f) = \Phi(g) \quad \textit{whenever} \quad f^* \equiv g^*.$$

Relations among the above notions of monotonicity are established in the next proposition.

PROPOSITION 2.3. *Let* Φ *be a functional from* $L^1_+(\mathbb{R}^n)$ *into* $[0, \infty]$. *Then*

(i) Φ *is* $*$*-increasing [respectively* $*$*-strictly increasing] if and only if* Φ *is pointwise increasing [strictly pointwise increasing] and rearrangement invariant.*

(ii) Φ *is* $**$*-increasing [*$**$*-strictly increasing] if and only if* Φ *is* $*$*-increasing [*$*$*-strictly increasing] and* $\prec$*-increasing [*$\prec$*-strictly increasing].*

Proof. (i) If Φ is $*$-increasing, it is obviously rearrangement invariant. It is also pointwise increasing, since if $f, g \in L^1_+(\mathbb{R}^n)$ and $f(x) \le g(x)$ for $\mathcal{L}^n$-a.e. $x \in \mathbb{R}^n$, then $f^*(s) \le g^*(s)$ for $s > 0$, and hence $\Phi(f) \le \Phi(g)$.

Conversely, assume that Φ is pointwise increasing and rearrangement invariant. Let f and g be functions from $L^1_+(\mathbb{R}^n)$ such that $f^*(s) \le g^*(s)$ for $s > 0$. Consequently,

$$\Phi(f) = \Phi(f^\bigstar) \le \Phi(g^\bigstar) = \Phi(g),$$

where the equalities hold because Φ is rearrangement invariant, and the inequality is due to the fact that Φ is pointwise increasing. Thus, Φ is $*$-increasing.

As for the assertions about strict monotonicity, if Φ is $*$-strictly increasing, then it must be also strictly pointwise increasing. Otherwise, there would exist $f, g \in L^1_+(\mathbb{R}^n)$, satisfying $f(x) \le g(x)$ for $\mathcal{L}^n$-a.e. $x \in \mathbb{R}^n$, $f(x) < g(x)$ for x in a set of positive Lebesgue measure, and $\Phi(f) = \Phi(g)$. The first two conditions entail that $f^*(s) \le g^*(s)$ for $s > 0$ and that $f^*(s) < g^*(s)$ for s in a set of positive measure, and these facts, combined with the equality $\Phi(f) = \Phi(g)$, contradict the $*$-strict monotonicity of Φ.

Suppose now that Φ is strictly pointwise increasing and rearrangement invariant, but not $*$-strictly increasing. Then there exist $f, g \in L^1_+(\mathbb{R}^n)$ fulfilling $f^*(s) \le g^*(s)$ for $s > 0$, $f^*(s) < g^*(s)$ for s in a set of positive measure, and $\Phi(f) = \Phi(g)$. The same argument as above shows that the functions $f^\bigstar$ and $g^\bigstar$ violate the strict pointwise monotonicity of Φ.

(ii) The fact that any $**$-increasing functional is both $*$-increasing and $\prec$-increasing is trivial. To prove the converse, choose any $f, g \in L^1_+(\mathbb{R}^n)$ such that

$$f^{**}(s) \leq g^{**}(s) \qquad \text{for } s > 0. \tag{2.3}$$

If $\int_{\mathbb{R}^n} f(x)dx = \int_{\mathbb{R}^n} g(x)dx$, then $f \prec g$, whence $\Phi(f) \leq \Phi(g)$ by the $\prec$-monotonicity of Φ. Suppose, instead, that $\int_{\mathbb{R}^n} f(x)dx < \int_{\mathbb{R}^n} g(x)dx$. Then, there exists $s_0 > 0$ such that

$$\int_{\mathbb{R}^n} f(x)dx = \int_0^{s_0} g^*(s)ds. \tag{2.4}$$

Set $h(x) = g^\bigstar(x)\chi_{[0,s_0]}(\omega_n|x|^n)$, where χ_E denotes the characteristic function of a set E. Clearly, $h^*(s) = g^*(s)\chi_{[0,s_0]}(s) \leq g^*(s)$ for $s > 0$. Thus, since Φ is $*$-increasing,

$$\Phi(h) \leq \Phi(g). \tag{2.5}$$

On the other hand, by (2.3) and (2.4), $f \prec h$, whence, by the $\prec$-monotonicity of Φ,

$$\Phi(f) \leq \Phi(h). \tag{2.6}$$

Combining (2.5) and (2.6) yields the $**$-monotonicity of Φ.

As far as strict monotonicity properties are concerned, it is again clear that any $**$-strictly increasing functional is also $*$-strictly increasing and $\prec$-strictly increasing. Conversely, if Φ is both $*$-strictly increasing and $\prec$-strictly increasing, and if $f, g \in L^1_+(\mathbb{R}^n)$ satisfy (2.3) and $\Phi(f) = \Phi(g)$, then equality holds in (2.5) and (2.6). The $*$-strict monotonicity and the $\prec$-strict monotonicity then imply that $g^* \equiv h^* \equiv f^*$. Hence, Φ is $**$-strictly increasing. □

We now come to the proof of Theorem 1.2. Basic tools are the isoperimetric inequality in $\mathbb{R}^n$ and the coarea formula for Sobolev functions. De Giorgi's isoperimetric theorem in $\mathbb{R}^n$ states that

$$\mathcal{H}^{n-1}(\partial^* E) \geq n\omega_n^{1/n}\mathcal{L}^n(E)^{1-1/n} \tag{2.7}$$

for every measurable subset E of $\mathbb{R}^n$ having finite measure, and that equality holds in (2.7) if and only if E is equivalent to a ball. Here, $\partial^* E$ denotes the reduced boundary of E.

The coarea formula entails that if $u \in W^1_+(\mathbb{R}^n)$ and $f : \mathbb{R}^n \to [0, \infty)$ is a Borel function, then

$$\int_{\mathbb{R}^n} |\nabla u|(x) f(x)dx = \int_0^\infty dt \int_{\partial^*\{u>t\}} f(x)d\mathcal{H}^{n-1}(x) \tag{2.8}$$

(see, e.g., [AFP], [Z]). Notice that, if a suitable precise representative of u is chosen, a choice that will be always tacitly made throughout, then $\partial^*\{u > t\}$ and $\{u = t\}$ agree, up to a set of zero $\mathcal{H}^{n-1}$-measure, for $\mathcal{L}^1$-a.e. $t > 0$, and hence (2.8) reads

$$\int_{\mathbb{R}^n} |\nabla u|(x) f(x)dx = \int_0^\infty dt \int_{\{u=t\}} f(x)d\mathcal{H}^{n-1}(x). \tag{2.9}$$

Another ingredient in our proof of Theorem 1.2 is the following lemma (see [AT], [FV2]).

LEMMA 2.4. *Let $u \in W^1_+(\mathbb{R}^n)$. Then the function given by*

$$s \mapsto \int_{u>u^*(s)} |\nabla u(x)|dx \qquad \text{for } s \geq 0$$

is absolutely continuous in $[0,\infty)$. Moreover, on setting

$$|\nabla u|_{*u}(s) = \frac{d}{ds}\int_{u>u^*(s)} |\nabla u|(x)dx \qquad \text{for } s \geq 0$$

and

$$|\nabla u|_{\bigstar u}(x) = |\nabla u|_{*u}(\omega_n |x|^n) \qquad \text{for } x \in \mathbb{R}^n$$

we have

$$|\nabla u|_{\bigstar u} \prec |\nabla u|. \tag{2.10}$$

Proof of Theorem 1.2. By the coarea formula (2.8),

$$|\nabla u|_{*u}(s) = -\frac{du^*}{ds}(s)\mathcal{H}^{n-1}(\partial^*\{u > u^*(s)\}) \qquad \text{for } \mathcal{L}^1\text{-a.e. } s > 0. \tag{2.11}$$

Denote by $\{t_i\}_{i\in P}$, with $P \subseteq \mathbb{N}$, the (possibly empty) family of all levels $t > 0$ satisfying

$$\mathcal{L}^n(\{u = t_i\}) = \mathcal{L}^1(\{u^* = t_i\}) > 0 \qquad \text{for } i \in P,$$

and by $\{I_i\}_{i\in P}$ the corresponding family of intervals in $[0, \mathcal{L}^n(\{u > 0\})]$ such that $(u^*)_{|I_i} = t_i$ for $i \in P$. Then,

$$\mathcal{L}^n(\{u > u^*(s)\}) = s \quad \text{for every} \quad s \in [0, \mathcal{L}^n(\{u > 0\})] \setminus \left(\cup_{i\in P} I_i\right) \tag{2.12}$$

and

$$\frac{du^*}{ds}(s) = 0 \qquad \text{for } \mathcal{L}^1\text{-a.e. } s \in \cup_{i\in P} I_i.$$

Thus, from (2.11) and from the isoperimetric inequality (2.7), we deduce that

$$n\omega_n^{1/n} s^{1-1/n}\left(-\frac{du^*}{ds}(s)\right) \leq |\nabla u|_{*u}(s) \qquad \text{for } \mathcal{L}^1\text{-a.e. } s > 0. \tag{2.13}$$

Hence, owing to (2.1),

$$|\nabla u^\bigstar(x)| \leq |\nabla u|_{\bigstar u}(x) \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \mathbb{R}^n. \tag{2.14}$$

Combining (2.14) and (2.10) yields

$$\Phi(|\nabla u^\bigstar|) \leq \Phi(|\nabla u|_{\bigstar u}) \leq \Phi(|\nabla u|), \tag{2.15}$$

where the first inequality holds because Φ is, in particular, pointwise increasing, and the second one because Φ is also $\prec$-increasing (see Proposition 2.3). □

3. Extremals of $*$-strictly increasing functionals. This section is devoted to Theorem 1.4. Its proof, as well as the proof of Theorem 1.6 contained in the next section, makes use of measure preserving maps. Recall that, given subsets Ω of $\mathbb{R}^n$ and G of $\mathbb{R}^m$ having finite Lebesgue measure, a function $\gamma : G \to \Omega$ is called a *measure preserving* (briefly *m.p.*) *map* if

$$\mathcal{L}^m(\gamma^{-1}(E)) = \mathcal{L}^n(E)$$

for every measurable subset E of Ω. Notice that, on modifying, if necessary, γ on a set of Lebesgue measure zero, we may assume without loss of generality that γ is a Borel function.

Since any measurable function $f : \Omega \to [0, +\infty)$ is equidistributed with the function $f \circ \gamma : G \to [0, +\infty)$, we have

$$\int_G f(\gamma(y))dy = \int_\Omega f(x)dx. \tag{3.1}$$

If $\sigma : \Omega \to [0, \mathcal{L}^n(\Omega)]$ is a bijective m.p. map, then $\sigma^{-1} : [0, \mathcal{L}^n(\Omega)] \to \Omega$ is also an m.p. map, and hence (3.1) yields

$$\int_0^{\mathcal{L}^n(\Omega)} f(\sigma^{-1}(s))ds = \int_\Omega f(x)dx. \tag{3.2}$$

In what follows, we shall repeatedly exploit the fact that if $\sigma : \Omega \to [0, \mathcal{L}^n(\Omega)]$ is any m.p. map, $\phi : [0, \mathcal{L}^n(\Omega)] \to [0, \infty)$ is nondecreasing and right-continuous, and $u : \mathbb{R}^n \to [0, \infty)$ is given by

$$u(x) = \begin{cases} \phi \circ \sigma(x) & \text{if } x \in \Omega, \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

then, by the equimeasurability of u and ϕ,

$$u^*(s) = \begin{cases} \phi(s) & \text{if } s \in [0, \mathcal{L}^n(\Omega)), \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Proof of Theorem 1.4. Assume that Φ is $*$-strictly increasing. Let u be any minimal rearrangement of Φ. Since u satisfies (1.11), equality holds in the first inequality in (2.15). Inasmuch as Φ is $*$-strictly increasing, by Proposition 2.3 it is also pointwise strictly increasing. An inspection of the proof of Theorem 1.2 then reveals that equality must hold in (2.14) for $\mathcal{L}^n$-a.e. $x \in \mathbb{R}^n$, and hence in (2.13) for $\mathcal{L}^1$-a.e. $s > 0$. In conclusion, owing to (2.11), one has that

$$\mathcal{H}^{n-1}\left(\partial^*\{u > u^*(s)\}\right) = n\omega_n^{1/n} s^{1-1/n} \tag{3.5}$$

for $\mathcal{L}^1$-a.e. $s > 0$ such that $\frac{du^*}{ds}(s)$ exists and is different from zero. Set

$$D_1 = \{s > 0 : u^* \text{ is not differentiable at } s\}$$

and

$$D_2 = \left\{ s > 0 : \frac{du^*}{ds}(s) = 0 \right\},$$

and observe that the intervals $\{I_i\}_{i\in P}$ introduced in the proof of Theorem 1.2 satisfy $\cup_{i\in P} I_i \subseteq D_1 \cup D_2$. Thus, owing to (2.12), equation (3.5) reads

$$\mathcal{H}^{n-1}\left(\partial^*\{u > u^*(s)\}\right) = n\omega_n^{1/n}\mathcal{L}^n\left(\{u > u^*(s)\}\right)^{1-1/n} \tag{3.6}$$

for $\mathcal{L}^1$-a.e. $s \notin D_1 \bigcup D_2$. Since u^* is locally absolutely continuous on $(0,\infty)$, then $\mathcal{L}^1(D_1) = 0$, and since absolutely continuous functions map sets of measure zero into sets of measure zero, $\mathcal{L}^1(u^*(D_1)) = 0$ as well. We also have $\mathcal{L}^1(u^*(D_2)) = 0$ (see, e.g., [CF1, Lemma 2.4]). Consequently, one can infer from (3.6) that

$$\mathcal{H}^{n-1}\left(\partial^*\{u^* > t\}\right) = n\omega_n^{1/n}\mathcal{L}^n\left(\{u > t\}\right)^{1-1/n} \qquad \text{for } \mathcal{L}^1\text{-a.e. } t \in (0, \operatorname{ess\,sup} u).$$

In other words, equality holds in the isoperimetric inequality (2.7) with $E = \{u > t\}$ for $\mathcal{L}^1$-a.e. $t \in (0, \operatorname{ess\,sup} u)$. Hence, (1.12) follows.

Conversely, assume that every function satisfying (1.11) necessarily fulfills (1.12), and suppose, by contradiction, that Φ is not $*$-strictly increasing. Then there exist $f, g \in L^1_+(\mathbb{R}^n)$ and $s_0 > 0$ satisfying

$$f^*(s) \le g^*(s) \qquad \text{for every } s > 0, \tag{3.7}$$

$$\Phi(f) = \Phi(g) < +\infty, \tag{3.8}$$

and

$$0 < f^*(s_0) < g^*(s_0).$$

This information will enable us to exhibit a function $u \in W^1_+(\mathbb{R}^n)$, in the form (3.3), fulfilling (1.11) but not (1.12). First, choose $\Omega = \{x : \omega_N|x|^n \le \mathcal{L}^n(\{g > 0\})\}$ and define $\phi : [0, \mathcal{L}^n(\Omega)] \to [0, +\infty)$ as

$$\phi(s) = \int_s^{\mathcal{L}^n(\Omega)} \frac{f^*(r)}{n\omega_n^{1/n} r^{1-1/n}}\, dr \qquad \text{for } s \in [0, \mathcal{L}^n(\Omega)]. \tag{3.9}$$

Next, we construct σ. Consider any family $\mathcal{D} = \{D(s)\}_{s\in[0,\mathcal{L}^n(\Omega)]}$ of measurable subsets of Ω satisfying

$$D(s_1) \subseteq D(s_2) \qquad \text{if} \quad 0 \le s_1 \le s_2 \le \mathcal{L}^n(\Omega) \tag{3.10}$$

and

$$\mathcal{L}^n(D(s)) = s \qquad \text{for } s \in [0, \mathcal{L}^n(\Omega)]. \tag{3.11}$$

If σ is given by

$$\sigma(x) = \inf\{s > 0 : x \in D(s)\} \qquad \text{for } x \in \Omega, \tag{3.12}$$

then

$$\sigma^{-1}\left((0, s)\right) = D(s) \qquad \text{for } s \in [0, \mathcal{L}^n(\Omega)].$$

Hence, σ is an m.p. map, and its level sets belong to the family $\mathcal{D}$. The point is now to choose the sets $D(s)$ in such a way that the function u given by (3.3), with ϕ as

in (3.9), enjoys the desired properties. To this purpose, suppose for a moment that $D(s) = \{x \in \mathbb{R}^n : \omega_n |x|^n < s\}$ for $s > 0$, and observe that, in this case, $\sigma(x) = \omega_n |x|^n$, $u \equiv u^\bigstar$, and hence, in particular, (1.11) certainly holds. The idea is then to slightly perturb the balls $\{x \in \mathbb{R}^n : \omega_n |x|^n < s\}$ for s close to s_0, and exploit (3.8) to deduce that the function u, associated with this perturbed function σ, still fulfills (1.11). In order to make this idea precise, note that, since f^* and g^* are right continuous, there exist positive numbers ϵ and δ such that

$$f^*(s)(1+\delta) < g^*(s) \qquad \text{for every } s \in [s_0, s_0+\epsilon]. \tag{3.13}$$

Then we define

$$D(s) = \begin{cases} \{x \in \mathbb{R}^n : \omega_n |x|^n < s\} & \text{if } s \notin [s_0, s_0+\epsilon], \\ \left\{ x \in \mathbb{R}^n : \omega_n \left[\dfrac{x_1^2}{a^2(s)} + \displaystyle\sum_{i=2}^n \dfrac{x_i^2}{b^2(s)} \right]^{n/2} < s \right\} & \text{if } s \in [s_0, s_0+\epsilon], \end{cases} \tag{3.14}$$

where the functions a and b are sufficiently smooth, nonconstant in $[s_0, s_0+\epsilon]$, and satisfy the following properties:

$$a(s) b^{n-1}(s) = 1 \quad \text{for} \quad s \in [s_0, s_0+\epsilon], \tag{3.15}$$

$$a(s_0) = a(s_0+\epsilon) = 1, \tag{3.16}$$

$$\min\left\{ \left(a(s)s^{1/n}\right)', \left(b(s)s^{1/n}\right)' \right\} \left(\frac{s_0}{s_0+\epsilon} \right)^{1/n} \geq \frac{s^{-1+1/n}}{n(1+\delta)} \qquad \text{for } s \in [s_0, s_0+\epsilon]. \tag{3.17}$$

Here, prime stands for differentiation. In particular, by (3.17), the functions $s \mapsto a(s)s^{1/n}$ and $s \mapsto b(s)s^{1/n}$ are increasing in $[s_0, s_0+\epsilon]$. A possible choice is, for instance,

$$a(s) = \left[\frac{(s_0+\epsilon)^{1/n} - s_0^{1/n}}{\epsilon} (s - s_0) + s_0^{1/n} \right] s^{-1/n}$$

and

$$b(s) = a(s)^{-1/(n-1)},$$

provided that ϵ is sufficiently small. Notice that (3.16) along with the monotonicity of $a(s)s^{1/n}$ and $b(s)s^{1/n}$ ensure that (3.10) is fulfilled, whereas (3.15) implies (3.11). Condition (3.17), combined with (3.13), will be used to show that

$$\Phi(|\nabla u^\bigstar|) = \Phi(|\nabla u|). \tag{3.18}$$

Clearly, this yields the announced contradiction, since the level sets of u whose measure lies between s_0 and $s_0+\epsilon$ are not balls.

To verify (3.18), observe that, by (3.3) and (3.9),

$$|\nabla u(x)| = \frac{f^*(\sigma(x))}{n\omega_n^{1/n} \sigma(x)^{1-1/n}} |\nabla \sigma(x)| \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \Omega. \tag{3.19}$$

Such an equation holds since σ is Lipschitz continuous. The Lipschitz continuity of σ is, in turn, a consequence of the fact that, by (3.12) and (3.14), $\sigma(x)$ equals $\omega_n|x|^n$ when $\omega_n|x|^n \notin [s_0, s_0+\epsilon]$, and that it is implicitly defined by the equation

$$K(x, \sigma(x)) = 1 \tag{3.20}$$

otherwise, where $K : \mathbb{R}^n \times (0,\infty) \to \mathbb{R}$ is given by

$$K(x,s) = \Big(\frac{\omega_n}{s}\Big)^{2/n} \frac{x_1^2}{a(s)^2} + \Big(\frac{\omega_n}{s}\Big)^{2/n} \frac{1}{b(s)^2} \sum_{i=2}^n x_i^2 \qquad \text{for } (x,s) \in \mathbb{R}^n \times (0,\infty).$$

Now we have

$$|\nabla \sigma(x)| = \left[\frac{\left[\left(\frac{\omega_n^{1/n}}{s^{1/n} a(s)} \right)^4 x_1^2 + \left(\frac{\omega_n^{1/n}}{s^{1/n} b(s)} \right)^4 \sum_{i=2}^n x_i^2 \right]^{1/2}}{\left(\frac{\omega_n^{1/n}}{s^{1/n} a(s)} \right)^2 x_1^2 \frac{\big(s^{1/n} a(s)\big)'}{s^{1/n} a(s)} + \left(\frac{\omega_n^{1/n}}{s^{1/n} b(s)} \right)^2 \frac{\big(s^{1/n} b(s)\big)'}{s^{1/n} b(s)} \sum_{i=2}^n x_i^2} \right]_{s=\sigma(x)} \tag{3.21}$$

if $s_0 \le \omega_n|x|^n \le s_0+\epsilon$, by the implicit function theorem, and

$$|\nabla \sigma(x)| = n \omega_n^{1/n} \sigma(x)^{1-1/n} \qquad \text{if} \quad \omega_n|x|^n \notin [s_0, s_0+\epsilon]. \tag{3.22}$$

From (3.21), via the monotonicity of $a(s)s^{1/n}$ and $b(s)s^{1/n}$, and due to (3.16), (3.17), and (3.20), one can easily deduce that

$$|\nabla \sigma(x)| \le n \omega_n^{1/n} \sigma(x)^{1-1/n} (1+\delta) \qquad \text{if } s_0 \le \omega_n|x|^n \le s_0+\epsilon. \tag{3.23}$$

Thus (3.19), (3.22), and (3.23) yield

$$|\nabla u(x)| \le h(x) \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \Omega, \tag{3.24}$$

where

$$h(x) = \begin{cases} f^*(\sigma(x))(1+\delta) & \text{if } s_0 \le \omega_n|x|^n \le s_0+\epsilon, \\ f^*(\sigma(x)) & \text{otherwise.} \end{cases}$$

By (3.7) and (3.13),

$$h(x) \le g^*(\sigma(x)) \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \Omega. \tag{3.25}$$

Combining (3.24) and (3.25), and exploiting the fact that σ is an m.p. map tell us that

$$|\nabla u|^*(s) \le g^*(s) \qquad \text{for } s > 0. \tag{3.26}$$

On the other hand,

$$f^{**}(s) = |\nabla u^\bigstar|^{**}(s) \le |\nabla u|^{**}(s) \le g^{**}(s) \qquad \text{for } s > 0,$$

where the equality is due to (2.1) and to (3.4) and (3.9), the first inequality to (1.10), and the last inequality to (3.26). Hence, owing to (3.8), equation (3.18) follows. □

4. Extremals of $**$-strictly increasing functionals. The proof of Theorem 1.6, which will be accomplished in this section, requires some preliminary results. The first one appears in [ALT, Theorem 2.3] and deals with weak limits in L^1 of equidistributed sequences of functions.

LEMMA 4.1. *Let Ω be a measurable subset of $\mathbb{R}^n$ having finite measure, and let $f \in L^1(\Omega)$. Let $\{f_k\}_{k\in\mathbb{N}}$ be any sequence of functions in $L^1(\Omega)$ such that*

$$f_k \rightharpoonup f \quad \text{weakly in } L^1(\Omega) \quad \text{and} \quad f_k^* \equiv f^* \qquad \text{for every } k \in \mathbb{N}.$$

Then

$$f_k \to f \quad \text{strongly in } L^1(\Omega).$$

The convergence in L^1 of the composition of a fixed function with a convergent sequence of m.p. maps is the object of the following lemma.

LEMMA 4.2. *Let Ω be a measurable subset of $\mathbb{R}^n$ having finite measure, and let σ and $\{\sigma_k\}_{k\in\mathbb{N}}$ be, respectively, an m.p. map and a sequence of m.p. maps from Ω into $[0, \mathcal{L}^n(\Omega)]$ such that*

$$\lim_{k\to+\infty} \sigma_k(x) = \sigma(x) \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \Omega.$$

Let ϕ be any function from $L^1(0, \mathcal{L}^n(\Omega))$. Then

$$\phi \circ \sigma_k \to \phi \circ \sigma \qquad \text{strongly in } L^1(\Omega). \tag{4.1}$$

Proof. When ϕ is continuous, $\lim_{k\to+\infty} \phi(\sigma_k(x)) = \phi(\sigma(x))$ for $\mathcal{L}^n$-a.e. $x \in \Omega$. Moreover, since the functions $\phi\circ\sigma_k$, $\phi\circ\sigma$, and ϕ are equimeasurable for every $k \in \mathbb{N}$, $\| \phi\circ\sigma_k \|_{L^1(\Omega)} = \| \phi\circ\sigma \|_{L^1(\Omega)} = \| \phi \|_{L^1(0,\mathcal{L}^n(\Omega))}$ for every $k \in \mathbb{N}$. Hence (4.1) follows (see, e.g., [AFP, Proposition 1.33]). In the general case, where $\phi \in L^1(0, \ln(\Omega))$, fix any $\epsilon > 0$, and choose a continuous function $\tilde\phi : [0, \mathcal{L}^n(\Omega)] \to \mathbb{R}$ such that $\| \phi - \tilde\phi \|_{L^1(0,\mathcal{L}^n(\Omega))} < \epsilon$. Hence,

$$\begin{aligned}\int_\Omega |\phi(\sigma_k(x)) - \phi(\sigma(x))|dx &\le \int_\Omega |\phi(\sigma_k(x)) - \tilde\phi(\sigma_k(x))|dx \\ &+ \int_\Omega |\tilde\phi(\sigma_k(x)) - \tilde\phi(\sigma(x))|dx + \int_\Omega |\tilde\phi(\sigma(x)) - \phi(\sigma(x))|dx \\ &= 2 \| \phi - \tilde\phi \|_{L^1(0,\mathcal{L}^n(\Omega))} + \| \tilde\phi\circ\sigma_k - \tilde\phi\circ\sigma \|_{L^1(\Omega)} \le 2\epsilon + \| \tilde\phi\circ\sigma_k - \tilde\phi\circ\sigma \|_{L^1(\Omega)}.\end{aligned}$$

Notice that the equality holds by property (3.1). Owing to the arbitrariness of ϵ, (4.1) for ϕ is now a consequence of (4.1) applied to $\tilde\phi$. □

The next lemma provides us with a weak approximation $|\nabla u|_{*u}$ involving a sequence of m.p. maps related to the level sets of u.

LEMMA 4.3. *Let $u \in W^1_+(\mathbb{R}^n)$. Let σ and $\{\sigma_k\}_{k\in\mathbb{N}}$ be an m.p. map and a sequence of bijective m.p. maps, respectively, from $\{u > 0\}$ into $[0, \mathcal{L}^n(\{u > 0\})]$, such that*

$$u(x) = u^*(\sigma(x)) \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \{u > 0\} \tag{4.2}$$

and

$$\lim_{k\to+\infty} \sigma_k(x) = \sigma(x) \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \{u > 0\}. \tag{4.3}$$

Then

$$|\nabla u| \circ \sigma_k^{-1} \rightharpoonup |\nabla u|_{*u} \qquad \text{weakly in } L^1(0, \mathcal{L}^n(\{u > 0\})).$$

Remark 4.4. Let us notice that σ and $\{\sigma_k\}$, as in the statement of Lemma 5.1, certainly exist owing to [FV1].

Proof. Set $V = \mathcal{L}^n(\{u > 0\})$. We begin by showing that

$$\lim_{k\to+\infty} \int_0^V |\nabla u|(\sigma_k^{-1}(t))\varphi(t)dt = \int_0^V |\nabla u|_{*u}(t)\varphi(t)dt \tag{4.4}$$

for every $\varphi \in C^0(0, V)$. In order to prove (4.4), observe first that

$$\begin{aligned} \lim_{k\to+\infty} \int_0^V |\nabla u|(\sigma_k^{-1}(t))\varphi(t)dt &= \lim_{k\to+\infty} \int_{\{u>0\}} |\nabla u|(x)\varphi(\sigma_k(x))dx \\ &= \int_{\{u>0\}} |\nabla u|(x)\varphi(\sigma(x))dx, \end{aligned} \tag{4.5}$$

where the former equality is due to (3.2) and the latter follows from (4.3) via the dominated convergence theorem for integrals. Now, let $\{t_i\}_{i\in P}$ and $\{I_i\}_{i\in P}$ be the families of levels and intervals, respectively, introduced in the proof of Theorem 1.2. It is easily seen that

$$u^{*-1}(\{t\}) = \mu(t) \qquad \text{if } t \neq t_i \text{ for every } i \in P \tag{4.6}$$

and (see, e.g., [BZ, proof of Lemma 2.4])

$$\mu(u^*(s)) = s \qquad \text{if } s \notin \cup_{i\in P} I_i. \tag{4.7}$$

By the coarea formula (2.9),

$$\int_{\{u>0\}} |\nabla u|(x)\varphi(\sigma(x))dx = \int_0^\infty dt \int_{\{u=t\}} \varphi(\sigma(x))d\mathcal{H}^{n-1}(x). \tag{4.8}$$

Furthermore, by (4.2) and (4.6),

$$\{u = t\} = \{\sigma = \mu(t)\} \qquad \text{if } t \neq t_i \text{ for every } i \in P. \tag{4.9}$$

Consequently, (4.8) reads

$$\int_{\{u>0\}} |\nabla u|(x)\varphi(\sigma(x))dx = \int_0^\infty \varphi(\mu(t))\mathcal{H}^{n-1}(\{u = t\})dt. \tag{4.10}$$

The change of variable $t = u^*(s)$ in the last integral and (4.7) entail that

$$\int_0^\infty \varphi(\mu(t))\mathcal{H}^{n-1}(\{u = t\})dt = \int_0^V \varphi(s)\mathcal{H}^{n-1}(\{u = u^*(s)\}) \left(-\frac{du^*}{ds}(s)\right) ds. \tag{4.11}$$

Thanks to the coarea formula (2.9) again,

$$|\nabla u|_{*u}(s) = \left(-\frac{du^*}{ds}(s)\right) \mathcal{H}^{n-1}(\{u = u^*(s)\}) \qquad \text{for } \mathcal{L}^1\text{-a.e. } s > 0. \tag{4.12}$$

Combining (4.5), (4.8), (4.10), (4.11), and (4.12) yields (4.4).

In order to conclude, we need to extend (4.4) to every $\varphi \in L^\infty(0,V)$. Fixing any such φ and any $\epsilon > 0$, by Lusin's theorem there exists $\tilde{\varphi} \in C^0(0,V)$ such that

$$\mathcal{L}^1(\{t : \varphi(t) \neq \tilde{\varphi}(t)\}) < \epsilon \tag{4.13}$$

and

$$\|\tilde{\varphi}\|_{L^\infty(0,V)} \leq \|\varphi\|_{L^\infty(0,V)}. \tag{4.14}$$

By (4.14),

(4.15)
$$\begin{aligned}
&\left| \int_0^V \big(|\nabla u|(\sigma_k^{-1}(t)) - |\nabla u|_{*u}(t)\big)\varphi(t)dt \right| \leq \left| \int_0^V \big(|\nabla u|(\sigma_k^{-1}(t)) - |\nabla u|_{*u}(t)\big)\tilde{\varphi}(t)dt \right| \\
&\qquad + \left| \int_{\{\varphi \neq \tilde{\varphi}\}} \big(|\nabla u|(\sigma_k^{-1}(t)) - |\nabla u|_{*u}(t)\big)\big(\varphi(t) - \tilde{\varphi}(t)\big)dt \right| \\
&\leq \left| \int_0^V \big(|\nabla u|(\sigma_k^{-1}(t)) - |\nabla u|_{*u}(t)\big)\tilde{\varphi}(t)dt \right| \\
&\qquad + 2\|\varphi\|_{L^\infty(0,V)} \int_{\{\varphi \neq \tilde{\varphi}\}} \big(|\nabla u|(\sigma_k^{-1}(t)) + |\nabla u|_{*u}(t)\big)\, dt.
\end{aligned}$$

Equations (3.2) and (2.10) yield

$$\int_{\{\varphi \neq \tilde{\varphi}\}} |\nabla u|_{*u}(t)dt \leq \int_0^{\mathcal{L}^1(\{\varphi \neq \tilde{\varphi}\})} (|\nabla u|_{*u})^*(s)ds \leq \int_0^{\mathcal{L}^1(\{\varphi \neq \tilde{\varphi}\})} |\nabla u|^*(s)ds. \tag{4.16}$$

Moreover, since σ_k is a bijective m.p. map, $(|\nabla u| \circ \sigma_k^{-1})^* \equiv |\nabla u|^*$ for every $k \in \mathbb{N}$. Hence, by (3.2) and (4.13),

$$\int_{\{\varphi \neq \tilde{\varphi}\}} |\nabla u|(\sigma_k^{-1}(t))dt \leq \int_0^{\mathcal{L}^1(\{\varphi \neq \tilde{\varphi}\})} \big(|\nabla u| \circ \sigma_k^{-1}\big)^*(s)ds \leq \int_0^{\epsilon} |\nabla u|^*(s)ds. \tag{4.17}$$

From (4.15), (4.16), (4.17), and (4.13) one infers that

$$\begin{aligned}
&\left| \int_0^V \big(|\nabla u|(\sigma_k^{-1}(t)) - |\nabla u|_{*u}(t)\big)\varphi(t)dt \right| \\
&\qquad \leq \left| \int_0^V \big(|\nabla u|(\sigma_k^{-1}(t)) - |\nabla u|_{*u}(t)\big)\tilde{\varphi}(t)dt \right| + 4\|\varphi\|_{L^\infty(0,V)} \int_0^{\epsilon} |\nabla u|^*(s)ds.
\end{aligned}$$

Owing to the arbitrariness of ϵ, equation (4.4) follows from the same equation applied with φ replaced by $\tilde{\varphi}$. □

A technical, but crucial, result in preparation for Theorem 1.6 is the following perturbation lemma.

LEMMA 4.5. *Let f and g be any functions in $L^1_+(\mathbb{R}^n)$ satisfying*

$$f \prec g \tag{4.18}$$

and

$$f^* \not\equiv g^*. \tag{4.19}$$

Then there exist numbers $\bar{s} > 0$, $\epsilon > 0$, *and* $\delta \in (0, \frac{1}{2})$ *having the following property. Fix any point* $x_0 \in \mathbb{R}^n$ *such that* $|x_0| < \epsilon$, *and set*

$$\Omega_{\bar{s},\epsilon} = \{x : \omega_n |x - x_0|^n \geq \bar{s} \ \textit{and} \ \omega_n |x|^n \leq \bar{s} + \epsilon\}.$$

If

$$\sigma : \Omega_{\bar{s},\epsilon} \to [\bar{s}, \bar{s} + \epsilon]$$

is an m.p. map and h *is a function from* $L^1_+(\mathbb{R}^n)$ *fulfilling*

$$\int_{\mathbb{R}^n} h(x)dx = \int_{\mathbb{R}^n} f(x)dx, \tag{4.20}$$

$$h(x) = \begin{cases} f^\bigstar(x - x_0) & \textit{if } \omega_n |x - x_0|^n < \bar{s}, \\ f^\bigstar(x) & \textit{if } \omega_n |x|^n > \bar{s} + \epsilon, \end{cases} \tag{4.21}$$

and

$$f^*(\sigma(x))(1 - \delta) \leq h(x) \leq f^*(\sigma(x))(1 + \delta) \qquad \textit{for } \mathcal{L}^n\textit{-a.e.} \quad x \in \Omega_{\bar{s},\epsilon}, \tag{4.22}$$

then

$$h \prec g. \tag{4.23}$$

Proof. By (4.18) and (4.19), there exists $\hat{s} > 0$ such that $\int_0^{\hat{s}} f^*(r)dr < \int_0^{\hat{s}} g^*(r)dr$. Let $[\hat{a}, \hat{b}]$ be the maximal interval in $[0, +\infty)$ containing $\hat{s}$ and having the property that

$$\int_0^s f^*(r)dr < \int_0^s g^*(r)dr \qquad \text{for } s \in (\hat{a}, \hat{b}).$$

Then,

$$\int_0^{\hat{a}} f^*(r)dr = \int_0^{\hat{a}} g^*(r)dr, \qquad \int_0^{\hat{b}} f^*(r)dr = \int_0^{\hat{b}} g^*(r)dr,$$

and

$$f^*(\hat{a}) \leq g^*(\hat{a}), \qquad f^*(\hat{b}^-) \geq g^*(\hat{b}^-).$$

Here, $f^*(\hat{b}^-)$ and $g^*(\hat{b}^-)$ stand for the limits of f^* and g^* at $\hat{b}$ from the left. It is not difficult to see that a subinterval $[a, b]$ of $[\hat{a}, \hat{b}]$ can be found such that

$$\int_0^s f^*(r)dr < \int_0^s g^*(r)dr \qquad \text{for } s \in [a, b],$$

and

$$f^*(a) < g^*(a), \qquad f^*(b) > g^*(b).$$

Fix any $\eta > 0$ fulfilling

$$f^*(a)(1+\eta) < g^*(a), \qquad f^*(b)(1-\eta) > g^*(b). \tag{4.24}$$

We shall prove that the conclusion holds with

$$\bar{s} = b \tag{4.25}$$

$$\delta = \min\left\{\frac{\eta}{2}, \frac{1}{4}, \frac{1}{f^*(b)} \min_{s\in[a,b]} \int_0^s (g^*(r) - f^*(r))\, dr\right\} \tag{4.26}$$

and any $\epsilon \in (0,1)$ satisfying

$$\frac{f^*(b+\epsilon)}{f^*(b)} > \frac{(1-2\delta)}{(1-\delta)}. \tag{4.27}$$

Notice that, since $\delta < \frac{1}{2}$, such an ϵ certainly exists, due to the continuity of f^* from the right. Let h be any function as in the statement. By (4.18) and (4.20),

$$\int_{\mathbb{R}^n} h(x)dx = \int_{\mathbb{R}^n} g(x)dx.$$

Thus, by (2.2), equation (4.23) will follow if we show that

$$\int_E h(x)dx \le \int_0^{\mathcal{L}^n(E)} g^*(s)ds \tag{4.28}$$

for every measurable subset E of $\mathbb{R}^n$. Given any such E, define

$$E_1 = E \cap G_1, \qquad E_2 = E \cap G_2, \qquad E_3 = E \cap G_3,$$

where

$$G_1 = \{x \in \mathbb{R}^n : \omega_n |x - x_0|^n < b\}, \qquad G_2 = \Omega_{b,\epsilon}, \qquad G_3 = \{x \in \mathbb{R}^n : \omega_n |x|^n > b + \epsilon\},$$

and decompose $\int_E h(x)dx$ as

$$\int_E h(x)dx = \int_{E_1} h(x)dx + \int_{E_2} h(x)dx + \int_{E_3} h(x)dx. \tag{4.29}$$

The first and the third integral on the right-hand of (4.29) can be easily estimated in terms of f^* via (4.21). Indeed,

$$\int_{E_1} h(x)dx = \int_{E_1} f^\bigstar(x - x_0)dx \le \int_0^{\mathcal{L}^n(E_1)} f^*(s)ds \tag{4.30}$$

and

$$\int_{E_3} h(x)dx = \int_{E_3} f^\bigstar(x)dx \le \int_{b+\epsilon}^{b+\epsilon+\mathcal{L}^n(E_3)} f^*(s)ds. \tag{4.31}$$

We now accomplish the proof of (4.28) on distinguishing the following three cases.

Case I. $\mathcal{L}^n(E) \le a$. Since σ is an m.p. map and $\sigma(x) \in [b, b+\epsilon]$ for $\mathcal{L}^n$-a.e. $x \in G_2$, by (4.22)

$$\int_{E_2} h(x)dx \le (1+\delta)\int_{E_2} f^*(\sigma(x))dx \le (1+\delta)\int_b^{b+\mathcal{L}^n(E_2)} f^*(s)ds. \tag{4.32}$$

Furthermore,

$$\begin{aligned}(1+\delta)\int_b^{b+\mathcal{L}^n(E_2)} f^*(s)ds &\le (1+\delta)f^*(a)\mathcal{L}^n(E_2) \le g^*(a)\mathcal{L}^n(E_2)\\ &\le \int_{a-\mathcal{L}^n(E_2)}^{a} g^*(s)ds,\end{aligned} \tag{4.33}$$

where the first and the third inequalities lean on the monotonicity of f^* and g^*, respectively, and the second inequality is a consequence of (4.24) and (4.26). The monotonicity of g^* again and the assumption $\mathcal{L}^n(E) \le a$ entail that

$$\int_{a-\mathcal{L}^n(E_2)}^{a} g^*(s)ds \le \int_{\mathcal{L}^n(E)-\mathcal{L}^n(E_2)}^{\mathcal{L}^n(E)} g^*(s)ds = \int_{\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)}^{\mathcal{L}^n(E)} g^*(s)ds. \tag{4.34}$$

Combining inequalities (4.29)–(4.34) yields

$$\int_E h(x)dx \le \int_0^{\mathcal{L}^n(E_1)} f^*(s)ds + \int_{b+\epsilon}^{b+\epsilon+\mathcal{L}^n(E_3)} f^*(s)ds + \int_{\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)}^{\mathcal{L}^n(E)} g^*(s)ds. \tag{4.35}$$

On the other hand,

$$\begin{aligned}\int_0^{\mathcal{L}^n(E_1)} f^*(s)ds + \int_{b+\epsilon}^{b+\epsilon+\mathcal{L}^n(E_3)} f^*(s)ds &\le \int_0^{\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)} f^*(s)ds\\ &\le \int_0^{\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)} g^*(s)ds,\end{aligned} \tag{4.36}$$

where the last inequality holds because of (4.18). Inequality (4.28) follows from (4.35) and (4.36).

Case II. $a < \mathcal{L}^n(E) \le b$. We have

$$\begin{aligned}&\int_{E_2} h(x)dx \le (1+\delta)\int_b^{\mathcal{L}^n(E_2)+b} f^*(s)ds \le \int_b^{b+\mathcal{L}^n(E_2)} f^*(s)ds + \delta\mathcal{L}^n(E_2)f^*(b)\\ &\quad\le \int_b^{b+\mathcal{L}^n(E_2)} f^*(s)ds + \delta f^*(b) \le \int_b^{b+\mathcal{L}^n(E_2)} f^*(s)ds + \min_{s\in[a,b]}\int_0^s (g^*(r)-f^*(r))\,dr,\end{aligned} \tag{4.37}$$

where the first inequality is due to (4.32), the second one holds thanks to the monotonicity of f^*, the third one is a consequence of the fact that $\mathcal{L}^n(E_2) \le \epsilon < 1$, and the last one follows from (4.26). Since we are assuming that $a < \mathcal{L}^n(E) \le b$, combining (4.29), (4.30), (4.31), and (4.37) ensures that

$$\begin{aligned}\int_E h(x)dx \le &\int_0^{\mathcal{L}^n(E_1)} f^*(s)ds + \int_{b+\epsilon}^{b+\epsilon+\mathcal{L}^n(E_3)} f^*(s)ds\\ &+ \int_b^{b+\mathcal{L}^n(E_3)} f^*(s)ds + \int_0^{\mathcal{L}^n(E)} (g^*(s)-f^*(s))\,ds.\end{aligned} \tag{4.38}$$

Since the first three integrals on the right-hand side of (4.38) are extended over disjoint intervals whose total measure is $\mathcal{L}^n(E)$, their sum does not exceed $\int_0^{\mathcal{L}^n(E)} f^*(s)ds$. Hence, (4.28) follows.

Case III. $\mathcal{L}^n(E) > b$. Assumptions (4.20) and (4.21) and the choice (4.25) ensure that

$$\begin{aligned}\int_{G_2} h(x)dx &= \int_{\mathbb{R}^n} f(x)dx - \int_{G_1} f^\bigstar(x-x_0)dx - \int_{G_3} f^\bigstar(x)dx \\ &= \int_0^{+\infty} f^*(s)ds - \int_0^b f^*(s)ds - \int_{b+\epsilon}^{+\infty} f^*(s)ds = \int_b^{b+\epsilon} f^*(s)ds. \end{aligned} \tag{4.39}$$

Now,

$$\begin{aligned}\int_{G_2\setminus E_2} h(x)dx &\geq (1-\delta)\int_{G_2\setminus E_2} f^\bigstar(\sigma(x))dx \\ &\geq (1-\delta)\int_{b+\epsilon-\mathcal{L}^n(G_2\setminus E_2)}^{b+\epsilon} f^*(s)ds = (1-\delta)\int_{b+\mathcal{L}^n(E_2)}^{b+\epsilon} f^*(s)ds,\end{aligned} \tag{4.40}$$

where the first inequality is a consequence of (4.22), and the second one holds since $\sigma(x) \in [b, b+\epsilon]$ for $\mathcal{L}^n$-a.e. $x \in G_2$. By (4.27), (4.24), and (4.26),

$$\begin{aligned}(1-\delta)\int_{b+\mathcal{L}^n(E_2)}^{b+\epsilon} f^*(s)ds &\geq (1-\delta) f^*(b+\epsilon)\,(\epsilon - \mathcal{L}^n(E_2)) \\ &\geq g^*(b)\,(\epsilon - \mathcal{L}^n(E_2)) \quad \geq \quad \int_b^{b+\epsilon+\mathcal{L}^n(E_2)} g^*(s)ds.\end{aligned} \tag{4.41}$$

From (4.39)–(4.41), and from the assumption $\mathcal{L}^n(E) > b$, one gets

$$\begin{aligned}\int_{E_2} h(x)dx &\leq \int_b^{b+\epsilon} f^*(s)ds - \int_b^{b+\epsilon+\mathcal{L}^n(E_2)} g^*(s)ds \\ &\leq \int_b^{b+\epsilon} f^*(s)ds - \int_{\mathcal{L}^n(E)}^{\epsilon+\mathcal{L}^n(E)-\mathcal{L}^n(E_2)} g^*(s)ds \\ &= \int_b^{b+\epsilon} f^*(s)ds - \int_{\mathcal{L}^n(E)}^{\epsilon+\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)} g^*(s)ds.\end{aligned} \tag{4.42}$$

Combining (4.29)–(4.31) and (4.42) yields

$$\int_E h(x)dx \leq \int_0^{\mathcal{L}^n(E_1)} f^*(s)ds + \int_b^{b+\epsilon+\mathcal{L}^n(E_3)} f^*(s)ds - \int_{\mathcal{L}^n(E)}^{\epsilon+\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)} g^*(s)ds. \tag{4.43}$$

Since $\mathcal{L}^n(E_1) \leq b$ and f^* is nonincreasing,

$$\int_b^{b+\epsilon+\mathcal{L}^n(E_3)} f^*(s)ds \leq \int_{\mathcal{L}^n(E_1)}^{\epsilon+\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)} f^*(s)ds. \tag{4.44}$$

Inequalities (4.43) and (4.44) and assumption (4.18) enable us to conclude that

$$\int_E h(x)dx \leq \int_0^{\epsilon+\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)} g^*(s)ds - \int_{\mathcal{L}^n(E)}^{\epsilon+\mathcal{L}^n(E_1)+\mathcal{L}^n(E_3)} g^*(s)ds = \int_0^{\mathcal{L}^n(E)} g^*(s)ds.$$

Hence, (4.28) holds also in this case. □

We are now in position to prove Theorem 1.6.

Proof of Theorem 1.6. Assume that Φ is $**$-strictly increasing. By Proposition 2.3, Φ is, in particular, $\prec$-strictly increasing. If u is a minimal rearrangement, i.e., satisfies (1.11), then, thanks to the second of inequalities (2.15),

$$\Phi\left(|\nabla u|_{\bigstar u}\right) = \Phi\left(|\nabla u|\right) < +\infty. \tag{4.45}$$

From (4.45), (2.10), and the $\prec$-strict monotonicity of Φ, one deduces that

$$\left(|\nabla u|_{\bigstar u}\right)^* \equiv |\nabla u|^*. \tag{4.46}$$

Let $\{\sigma_k\}_{k\in\mathbb{N}}$ and σ be as in Lemma 4.3. Since σ_k is an m.p. map for every $k \in \mathbb{N}$, $\left(|\nabla u| \circ \sigma_k^{-1}\right)^* \equiv |\nabla u|^*$ for every $k \in \mathbb{N}$. Hence, inasmuch as $\left(|\nabla u|_{\bigstar u}\right)^* \equiv \left(|\nabla u|_{*u}\right)^*$, equation (4.46) entails that

$$\left(|\nabla u| \circ \sigma_k^{-1}\right)^* \equiv \left(|\nabla u|_{*u}\right)^* \qquad \text{for every } k \in \mathbb{N}. \tag{4.47}$$

From Lemma 4.3, (4.47), and Lemma 4.1, we infer that

$$\lim_{k\to+\infty} \| |\nabla u| \circ \sigma_k^{-1} - |\nabla u|_{*u}\|_{L^1(0,\mathcal{L}^n(\{u>0\}))} = 0,$$

whence, by (3.2),

$$\lim_{k\to+\infty} \| |\nabla u| - |\nabla u|_{*u} \circ \sigma_k\|_{L^1(\{u>0\})} = 0. \tag{4.48}$$

On the other hand, by Lemma 4.2,

$$\lim_{k\to+\infty} \| |\nabla u| - |\nabla u|_{*u} \circ \sigma_k\|_{L^1(\{u>0\})} = \| |\nabla u| - |\nabla u|_{*u} \circ \sigma\|_{L^1(\{u>0\})}. \tag{4.49}$$

Equations (4.48) and (4.49) tell us that

$$|\nabla u(x)| = |\nabla u|_{*u}(\sigma(x)) \qquad \text{for } \mathcal{L}^n\text{-a.e. } x \in \{u > 0\}. \tag{4.50}$$

Denote by N the set of those points in $\{u > 0\}$ where (4.50) does not hold. Thus $\mathcal{L}^n(N) = 0$. Since both $|\nabla u|$ and $|\nabla u|_{*u} \circ \sigma$ are Borel functions, N is a Borel set. Hence, an application of the coarea formula (2.9) with $f = \chi_N$ entails that

$$\mathcal{H}^{n-1}\left(\{u = t\} \cap N\right) = 0 \qquad \text{for } \mathcal{L}^1\text{-a.e. } t > 0. \tag{4.51}$$

Now, let $\{t_i\}_{i\in P}$ be the family of levels introduced in the proof of Theorem 1.2. Then, by (4.6), we again deduce (4.9). Equations (4.50), (4.51), and (4.9) ensure that, for $\mathcal{L}^1$-a.e. $t > 0$,

$$|\nabla u(x)| = |\nabla u|_{*u}(\mu(t)) \qquad \text{for } \mathcal{H}^{n-1}\text{-a.e. } x \in \{u = t\}.$$

Hence, (1.13) follows.

Conversely, suppose that any minimal rearrangement of Φ necessarily fulfills (1.13). We shall prove that Φ is $**$-strictly increasing. By Proposition 2.3, this is equivalent to showing that Φ is simultaneously $*$-strictly increasing and $\prec$-strictly increasing.

The functional Φ has to be $*$-strictly increasing, since otherwise the same function u constructed in the proof of Theorem 1.4 would satisfy (1.11) but not (1.13).

In order to show that Φ is also $\prec$-strictly increasing, we argue by contradiction, and suppose that $f, g \in L^1_+(\mathbb{R}^n)$ exist such that $f \prec g$ and

$$\Phi(f) = \Phi(g) < +\infty, \tag{4.52}$$

but $f^* \not\equiv g^*$. Our aim is to exhibit a function $u \in W^1_+(\mathbb{R}^n)$ satisfying

$$|\nabla u^\bigstar| \equiv f^\bigstar \tag{4.53}$$

and

$$|\nabla u| \prec g \tag{4.54}$$

but not (1.13). This will lead to a contradiction, since, by (2.15), (4.52)–(4.54), and the $**$-monotonicity of Φ, such a function u is a minimal rearrangement of Φ. As in the proof of Theorem 1.4, we look for a function u in the form (3.3), where

$$\phi(s) = \int_s^{\mathcal{L}^n(\Omega)} \frac{f^*(r)}{n\omega_n^{1/n} r^{1-1/n}} dr \qquad \text{for } s \in [0, \mathcal{L}^n(\Omega)] \tag{4.55}$$

and $\Omega = \{x : \omega_n |x|^n \le \mathcal{L}^n(\{g > 0\})\}$. Owing to (3.4) and (4.55), any such function satisfies (4.53). Thus, our task is to construct σ in such a way that (4.54) holds. The idea is again to obtain u as a perturbation of $u^\bigstar$, namely, to define $\sigma(x)$ by a slight modification of $\omega_n |x|^n$. It is at this stage that Lemma 4.5 comes into play. Actually, the functions f and g satisfy the assumptions of this lemma. Let $\bar{s}$, ϵ, and δ be the numbers appearing in its statement. Consider any $\lambda > 0$ and any sufficiently smooth function

$$p : [\bar{s}, \bar{s} + \epsilon] \to [0, \lambda]$$

enjoying the following properties: p is strictly decreasing,

$$p(\bar{s}) = \lambda, \quad p(\bar{s} + \epsilon) = 0, \tag{4.56}$$

and

$$p'(s) \ge -\frac{\delta}{n(1+\delta)} \frac{s^{-1+1/n}}{\omega_n^{1/n}} \qquad \text{for } s \in [\bar{s}, \bar{s} + \epsilon]. \tag{4.57}$$

In particular, by (4.57), the function $s \mapsto p(s) + (\frac{s}{\omega_n})^{1/n}$ is strictly increasing in $[\bar{s}, \bar{s} + \epsilon]$. For instance, the choice

$$\lambda = \frac{\delta}{1+\delta} \frac{(\bar{s} + \epsilon)^{1/n} - \bar{s}^{1/n}}{\omega_n^{1/n}} \qquad \text{and} \qquad p(s) = \frac{\delta}{1+\delta} \frac{(\bar{s} + \epsilon)^{1/n} - s^{1/n}}{\omega_n^{1/n}}$$

is admissible. Now, define $x_0 \equiv (\lambda, \dots, 0) \in \mathbb{R}^n$ and $P(s) \equiv (p(s), 0, \dots, 0) \in \mathbb{R}^n$ for $s \in [\bar{s}, \bar{s} + \epsilon]$, and set

$$D(s) = \begin{cases} \{x \in \mathbb{R}^n : \omega_n |x - x_0|^n < s\} & \text{if } s < \bar{s}, \\ \{x \in \mathbb{R}^n : \omega_n |x - P(s)|^n < s\} & \text{if } \bar{s} \le s \le \bar{s} + \epsilon, \\ \{x \in \mathbb{R}^n : \omega_n |x|^n < s\} & \text{if } s > \bar{s} + \epsilon. \end{cases}$$

The m.p. map σ associated with the family of balls $\mathcal{D} = \{D(s)\}_{s\in[0,\mathcal{L}^n(\Omega)]}$ as in (3.12) obeys

$$\sigma(x) = \begin{cases} \omega_n|x-x_0|^n & \text{if } \omega_n|x-x_0|^n < \bar{s}, \\ \omega_n|x|^n & \text{if } \omega_n|x|^n > \bar{s}+\epsilon, \end{cases} \tag{4.58}$$

and it is otherwise implicitly defined by the equation

$$H(x,\sigma(x)) = 1,$$

where $H : \mathbb{R}^n \times (0,\infty) \to \mathbb{R}$ is given by

$$H(x,s) = \Big(\frac{\omega_n}{s}\Big)^{2/n} \left[(x_1 - p(s))^2 + \sum_{i=2}^n x_i^2\right] \qquad \text{for} \quad (x,s) \in \mathbb{R}^n \times (0,\infty).$$

Such a function σ is clearly Lipschitz continuous. Moreover, one can easily verify from (4.58) that

$$|\nabla\sigma(x)| = n\omega_n^{1/n}\sigma(x)^{1-1/n} \qquad \text{for} \quad \mathcal{L}^n\text{-a.e. } x \text{ such that } \ \sigma(x) \notin [\bar{s},\bar{s}+\epsilon], \tag{4.59}$$

and that, by the implicit function theorem,

$$|\nabla\sigma(x)| = \left[\left(\frac{1}{n}\frac{s^{2/n-1}}{\omega_n^{2/n}} + (x_1 - p(s))p'(s)\right)^{-1}\left(\frac{s}{\omega_n}\right)^{1/n}\right]_{s=\sigma(x)} \tag{4.60}$$

for $\mathcal{L}^n$-a.e. x such that $\sigma(x) \in [\bar{s},\bar{s}+\epsilon]$. We claim that

$$1-\delta \le \frac{|\nabla\sigma(x)|}{n\omega_n^{1/n}\sigma(x)^{1-1/n}} \le 1+\delta \tag{4.61}$$

for $\mathcal{L}^n$-a.e. x such that $\sigma(x) \in [\bar{s},\bar{s}+\epsilon]$. Indeed, since $|x_1 - p(s)| \le \big(\frac{s}{\omega_n}\big)^{1/n}$ if $s \in [\bar{s},\bar{s}+\epsilon]$, by (4.56) and (4.57) we have

$$\frac{1}{1+\delta}\,\frac{s^{2/n-1}}{n\omega_n^{2/n}} \le \frac{s^{2/n-1}}{n\omega_n^{2/n}} + (x_1 - p(s))p'(s) \le \frac{1+2\delta}{1+\delta}\,\frac{s^{2/n-1}}{n\omega_n^{2/n}}. \tag{4.62}$$

Inequality (4.61) follows via (4.60) and (4.62). From (3.19) and (4.61) we deduce that

$$(1-\delta)f^*(\sigma(x)) \le |\nabla u(x)| \le (1+\delta)f^*(\sigma(x)) \tag{4.63}$$

for $\mathcal{L}^n$-a.e. x such that $\sigma(x) \in [\bar{s},\bar{s}+\epsilon]$. On the other hand, by (4.59) and (3.19),

$$|\nabla u(x)| = \begin{cases} f^\bigstar(x-x_0) & \text{if } \omega_n|x-x_0|^n < \bar{s}, \\ f^\bigstar(x) & \text{if } \omega_n|x|^n > \bar{s}+\epsilon. \end{cases} \tag{4.64}$$

Finally, by the coarea formula (2.8),

$$\begin{aligned} \int_{\mathbb{R}^n} |\nabla u(x)|dx &= \int_{\mathbb{R}^n} \frac{f^*(\sigma(x))}{n\omega_n^{1/n}\sigma(x)^{1-1/n}}|\nabla\sigma(x)|dx \\ &= \int_0^\infty \frac{f^*(t)}{n\omega_n^{1/n}t^{1-1/n}}\mathcal{H}^{n-1}(\partial^*\{\sigma > t\})dt = \int_0^\infty f^*(t)dt = \int_{\mathbb{R}^n} f(x)dx. \end{aligned} \tag{4.65}$$

Notice that in the third equality we have made use of the fact that $\{\sigma > t\}$ is a ball of measure t for every $t > 0$. Equation (4.54) follows from (4.63)–(4.65), via Lemma 4.5 applied with $h = |\nabla u|$. The proof is complete. □

5. Examples. We conclude by discussing the monotonicity and strict monotonicity properties involved in Theorems 1.2, 1.4, and 1.6 for the functionals considered in section 1.

We preliminarily present a few general propositions linking $**$-monotonicity and strict monotonicity to customary notions appearing in the literature. The first one provides sufficient conditions for a functional Φ to be $**$-increasing in terms of convexity and lower-semicontinuity.

PROPOSITION 5.1. *Assume that the functional $\Phi : L^1_+(\mathbb{R}^n) \to [0,\infty]$ is convex, weakly lower-semicontinuous in L^1, and $*$-increasing. Then Φ is $**$-increasing.*

Proof. By [CR, Chapter VI, Lemma 20.2], our assumptions on Φ ensure that a family $\{h_j\}_{j\in J}$, with $J \subseteq \mathbb{N}$, of nonnegative functions $h_j \in L^\infty(\mathbb{R}^n)$ and a family $\{r_j\}_{j\in J}$ of real numbers r_j exist such that

$$\Phi(f) = \sup_{j\in J}\left\{\int_0^\infty f^*(s)h_j^*(s)ds + r_j\right\} \qquad \text{for every } f \in L^1_+(\mathbb{R}^n). \tag{5.1}$$

On the other hand, Hardy's lemma [BS, Chapter 2, Proposition 3.6] entails that if $f^{**}(s) \le g^{**}(s)$ for $s > 0$, then

$$\int_0^\infty f^*(s)\phi(s)ds \le \int_0^\infty g^*(s)\phi(s)ds \tag{5.2}$$

for every nonincreasing function $\phi : [0,\infty) \to [0,\infty)$. Combining (5.1) and (5.2) tells us that Φ is $**$-increasing. □

The next result can be found, e.g., in [BS, Chapter 2, Corollary 4.7] (see also [CR]) and concerns the $**$-monotonicity of rearrangement invariant (r.i.) norms, namely norms in Banach function spaces which are r.i. according to Definition 3.2.

PROPOSITION 5.2. *Any r.i. norm is a $**$-increasing functional.*

The last general property that will be established yields a sufficient condition for the $**$-strict monotonicity of a functional Φ in terms of standard strict convexity and, more generally, of rotundity. Recall that a functional $\Phi : L^1_+(\mathbb{R}^n) \to [0,+\infty]$ is called rotund if

$$\Phi(f) = \Phi(g) = \Phi\left(\frac{f+g}{2}\right) < +\infty \quad \text{implies that} \quad f = g \quad \mathcal{L}^n\text{-a.e. in } \mathbb{R}^n.$$

Clearly, any strictly convex functional Φ is rotund.

PROPOSITION 5.3. *Any rotund and $**$-increasing functional $\Phi : L^1_+(\mathbb{R}^n) \to [0,+\infty]$ is $**$-strictly increasing. In particular, Φ is $**$-strictly increasing, provided that it is $**$-increasing and strictly convex.*

Proof. Let f and g be functions in $L^1_+(\mathbb{R}^n)$ such that

$$f^{**}(s) \le g^{**}(s) \qquad \text{for } s > 0 \tag{5.3}$$

and

$$\Phi(f) = \Phi(g) < +\infty. \tag{5.4}$$

By (5.3),

$$f^{**}(s) \le \frac{f^{**}(s) + g^{**}(s)}{2} \le g^{**}(s) \qquad \text{for } s > 0. \tag{5.5}$$

Moreover,

$$\frac{f^{**}(s)+g^{**}(s)}{2}=\left[\frac{f^{\bigstar}+g^{\bigstar}}{2}\right]^{**}(s)\qquad\text{for } s>0. \tag{5.6}$$

Since Φ is $*$-increasing, and hence rearrangement invariant, equations (5.4)–(5.6) imply that

$$\Phi(f^{\bigstar})=\Phi(f)=\Phi\left(\frac{f^{\bigstar}+g^{\bigstar}}{2}\right)=\Phi(g)=\Phi(g^{\bigstar}).$$

The rotundity of Φ then forces $f^{\bigstar}$ to agree with $g^{\bigstar}$, whence $f^*\equiv g^*$. □

We now examine the concrete examples provided by $\|\ \|_{(p,q)}$, $\|\ \|_{p,q}$, J_A, $\|\ \|_A$, and $\|\ \|_{(A)}$. In order to avoid technical complications, we are not going to consider the most general possible choices of the exponents p, q and of the Young function A. However, the cases discussed below cover most situations arising in applications and interestingly show that even equivalent functionals may enjoy different monotonicity properties.

The following two propositions deal with $\|\ \|_{(p,q)}$ and $\|\ \|_{p,q}$, respectively, for $p, q \in [1,\infty]$.

PROPOSITION 5.4. *Let $1\le p,q\le+\infty$. Then*

(i) $\|\ \|_{(p,q)}$ *is $**$-increasing;*

(ii) $\|\ \|_{(p,q)}$ *is $*$-strictly increasing if and only if $q<+\infty$;*

(iii) $\|\ \|_{(p,q)}$ *is $**$-strictly increasing if and only if $q<+\infty$.*

Proof. The facts that $\|\ \|_{(p,q)}$ is $**$-increasing and that, if $q<+\infty$, it is also $**$-strictly increasing (and hence $*$-strictly increasing) are obvious consequences of the definitions.

To verify that $\|\ \|_{(p,q)}$ is not $*$-strictly increasing (and hence also not $**$-strictly increasing) when $q=+\infty$ and $1\le p\le+\infty$, choose any functions $f, g\in L^1_+(\mathbb{R}^n)$ such that

$$f^*(s)=\begin{cases}s^{-1/p} & \text{if } s\in(0,1),\\ 0 & \text{otherwise,}\end{cases}\qquad g^*(s)=\begin{cases}s^{-1/p} & \text{if } s\in(0,1),\\ 1 & \text{if } s\in[1,2),\\ 0 & \text{otherwise.}\end{cases}$$

(Here, we agree that $s^{-1/p}\equiv 1$ if $p=+\infty$.) Obviously, $f^*(s)\le g^*(s)$ for $s>0$, and $f^*\not\equiv g^*$, but nevertheless, $\|f\|_{L^{p,\infty}}=\|g\|_{L^{p,\infty}}$. □

PROPOSITION 5.5. *Let $1\le p,q\le+\infty$. Then*

(i) $\|\ \|_{p,q}$ *is $**$-increasing if and only if $q\le p$;*

(ii) $\|\ \|_{p,q}$ *is $*$-strictly increasing if and only if $q\le p$ and $q<+\infty$;*

(iii) $\|\ \|_{p,q}$ *is $**$-strictly increasing if and only if $q\le p$, $q<+\infty$, and $p>1$.*

Proof. (i) If $q\le p$, then the function $s^{\frac{q}{p}-1}$ is nonincreasing. Thus, by Hardy's lemma [BS, Chapter 2, Proposition 3.6],

$$\int_0^\infty f^*(s)^q s^{\frac{q}{p}-1}ds\le\int_0^\infty g^*(s)^q s^{\frac{q}{p}-1}ds$$

whenever $f^{**}(s)\le g^{**}(s)$. Hence, $\|\ \|_{p,q}$ is $**$-increasing. Assume now that $q>p$. Fix $\alpha\in(0,1)$, and choose any $f, g\in L^1_+(\mathbb{R}^n)$ such that

$$g^*(s)=\begin{cases}s^{-\alpha} & \text{if } s\in(0,1),\\ 0 & \text{otherwise,}\end{cases}\qquad\text{and}\qquad f^*(s)=\begin{cases}\frac{1}{1-\alpha} & \text{if } s\in(0,1),\\ 0 & \text{otherwise.}\end{cases}$$

Then $f \prec g$, but $\| f \|_{p,q} > \| g \|_{p,q}$ if α is sufficiently small. This shows that $\| \quad \|_{p,q}$ is not $**$-increasing.

(ii) When $q \le p$ and $q < +\infty$, the functional $\| \quad \|_{p,q}$ is obviously $*$-strictly increasing. If $p = q = +\infty$, then $\| \quad \|_{\infty,\infty} = \| \quad \|_{(\infty,\infty)} = \| \quad \|_{L^\infty}$, and Proposition 5.4 tells us that it is not $*$-strictly increasing.

(iii) Assume that $q \le p$, $q < \infty$, $p > 1$. Let $f, g \in L^1_+(\mathbb{R}^n) \cap L^{p,q}(\mathbb{R}^n)$. Suppose that

$$f^{**}(s) \le g^{**}(s) \tag{5.7}$$

and

$$\| f \|_{p,q} = \| g \|_{p,q}. \tag{5.8}$$

Inequality (5.7) and [ALT, Proposition 2.1] entail that

$$\int_0^s f^*(t)^q dt \le \int_0^s g^*(t)^q dt \qquad \text{for } s > 0. \tag{5.9}$$

Integration by parts in the integral defining $\| \quad \|_{p,q}$ tells us that (5.8) is equivalent to

$$\int_0^\infty s^{\frac{q}{p}-2} \int_0^s f^*(t)^q dt ds = \int_0^\infty s^{\frac{q}{p}-2} \int_0^s g^*(t)^q dt ds. \tag{5.10}$$

From (5.9) and (5.10) we infer that equality holds in (5.9) for every $s > 0$, and hence that $f^*(s) = g^*(s)$ for $s > 0$. Thus, $\| \quad \|_{p,q}$ is $**$-strictly increasing.

If $p = q = +\infty$, then $\| \quad \|_{\infty,\infty}$ is not $**$-strictly increasing by (ii).

Finally, if $p = 1$, then also $q = 1$. Therefore, $\| \quad \|_{1,1} = \| \quad \|_{L^1}$, and the latter is not $**$-strictly increasing as demonstrated by any couple of functions $f, g \in L^1_+(\mathbb{R}^n)$ satisfying $\int_0^s f^*(t)dt < \int_0^s g^*(t)dt$ for $s \ge 0$ and $\int_0^\infty f^*(t)dt = \int_0^\infty g^*(t)dt$. □

The conclusions of Propositions 5.4–5.5 are summarized in the following table.

$1 \le p, q \le \infty$	$**$-INCREASING	$*$-STRICTLY INCREASING	$**$-STRICTLY INCREASING
$\| \quad \|_{(p,q)}$	any p, q	$q < \infty$	$q < \infty$
$\| \quad \|_{p,q}$	$q \le p$	$q \le p$ and $q < \infty$	$q \le p$, $p > 1$ and $q < \infty$

The functional J_A and the norms $\| \quad \|_A$ and $\| \quad \|_{(A)}$ are the object of the next propositions. Henceforth, we take into account finite-valued Young functions A, which are strictly positive in $(0, +\infty)$ and satisfy

$$\lim_{t \to 0} \frac{A(t)}{t} = 0 \quad \text{and} \quad \lim_{t \to +\infty} \frac{A(t)}{t} = +\infty.$$

According to usage, Young functions fulfilling these mild additional assumptions will be called N-functions.

PROPOSITION 5.6. *Let A be an N-function. Then*

(i) *J_A is $**$-increasing;*

(ii) *J_A is $*$-strictly increasing;*

(iii) *J_A is $**$-strictly increasing if and only if A is strictly convex.*

Proof. (i) The fact that J_A is $**$-increasing for every Young function A is proved, e.g., in [ALT, Corollary 2.1].

(ii) The $*$-strict monotonicity of J_A is a straightforward consequence of the equation

$$J_A(f) = \int_0^\infty A(f^*(s))ds \qquad \text{for every } f \in L^1_+(\mathbb{R}^n)$$

and of the strict monotonicity of A.

(iii) If A is strictly convex, then J_A is strictly convex, and its $**$-strict monotonicity follows from Proposition 5.3.

Conversely, assume that A is not strictly convex. Then there exists an interval $(t_0, t_1) \subset (0, \infty)$ and numbers $a > 0$ and $b \leq 0$ such that

$$A(t) = at + b \qquad \text{if } t_0 < t < t_1. \tag{5.11}$$

Fixing any $C \in (t_0, t_1)$ and any $\epsilon > 0$ such that $(C - \epsilon, C + \epsilon) \subset (t_0, t_1)$, choose any disjoint measurable subsets E_1 and E_2 of $\mathbb{R}^n$ satisfying $\mathcal{L}^n(E_1) = \mathcal{L}^n(E_2) < +\infty$. Define $f, g : \mathbb{R}^n \to [0, +\infty)$ as

$$f(x) = \begin{cases} C & \text{if } x \in E_1 \bigcup E_2, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad g(x) = \begin{cases} C + \epsilon & \text{if } x \in E_1, \\ C - \epsilon & \text{if } x \in E_2, \\ 0 & \text{otherwise.} \end{cases}$$

It is easily verified that $f \prec g$. Moreover,

$$J_A(f) = 2aC\mathcal{L}^n(E_1) + b = a(C + \epsilon)\mathcal{L}^n(E_1) + a(C - \epsilon)\mathcal{L}^n(E_2) + b = J_A(g).$$

Since $f^* \not\equiv g^*$, J_A is not $**$-strictly monotone. □

The notion of Δ_2-condition near infinity plays a role in the characterization of the monotonicity properties of the Luxemburg norm $\| \quad \|_A$. Recall that a Young function A is said to satisfy the Δ_2 condition near infinity if there exist positive constants K and s_0 such that

$$A(2s) \leq KA(s) \qquad \text{for every } s \geq s_0.$$

We shall also write "$A \in \Delta_2$ near infinity" to denote that A satisfies the Δ_2-condition near infinity. Notice that, as a consequence of [RR, Chapter VII, Theorem 4], $A \in \Delta_2$ near infinity if and only if

$$\int_{\mathbb{R}^n} A\left(\frac{f(x)}{\| f \|_A}\right) dx = 1 \quad \text{for every } f \in L^1_+(\mathbb{R}^n) \cap L^A(\mathbb{R}^n) \setminus \{0\}. \tag{5.12}$$

Clearly, any power t^p, with $p \geq 1$, satisfies the Δ_2 condition near infinity. Functions $A(t)$ behaving like $t^p \log^\alpha(1 + t)$ for large t are in Δ_2 near infinity as well. On the other hand, if $A(t)$ has an exponential growth at infinity of type e^{t^α} for some $\alpha > 0$, then $A \notin \Delta_2$ near infinity. Other instances of Young functions which are not in Δ_2 near infinity are those "oscillating" between different powers.

PROPOSITION 5.7. *Let A be an N-function. Then*

(i) $\| \quad \|_A$ *is $**$-increasing;*

(ii) $\| \quad \|_A$ *is $*$-strictly increasing if and only if $A \in \Delta_2$ near infinity;*

(iii) $\| \quad \|_A$ *is $**$-strictly increasing if and only if $A \in \Delta_2$ near infinity and A is strictly convex.*

Proof. (i) $\|\quad\|_A$ is an r.i. norm. Thus, by Proposition 5.2, it is $**$-increasing.

(ii) Assume that $A \in \Delta_2$ near infinity. Suppose, by contradiction, that $\|\quad\|_A$ is not $*$-strictly monotone. Then there exist two functions f and g in $L^1_+(\mathbb{R}^n) \cap L^A(\mathbb{R}^n)$ and a measurable set $E \subset (0,\infty)$ with $\mathcal{L}^1(E) > 0$ such that

$$f^*(s) \le g^*(s) \qquad \text{for } s \ge 0, \qquad f^*(s) < g^*(s) \qquad \text{for } s \in E, \tag{5.13}$$

and

$$\|f\|_A = \|g\|_A. \tag{5.14}$$

We have

$$\begin{aligned} 1 \ge \int_{\mathbb{R}^n} A\left(\frac{g(x)}{\|g\|_A}\right) dx &= \int_0^\infty A\left(\frac{g^*(s)}{\|g\|_A}\right) ds > \int_0^\infty A\left(\frac{f^*(s)}{\|g\|_A}\right) ds \\ &= \int_0^\infty A\left(\frac{f^*(s)}{\|f\|_A}\right) = \int_{\mathbb{R}^n} A\left(\frac{f(x)}{\|f\|_A}\right) dx = 1, \end{aligned} \tag{5.15}$$

a contradiction. Notice that the second inequality in (5.15) is due to (5.13) and to the strict monotonicity of A, the second equality to (5.14), and the last equality to (5.12).

Conversely, suppose that $\|\quad\|_A$ is $*$-strictly increasing. By (5.12) again, if $A \notin \Delta_2$ near infinity, then there exists $f \in L^1_+(\mathbb{R}^n) \cap L^A(\mathbb{R}^n)$ such that

$$\int_{\mathbb{R}^n} A\left(\frac{f(x)}{\|f\|_A}\right) dx < 1. \tag{5.16}$$

Given three positive numbers s_0, ϵ, and δ, consider the function $g \in L^1_+(\mathbb{R}^n)$ defined by

$$g(x) = \begin{cases} f^\bigstar(x) & \text{if } \omega_n|x|^n \notin [s_0, s_0+\delta], \\ f^*(s_0) + \epsilon & \text{if } \omega_n|x|^n \in [s_0, s_0+\delta]. \end{cases}$$

Clearly, $f^*(s) \le g^*(s)$ for $s > 0$, and $f^*(s) < g^*(s)$ if $s \in [s_0, s_0+\delta]$. Consequently,

$$\|f\|_A < \|g\|_A. \tag{5.17}$$

On the other hand,

$$\begin{aligned} \int_{\mathbb{R}^n} A\left(\frac{g(x)}{\|f\|_A}\right) dx &= \int_{\{x:\omega_n|x|^n \notin [s_0, s_0+\delta]\}} A\left(\frac{f^\bigstar(x)}{\|f\|_A}\right) dx \\ &\quad + \int_{\{x:\omega_n|x|^n \in [s_0, s_0+\delta]\}} A\left(\frac{f^*(s_0)+\epsilon}{\|f\|_A}\right) dx \\ &\le \int_{\mathbb{R}^n} A\left(\frac{f(x)}{\|f\|_A}\right) dx + \delta A\left(\frac{f^*(s_0)+\epsilon}{\|f\|_A}\right). \end{aligned} \tag{5.18}$$

Owing to (5.16) and (5.18), we may choose $\delta > 0$ so small that $\int_{\mathbb{R}^n} A\left(\frac{g(x)}{\|f\|_A}\right) dx \le 1$. Hence, $\|g\|_A \le \|f\|_A$, thus contradicting (5.17).

(iii) Suppose that A is strictly convex and satisfies the Δ_2-condition near infinity. In order to prove that $\|\quad\|_A$ is $**$-strictly increasing, observe that if functions $f, g \in L^1_+(\mathbb{R}^n) \cap L^A(\mathbb{R}^n)$ existed such that

$$f^{**}(s) \le g^{**}(s) \quad \text{for } s > 0, \quad f^* \not\equiv g^*, \quad \text{and } \|f\|_A = \|g\|_A,$$

then we would have the contradiction

$$1 = \int_{\mathbb{R}^n} A\left(\frac{g(x)}{\|g\|_A}\right) dx = \int_{\mathbb{R}^n} A\left(\frac{g(x)}{\|f\|_A}\right) dx > \int_{\mathbb{R}^n} A\left(\frac{f(x)}{\|f\|_A}\right) dx = 1,$$

where the first and the last equality hold thanks to (5.12), and the inequality is due to the fact that, by Proposition 5.6, the functional J_A is $**$-strictly increasing.

Conversely, if $\|\quad\|_A$ is $**$-strictly increasing, then it is, in particular, $*$-strictly increasing, and hence, by (ii), $A \in \Delta_2$ near infinity. We have to show that A is also strictly convex. Assume, by contradiction, that this is not the case. Then there exist $s_0, t_1 \in (0, +\infty)$, $a > 0$, and $b \leq 0$ such that (5.11) holds. Choose C, ϵ, E_1, and E_2 satisfying the same properties as in the proof of Proposition 5.6, and, in addition,

$$2aC\mathcal{L}^n(E_1) + b = 1.$$

Let f and g be defined as in the proof of Proposition 5.6. Then,

$$\int_{\mathbb{R}^n} A(f(x))dx = aC\mathcal{L}^n(E_1) + aC\mathcal{L}^n(E_2) + b = 2aC\mathcal{L}^n(E_1) + b = 1$$

and

$$\int_{\mathbb{R}^n} A\left(\frac{f(x)}{\lambda}\right) dx > 1 \qquad \text{if } \lambda < 1,$$

whence $\|f\|_A = 1$. On the other hand,

$$\int_{\mathbb{R}^n} A(g(x))dx = a\left((C+\epsilon)\mathcal{L}^n(E_1) + (C-\epsilon)\mathcal{L}^n(E_2)\right) + b = 1,$$

and consequently, $\|g\|_A \leq 1$. Since $f \prec g$, $\|f\|_A = \|g\|_A$, and since $f^* \not\equiv g^*$, this contradicts the $**$-strict monotonicity of $\|\quad\|_A$. □

The Orlicz norm $\|\quad\|_{(A)}$ is considered in Proposition 5.8 below. Its proof requires a characterization of $\|\quad\|_{(A)}$, which tells us that, if A is an N-function, then

$$\|f\|_{(A)} = \min\left\{\frac{1}{k}\left(1 + \int_{\mathbb{R}^n} A\left(kf(x)\right) dx\right) : k > 0\right\} \tag{5.19}$$

for every $f \in L^1_+(\mathbb{R}^n) \cap L^A(\mathbb{R}^n)$ (see [RR, Chapter III, Theorem 13]).

PROPOSITION 5.8. *Let A be an N-function. Then*

(i) $\|\quad\|_{(A)}$ *is $**$-increasing;*
(ii) $\|\quad\|_{(A)}$ *is $*$-strictly increasing;*
(iii) $\|\quad\|_{(A)}$ *is $**$-strictly increasing if and only if A is strictly convex.*

Proof. (i) $\|\quad\|_{(A)}$ is an r.i. norm, and hence, by Proposition 5.2, it is $**$-increasing.

(ii) Let $f, g \in L^1_+(\mathbb{R}^n) \cap L^A(\mathbb{R}^n)$ be such that

(5.20)

$$f^*(s) \leq g^*(s) \quad \text{for } s > 0 \quad \text{and } f^*(s) < g^*(s) \text{ for } s \text{ in a set of positive measure.}$$

By (5.19) applied to g, a positive number k_g exists such that

$$\|g\|_{(A)} = \frac{1}{k_g}\left(1 + \int_{\mathbb{R}^n} A(k_g g(x))dx\right). \tag{5.21}$$

We have

$$\|g\|_{(A)} = \frac{1}{k_g}\left(1+\int_{\mathbb{R}^n} A(k_g g(x))dx\right) = \frac{1}{k_g}\left(1+\int_0^\infty A(k_g g^*(s))ds\right) \tag{5.22}$$
$$> \frac{1}{k_g}\left(1+\int_0^\infty A(k_g f^*(s))ds\right) = \frac{1}{k_g}\left(1+\int_{\mathbb{R}^n} A(k_g f(x))dx\right) \geq \|f\|_{(A)},$$

where the first inequality holds thanks to (5.20), and the last inequality is a consequence of (5.19). The $*$-strict monotonicity of $\|\quad\|_{(A)}$ follows.

(iii) Assume that A is strictly convex. By Proposition 5.6, the functional J_A is $**$-strictly increasing. Thus, if $f, g \in L^1_+(\mathbb{R}^n) \cap L^A(\mathbb{R}^n)$ are such that $f^{**}(s) \leq g^{**}(s)$ for $s > 0$ but $f^* \not\equiv g^*$, then

$$\int_{\mathbb{R}^n} A(k_g g(x))dx > \int_{\mathbb{R}^n} A(k_g f(x))dx,$$

where k_g is the number appearing in (5.21). The same chain as in (5.22) tells us that $\|f\|_{(A)} < \|g\|_{(A)}$, and the $**$-strict monotonicity of $\|\quad\|_{(A)}$ follows.

Finally, assume that $\|\quad\|_{(A)}$ is $**$-strictly increasing, and suppose, by contradiction, that A is not strictly convex. Then there exist an interval $(t_0, t_1) \subset (0, +\infty)$ and $a > 0$, $b \leq 0$ such that (5.11) holds. Let f be any function in $L^1_+(\mathbb{R}^n) \bigcap L^A(\mathbb{R}^n)$ such that f^* is continuous and strictly decreasing in $(0, +\infty)$ and $\lim_{s\to 0^+} f^*(s) = +\infty$. Let k_f be a positive number such that

$$\|f\|_{(A)} = \frac{1}{k_f}\left(1+\int_{\mathbb{R}^n} A(k_f f(x))dx\right). \tag{5.23}$$

Since f^* is continuous and $f^*(0, +\infty) = [0, +\infty)$, there exist $\bar{s} > 0$ and $\delta > 0$ satisfying

$$t_0 < k_f f^*(s) < t_1 \qquad \text{if } \ \bar{s} - \delta \leq s \leq \bar{s} + \delta.$$

Let g be the function in $L^1_+(\mathbb{R}^n)$ obeying

$$g^*(s) = \begin{cases} f^*(s) & \text{if } s \notin (\bar{s}-\delta, \bar{s}+\delta), \\ f^*(\bar{s}-\delta) & \text{if } \bar{s}-\delta < s < s_0, \\ f^*(\bar{s}+\delta) & \text{if } s_0 \leq s < \bar{s}+\delta, \end{cases}$$

where s_0 is chosen in $(\bar{s}-\delta, \bar{s}+\delta)$ in such a way that

$$\int_{\bar{s}-\delta}^{\bar{s}+\delta} f^*(s)ds = \int_{\bar{s}-\delta}^{\bar{s}+\delta} g^*(s)ds. \tag{5.24}$$

Thus,

$$f \prec g \qquad \text{and} \qquad f^* \not\equiv g^*. \tag{5.25}$$

On the other hand, by (5.23) and (5.24),

$$\|f\|_{(A)} = \frac{1}{k_f}\left(1+\int_{[0,\infty)\setminus(\bar{s}-\delta,\bar{s}+\delta)} A(k_f f^*(s))ds + \int_{\bar{s}-\delta}^{\bar{s}+\delta}(ak_f f^*(s)+b)ds\right) \tag{5.26}$$
$$= \frac{1}{k_f}\left(1+\int_{[0,\infty)\setminus(\bar{s}-\delta,\bar{s}+\delta)} A(k_f g^*(s))ds + \int_{\bar{s}-\delta}^{\bar{s}+\delta}(ak_f g^*(s)+b)ds\right)$$
$$= \frac{1}{k_f}\left(1+\int_{\mathbb{R}^n} A(k_f g(x))dx\right) \geq \|g\|_{(A)}.$$

Equations (5.25) and (5.26) contradict the $**$-strict monotonicity of $\| \quad \|_{(A)}$. □

For ease of comparison, the results of Propositions 5.6–5.8 are collected in the following table.

A N-FUNCTION	$**$-INCREASING	$*$-STRICTLY INCREASING	$**$-STRICTLY INCREASING
J_A	any A	any A	A strictly convex
$\| \quad \|_A$	any A	$A \in \Delta_2$ near infinity	$A \in \Delta_2$ near infinity and A strictly convex
$\| \quad \|_{(A)}$	any A	any A	A strictly convex

REFERENCES

[A] A. Alvino, *Sulla disuguaglianza di Sobolev in spazi di Lorentz*, Boll. Un. Mat. Ital. A, 5 (1977), pp. 148–156.

[ALT] A. Alvino, P. L. Lions, and G. Trombetti, *On optimization problems with prescribed rearrangements*, Nonlinear Anal., 13 (1989), pp. 185–220.

[AT] A. Alvino and G. Trombetti, *Sulle migliori costanti di maggiorazione per una classe di equazioni ellittiche degeneri*, Ricerche Mat., 27 (1978), pp. 413–428.

[AFP] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, Oxford, UK, 2000.

[BS] C. Bennett and R. Sharpley, *Interpolation of Operators*, Academic Press, Boston, 1988.

[BZ] J. E. Brothers and W. P. Ziemer, *Minimal rearrangements of Sobolev functions*, J. Reine Angew. Math., 384 (1988), pp. 153–179.

[B] A. Burchard, *Steiner symmetrization is continuous in $W^{1,p}$*, Geom. Funct. Anal., 7 (1997), pp. 823–860.

[CR] K. M. Chong and N. M. Rice, *Equimeasurable Rearrangements of Functions*, Queens Paper in Pure and Appl. Math., vol. 28, Queen's University, Ontario, 1971.

[CF1] A. Cianchi and N. Fusco, *Functions of bounded variation and rearrangements*, Arch. Ration. Mech. Anal., 165 (2002), pp. 1–40.

[CF2] A. Cianchi and N. Fusco, *Minimal rearrangements, strict convexity and critical points*, Appl. Anal., to appear.

[CP] A. Cianchi and L. Pick, *Sobolev embeddings into BMO, VMO and L^∞*, Ark. Mat., 36 (1998), pp. 317–340.

[CNV] D. Cordero-Erasquin, B. Nazaret, and C. Villani, *A mass-transportation approach to sharp Sobolev and Gagliardo–Nirenberg inequalities*, Adv. Math., 182 (2004), pp. 307–332.

[FV1] A. Ferone and R. Volpicelli, *Polar factorization and pseudo-rearrangements: Applications to Pòlya–Szegö-type inequalities*, Nonlinear Anal., 53 (2003), pp. 929–949.

[FV2] A. Ferone and R. Volpicelli, *Minimal rearrangements of Sobolev functions: A new proof*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 333–339.

[F] J. J. F. Fournier, *Mixed norms and rearrangements: Sobolev's inequality and Littlewood's inequality*, Ann. Mat. Pura Appl., 4 (1987), pp. 51–76.

[K] V. I. Kolyiada, *Rearrangements of functions and embedding theorems*, Russ. Math. Surv., 44 (1989), pp. 349–374.

[M] F. Maggi, *A Remark on the Optimal Functions in the Sobolev Inequality*, preprint, 2003.

[Ra] J. M. Rakotoson, *General pointwise relations for the relative rearrangement and applications*, Appl. Anal., 80 (2001), pp. 201–232.

[RT] J. M. Rakotoson and R. Temam, *A co-area formula with applications to monotone rearrangement and to regularity*, Arch. Ration. Mech. Anal., 109 (1990), pp. 213–238.

[RR] M. M. Rao and Z. D. Ren, *Theory of Orlicz Spaces*, Marcel Dekker, New York, 1991.

[Ta] G.Talenti, *Best constant in Sobolev inequality*, Ann. Mat. Pura Appl., 110 (1976), pp. 353–372.

[Z] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 309–332

CONTRACTING LORENZ ATTRACTORS THROUGH RESONANT DOUBLE HOMOCLINIC LOOPS*

C. A. MORALES†, M. J. PACIFICO†, AND B. SAN MARTIN‡

Abstract. A *contracting Lorenz attractor* of a three-dimensional vector field is an attractor with a unique singularity whose eigenvalues are real and satisfy the eigenvalue conditions $\lambda_{ss} < \lambda_s < 0 < \lambda_u$ and $\lambda_s + \lambda_u < 0$. The study of contracting Lorenz attractors started in [A. Rovella, *Bol. Soc. Brasil. Mat. (N.S.),* 24 (1993), pp. 233–259]. In this paper we show that certain resonant double homoclinic loops in dimension three generate contracting Lorenz attractors in a positive Lebesgue subset of the parameter space. This gives a positive answer to a question posed in [C. Robinson, *SIAM J. Math. Anal.*, 32 (2000), pp. 119–141].

Key words. contracting Lorenz attractors, resonant double homoclinic loops

AMS subject classifications. 37G10, 37G15, 37G25

DOI. 10.1137/S0036141004443907

1. Introduction.

Inspired by the dynamical features of Lorenz and Henon-like attractors, Rovella [Rov] studied attractors with a *unique* singularity exhibiting real eigenvalues $\lambda_{ss}, \lambda_s, \lambda_u$ satisfying the eigenvalue conditions

$$\lambda_{ss} < \lambda_s < 0 < \lambda_u \quad \text{and} \quad \lambda_s + \lambda_u < 0.$$

Today these kinds of attractors are denominated *contracting Lorenz attractors*, as opposed to *expanding Lorenz attractors* [GW] in which the second eigenvalue condition is replaced by $\lambda_s + \lambda_u > 0$. The result obtained by Rovella was that there are contracting Lorenz attractors which are measure-theoretical persistent in parametrized families of codimension two. In the meantime Robinson [Rob1], [Rob2] developed a theory of expanding Lorenz attractors appearing in the unfolding of resonant double homoclinic loops in dimension three. These works are the motivation for our investigation. Indeed, we prove that positive Lebesgue measure sets of vector fields exhibiting contracting Lorenz attractors also appear in the unfolding of certain resonant double homoclinic loops. This gives a positive answer for a question posed by Robinson in [Rob1, Remark 5.1].

Let us state our main result in a precise way. Hereafter M will be a closed 3-manifold whose space of C^r-vector fields is denoted by $\mathcal{X}^r(M)$, $r \geq 1$. If $X \in \mathcal{X}^r(M)$, we denote by X^t the flow it generates. An invariant set Λ of X is *transitive* if $\Lambda = \omega_X(q)$ for some $q \in \Lambda$, where

$$\omega_X(q) = \left\{ x \in M : \lim_{n\to\infty} X^{t_n}(q) \text{ for some sequence } t_n \to \infty \right\}$$

*Received by the editors May 13, 2004; accepted for publication (in revised form) September 15, 2005; published electronically May 5, 2006. This work is partially supported by CNPq, FAPERJ, Pronex on Dyn. Systems, Fondecyt grant 1040682, and Mecesup.
http://www.siam.org/journals/sima/38-1/44390.html
†Instituto de Matemática, Universidade Federal do Rio de Janeiro, C. P. 68.530, CEP 21.945-970, Rio de Janeiro, R. J., Brazil (morales@impa.br, pacifico@impa.br).
‡Departamento de Matemáticas, Universidad Católica del Norte, Casilla 1280, Antofagasta, Chile (sanmarti@ucn.cl).

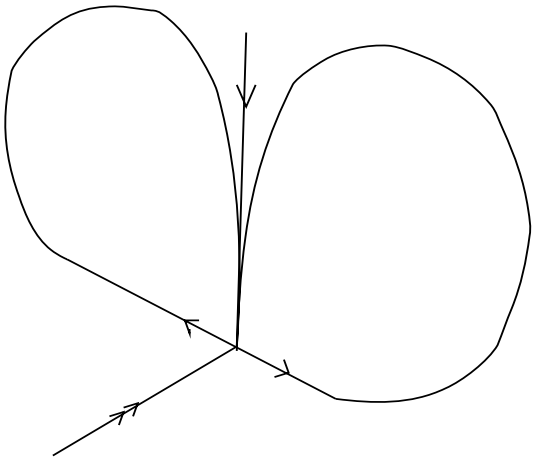


FIG. 1.1. *A butterfly loop for* X_{η_0}.

is the omega-limit set of q. An *attracting set* of X is a compact invariant set Λ such that

$$\Lambda = \bigcap_{t>0} X^t(U)$$

for some compact neighborhood U of Λ. We note that U above can be chosen *positively invariant*, i.e., $X^t(U) \subset \mathrm{Int}(U)$ for all $t > 0$. An *attractor* is a transitive attracting set.

Let Q be a singularity of $X \in \mathcal{X}^r(M)$ with real eigenvalues $\lambda_u, \lambda_{ss}, \lambda_s$ satisfying

$$\lambda_{ss} < \lambda_s < 0 < \lambda_u.$$

Then Q is hyperbolic, and so there are stable and unstable C^r manifolds $W^u(Q,X)$ and $W^s(Q,X)$ tangent to the eigenvectors associated to the eigenvalues λ_u and $\{\lambda_{ss}, \lambda_s\}$, respectively [HPS]. In particular, $W^u(Q,X)$ is one-dimensional and so $W^u(Q,X) \setminus \{Q\}$ consists of two regular orbits Γ^+, Γ^-. There are also a strong stable manifold $W^{ss}(Q,X)$ tangent to the eigenvector associated to λ_{ss} and a central unstable manifold $W^{cu}(Q,X)$ tangent to the eigendirection associated to λ_s, λ_u. The latter one is C^1, is nonunique, and contains $W^u(Q,X)$. Although $W^{cu}(Q,X)$ is not unique in general we can define the plane field $P = \{P(q) : q \in W^u(Q,X)\}$ by

$$P(q) = T_q W^{cu}(Q,X) \quad \text{for} \quad q \in W^u(Q,X).$$

We note that the plane field P defined above is continuous if $W^{cu}(Q,X)$ is transverse to $W^s(Q,X)$ along $W^u(Q,X)$.

We say that X has a *double homoclinic loop at* Q if

$$W^u(Q,X) \subset W^s(Q,X)$$

(equivalently, $\Gamma^+ \cup \Gamma^- \subset W^s(Q,X)$). The double homoclinic loop is said to be either *figure-eight* or *butterfly* depending on whether the regular orbits Γ^+, Γ^- are contained in the same connected component of $W^s(Q,X) \setminus W^{ss}(Q,X)$ or not. (See Figure 1.1.) All double homoclinic loops considered in this paper will satisfy the generic assumption

$$W^u(Q,X) \cap W^{ss}(Q,X) = \{Q\}.$$

Now suppose that X has a double homoclinic loop at Q. If X is C^r with $r \geq 3$, and the eigenvalue condition

$$\lambda_{ss} < 2\lambda_s$$

holds, then $W^{cu}(Q, X)$ is C^2. This permits us to define a constant $C^+(X)$ by taking a parametrization $q^+(t)$ of Γ^+ and defining

$$C^+(X) = \exp\left(\int_{-\infty}^{\infty} Div_2(q^+(t))dt\right),$$

where $Div_2(q^+(t))$ is the Jacobian of X at t restricted to $T_{\Gamma^+}W^{cu}(Q, X)$. Analogously we define the constant $C^-(X)$ just replacing + by − in the above expression. These constants represent the change in area within the planes $P(q)$ along the whole length of $\Gamma^{\pm}$. It happens that $C^+(X), C^-(X)$ are finite and positive when Q is *resonant*, i.e.,

$$\lambda_u + \lambda_s = 0.$$

Then, in the resonant case, we can define the constant $B(X)$ by

$$B(X) = \frac{C^+(X) + C^-(X)}{C^+(X)C^-(X)}.$$

DEFINITION 1.1. *Define $\mathcal{N}$ as the set of vector fields $X \in \mathcal{X}^r(M)$ satisfying the following six properties:*

(A1) *X has a singularity Q whose eigenvalues $\{\lambda_u, \lambda_s, \lambda_{ss}\}$ are real and satisfy*

$$\lambda_{ss} < \lambda_s < 0 < \lambda_u.$$

(A2) *X exhibits a* butterfly *double homoclinic loop at Q.*
(A3) *$W^{cu}(Q, X)$ is transverse to $W^s(Q, X)$ along $W^u(Q, X)$.*
(A4) *$\lambda_{ss} - \lambda_s + 2\lambda_u < 0$ and $\lambda_{ss} < 2\lambda_s$.*
(A5) *$B(X) < 1$.*
(A6) *$\lambda_u + \lambda_s = 0$.*

The implicit function theorem in Banach spaces implies that $\mathcal{N}$ is a codimension three submanifold of $\mathcal{X}^r(M)$ for all $r \geq 1$. Given a set A we denote its closure by $\mathrm{Cl}(A)$. Our main result is the following.

THEOREM 1.2. *Let $\{X_\eta\}$ be a C^k three-parametrized family of C^r-vector fields transverse to $\mathcal{N}$ at X_{η_0}, $k, r \geq 3$. Then, there is a positive Lebesgue measure set $\mathcal{L}$ in the parameter space with $\eta_0 \in \mathrm{Cl}(\mathcal{L})$ such that X_η has a contracting Lorenz attractor for all $\eta \in \mathcal{L}$.*

The proof of this theorem will be reduced to Theorem 2.2 in section 2 where one-parameter families of discontinuous maps are studied.

The present paper, together with [MPS], gives a full description of the bifurcation scenario in the case when the constant $B \neq 1$: The family generates expanding or contracting Lorenz attractors depending on whether $B > 1$ or $B < 1$. Viana posed the question of what bifurcation scenario can appear in the case $B = 1$. Another problem is to exhibit a polynomial vector field with a contracting Lorenz attractor. To find such a polynomial we expect to find a parametrized polynomial vector field family satisfying the hypotheses of Theorem 1.2.

2. One-dimensional reduction. In this section we deduce Theorem 1.2 from Theorem 2.2 stated below. Hereafter we let X_η be a one-parameter family satisfying

the hypotheses of Theorem 1.2. In particular, the parameter η_0 satisfies $X_{\eta_0} \in \mathcal{N}$ and then the corresponding vector field X_{η_0} satisfies assumptions (A1)–(A6).

Throughout the paper we fix the notation

$$C^+ = C^+(X_{\eta_0}), \quad C^- = C^-(X_{\eta_0}), \quad \text{and} \quad B = B(X_{\eta_0}).$$

Let P be the plane field used in the introduction to define the constants $C^{\pm}$. Consider the regular orbits Γ^+, Γ^- forming $W^u(Q_{\eta_0}, X_{\eta_0}) \setminus \{Q_{\eta_0}\}$.

Define $\nu^+ = 1$ (if P is orientable in $\Gamma^+ \cup \{Q_{\eta_0}\}$) and $\nu^+ = -1$ (otherwise). Analogously we define $\nu^- = 1$ (if P is orientable in $\Gamma^- \cup \{Q_{\eta_0}\}$) and $\nu^- = -1$ (otherwise).

By assumption (A2) we can fix a cross-section Σ of X_{η_0} with the following properties: Σ is close to Q_{η_0}, Σ is transversal to $W^s(Q_{\eta_0})$, and Σ intersects both Γ^+ and Γ^-. There is a neighborhood V of $\Sigma \cap W^s(Q_{\eta_0})$ in Σ such that the positive orbit of every point at $V \setminus W^s(Q_{\eta_0})$ intersects Σ for every parameter η near enough η_0. This defines a Poincaré map from $V \setminus W^s(Q_{\eta_0})$ to Σ.

As in [MPS] (see also [Rob2]) we can use (A1)–(A4) and the standard stable manifold theory [HPS], [Sh] to show the existence of a stable foliation in a small neighborhood of $W^s(Q_\eta)$ varying C^2 with the parameter. For convenience, we assume this neighborhood to be equal to V. Hence we can reduce the problem to a one-dimensional Poincaré map

$$f_\eta : V' \setminus \{c_\eta\} \subset [-1,1] \to [-1,1],$$

where c_η is the projection of $W^s(Q_{\eta_0}) \cap V$ onto V'. We assume $c_\eta = 0$ for every η. We shall use the notation

$$a_\eta^{\pm} = f_\eta(0\pm) = \lim_{\tau \to \pm 0} f_\eta(\tau).$$

As in [MPS] or [Rob2] we have the following.

LEMMA 2.1. *There is an interval J, $0 \in J$, such that for every η sufficiently near to η_0, the map $f_\eta : J \subset [-1,1] \to [-1,1]$ has the following form:*

$$f_\eta(\tau) = \begin{cases} a_\eta^+ + \nu^+ C_\eta^+ |\tau|^{\alpha(\eta)} + O_{\eta,1}(|\tau|^{\alpha(\eta)}) & \text{if} \quad \tau > 0, \\ a_\eta^- - \nu^- C_\eta^- |\tau|^{\alpha(\eta)} + O_{\eta,2}(|\tau|^{\alpha(\eta)}) & \text{if} \quad \tau < 0, \end{cases}$$

where $\alpha(\eta) = -\frac{\lambda_s(\eta)}{\lambda_u(\eta)}$, $O_{\eta,i}$ and $C_\eta^{\pm}$ are C^2, varying C^2 with respect to η, and $\lim_{x \to 0} \frac{O_{\eta,i}(x)}{x} = 0$ uniformly on η.

The proof of Theorem 1.2 is a direct consequence of the result below. We denote by $\mathrm{Int}(A)$ the interior and by $\mathrm{Cl}(A)$ the closure of a subset A.

THEOREM 2.2. *Let $\{X_\eta\}$ be a C^k three-parametrized family of C^r-vector fields transverse to $\mathcal{N}$ at X_{η_0}, $k, r \geq 3$. Let f_η be the corresponding three-parametrized family of one-dimensional maps as in Lemma 2.1. For all neighborhoods $\mathcal{V}$ of η_0 in the parameter space there is a positive Lebesgue measure set $\mathcal{L}' \subset \mathcal{V}$ with $\eta_0 \in \mathrm{Cl}(\mathcal{L}')$ such that for every $\eta \in \mathcal{L}'$ one has $\alpha(\eta) > 1$ and further there is a closed nontrivial interval Θ_η with the following properties:*

1. $0 \in \mathrm{Int}(\Theta_\eta)$.
2. *There is a compact interval $W \subset J$ with $\Theta_\eta \subset \mathrm{Int}(W)$ such that $\mathrm{Cl}(f_\eta(W \setminus \{0\})) \subset \mathrm{Int}(W)$ and $f_\eta^2(W \setminus \{0\}) \subset \Theta_\eta$.*
3. *$f_\eta/_{\Theta_\eta \setminus \{0\}}$ is transitive.*

Proof of Theorem 1.2. Let $\{X_\eta\}$ be a C^k three-parametrized family of C^r-vector fields transverse to $\mathcal{N}$ at X_{η_0}, $k, r \geq 3$. We have to prove that there is a positive Lebesgue measure set $\mathcal{L}$ in the parameter space with $\eta_0 \in \mathrm{Cl}(\mathcal{L})$ such that X_η has a contracting Lorenz attractor for all $\eta \in \mathcal{L}$. For this we proceed as follows.

As hypothesis (A1) is an open condition we can fix a neighborhood $\mathcal{V}$ of η_0 in the parameter space such that

$$\lambda_{ss}(\eta) < \lambda_s(\eta) < 0 < \lambda_u(\eta) \qquad \forall\, \eta \in \mathcal{V}. \tag{2.1}$$

Since Q_{η_0} in (A1) is hyperbolic we can consider the analytic continuation Q_η of Q_{η_0} which we assume to be defined for all $\eta \in \mathcal{V}$.

Now, let f_η be the three-parameter family of one-dimensional maps associated to X_η as in Lemma 2.1. Let $\mathcal{L}'$ be the positive Lebesgue measure set in Theorem 2.2 for the fixed neighborhood $\mathcal{V}$ of η_0. Define $\mathcal{L} = \mathcal{L}'$. We shall prove that $\mathcal{L}$ works. First we observe that $\eta_0 \in \mathrm{Cl}(\mathcal{L})$ by Theorem 2.2.

Next we claim that if $\eta \in \mathcal{L}$, then X_η has a contracting Lorenz attractor. Indeed, let π_η be the projection along the invariant stable foliation defining f_η. Define

$$\Sigma_\eta = \pi_\eta^{-1}(\Theta_\eta).$$

It follows that $\Sigma_\eta \subset \Sigma$ is a cross-section of X. We note that the set

$$\bigcup_{t>0} X_\eta^t(\Sigma_\eta)$$

is positively invariant but not closed since there exists a compact part W_η of $W^u(Q_\eta, X_\eta)$ which is not contained in it. However, by item 2 of Theorem 2.2, we can choose a small tubular neighborhood $\mathcal{U}_\eta$ of W_η in a way that

$$U_\eta = \mathcal{U}_\eta \cup \left(\bigcup_{t>0} X_\eta^t(\Sigma_\eta) \right)$$

is now compact and positively invariant. Moreover, by a suitable modification of U_η via the long tubular flow-box theorem [dMP], we can even assume that ∂U_η is smooth and the vector field X_η points inward to U_η in ∂U_η.

Now define

$$\Lambda_\eta = \bigcap_{t>0} X_\eta^t(U_\eta).$$

Item 3 of Theorem 2.2 implies that Λ_η is a transitive set of X_η because the foliation which defines π_η is contracting; see also the proof of Theorem 11.3 in [Rob3]. In particular, Λ_η is compact invariant. On the other hand, $\Lambda \subset \mathrm{Int}(U_\eta)$ because X_η points inward to U_η in ∂U_η. It follows that U_η is a compact neighborhood of Λ_η, and so Λ_η is an attractor.

To finish we prove that Λ is a contracting Lorenz attractor. By item 1 of Theorem 2.2 we have

$$Q_\eta \in \Lambda_\eta,$$

where Q_η is the continuation of Q_{η_0}. As $\eta \in \mathcal{L} \subset \mathcal{V}$ we obtain (2.1). Recalling that $\alpha(\eta) = -\frac{\lambda_s(\eta)}{\lambda_u(\eta)}$ in Lemma 2.1 we have $\lambda_s(\eta) + \lambda_u(\eta) < 0$ because $\alpha(\eta) > 1$ in Theorem 2.2. As $Q_\eta \in \Lambda_\eta$ we conclude that Λ_η is a contracting Lorenz attractor. This finishes the proof. □

The rest of the paper is devoted to the proof of Theorem 2.2. It is done in sections 3–8. In section 3 we restate Lemma 2.2 of [MPS] for $\alpha > 1$ because $B < 1$ in the present case. In section 4 we define new parameters $(\mu(\alpha), \nu(\alpha))$ and study their limits as $\alpha \to 1^+$. These limits will be preperiodic for the piecewise linear map g defined in Lemma 4.2. In section 5 we introduce the rescaling map g_α (see (5.3)). We prove in Lemma 5.7 the convergence g_α (in the C^0, C^1, and C^2 topologies) to a piecewise linear map. We finish section 5 by proving in Theorem 5.8 the existence of invariant intervals for $g_\alpha(\mu, \nu, \cdot)$ for almost all $\alpha \to 1^+$. In section 6 we define the Rovella map and state the Rovella theorem which asserts that a one-parameter unfolding of a Rovella map displays a full density set of parameters exhibiting transitive invariant closed intervals. We refer the reader to the proof of Rovella's theorem in the original paper [Rov]. In section 7 we define the map h_α and prove that it is a Rovella map (Theorem 7.1). In addition, in Theorem 7.3, we use g_α to define a one-parameter family of maps φ_t unfolding h_α and satisfying the hypothesis of Rovella's theorem. As a corollary we obtain that the invariant interval of $g_\alpha(\mu, \nu, \cdot)$ found in section 5 *contains* a transitive invariant interval for a positive Lebesgue measure set of parameters (μ, ν) (Corollary 7.4). We use this interval in section 8 to define the interval Θ_η required in Theorem 2.2. This complete the proof.

3. Preliminary properties of f_η. These are summarized in Lemma 3.1 below. From now on we assume that $\eta \in \mathbb{R}^3$ belongs to a neighborhood of η_0 and X_η satisfies the hypothesis of Theorem 1.2. In particular, $X = X_{\eta_0}$ satisfies (A1)–(A7) in Definition 1.1. By the transversal hypothesis (A7) we have that the map $\eta \mapsto (\alpha_\eta, a^+_\eta, a^-_\eta)$ is a diffeomorphism from a neighborhood of η_0 onto a neighborhood of $(1, 0, 0)$. This allows us to consider the inverse

$$\eta = \eta(\alpha, a^+, a^-) \tag{3.1}$$

of this diffeomorphism. In particular,

$$\eta_0 = \eta(1, 0, 0)$$

and also

$$\alpha(\eta(\alpha, a^+, a^-)) = \alpha, \quad a^+_{\eta(\alpha, a^+, a^-)} = a^+, \quad a^-_{\eta(\alpha, a^+, a^-)} = a^-. \tag{3.2}$$

The next lemma is analogous to Lemma 2.2 in [MPS]. The difference here is that in [MPS] we found $O \subset (0, 1)$ and here we found $O \subset (1, 1 + \Lambda)$ for some $\Lambda > 0$. The proof follows as in [MPS] except, for instance, in the case where $\nu^+ = \nu^- = 1$, when the system of equations $f_\eta(p(\alpha)) = p(\alpha)$, $f_\eta(q(\alpha)) = q(\alpha)$, $f_\eta(0+) = p(\alpha)$, $f_\eta(0-) = q(\alpha)$ can be solved only for $\alpha > 1$ and not for $\alpha < 1$ as in [MPS] because $B < 1$ here and $B > 1$ there.

LEMMA 3.1. *There are $\Lambda > 0$, an open full Lebesgue measure set $O \subset (1, 1 + \Lambda)$, and C^1 maps*

$$a^+(\cdot), a^-(\cdot), p(\cdot), q(\cdot) : O \to \mathbb{R}$$

with $p(\alpha) < 0 < q(\alpha)$ such that if

$$\eta = \eta(\alpha, a^+(\alpha), a^-(\alpha)),$$

then the following hold:

(a) *If $\nu^+ = 1$ and $\nu^- = 1$, then $f_\eta(p(\alpha)) = p(\alpha)$, $f_\eta(q(\alpha)) = q(\alpha)$, $f_\eta(0+) = p(\alpha)$, and $f_\eta(0-) = q(\alpha)$.*

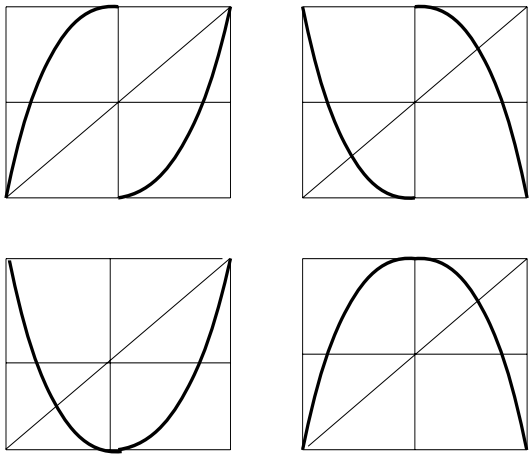


FIG. 3.1. *Possible dynamics for f_η in a neighborhood of* 0.

(b) *If $\nu^+ = 1$ and $\nu^- = -1$, then $f_\eta(p(\alpha)) = q(\alpha)$, $f_\eta(q(\alpha)) = q(\alpha)$, and $f_\eta(0+) = f_\eta(0-) = p(\alpha)$.*
(c) *If $\nu^+ = -1$ and $\nu^- = 1$, then $f_\eta(p(\alpha)) = p(\alpha)$, $f_\eta(q(\alpha)) = p(\alpha)$, and $f_\eta(0+) = f_\eta(0-) = q(\alpha)$.*
(d) *If $\nu^+ = -1$ and $\nu^- = -1$, then $f_\eta(p(\alpha)) = q(\alpha)$, $f_\eta(q(\alpha)) = p(\alpha)$, $f_\eta(0+) = q(\alpha)$, and $f_\eta(0-) = p(\alpha)$.*

In any case,

$$\lim_{\alpha\to 1^+} \frac{p(\alpha)}{q(\alpha)} = -\frac{C^+}{C^-},$$

$$\lim_{\alpha\to 1^+} q(\alpha) = 0 = \lim_{\alpha\to 1^+} p(\alpha), \text{ and } \lim_{\alpha\to 1^+} q(\alpha)^{\alpha-1} = \lim_{\alpha\to 1^+} |p(\alpha)|^{\alpha-1} = B.$$

(*See Figure* 3.1.)

4. Parameters $(\mu(\alpha), \nu(\alpha))$. Consider a parametrized family $\{X_\eta\}$ satisfying the hypotheses of Theorem 1.2. Let O be the set found in Lemma 3.1. Given $\alpha \in O$ let $a^-(\alpha)$, $a^+(\alpha)$, $p(\alpha)$, and $q(\alpha)$ be the maps in Lemma 3.1. Define

$$(\mu(\alpha), \nu(\alpha)) = \left(\frac{a^+(\alpha)}{q(\alpha)}, \frac{a^-(\alpha)}{q(\alpha)}\right). \tag{4.1}$$

The proofs of the next two lemmas follow from direct calculations and are left to the reader.

LEMMA 4.1. *The limits*

$$\mu(1) = \lim_{\alpha\to 1^+} \mu(\alpha) \quad \text{and} \quad \nu(1) = \lim_{\alpha\to 1^+} \nu(\alpha)$$

exist and satisfy the following properties:

1. *If $\nu^+ = 1$ and $\nu^- = 1$, then $\mu(1) = -\frac{C^+}{C^-}$ and $\nu(1) = 1$.*
2. *If $\nu^+ = 1$ and $\nu^- = -1$, then $\mu(1) = -\frac{C^+}{C^-}$ and $\nu(1) = -\frac{C^+}{C^-}$.*
3. *If $\nu^+ = -1$ and $\nu^- = 1$, then $\mu(1) = 1$ and $\nu(1) = 1$.*
4. *If $\nu^+ = -1$ and $\nu^- = -1$, then $\mu(1) = 1$ and $\nu(1) = -\frac{C^+}{C^-}$.*

LEMMA 4.2. *Let $\mu(1), \nu(1)$ be the limits in Lemma 4.1. Let $g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be defined by*

$$g(x) = \begin{cases} \mu(1) + \nu^+ C^+ Bx & \text{if} \quad x > 0, \\ \nu(1) + \nu^- C^- Bx & \text{if} \quad x < 0. \end{cases}$$

Then, the following properties hold:

1. *If $\nu^+ = 1$ and $\nu^- = 1$, then $g(\mu(1)) = \mu(1)$ and $g(\nu(1)) = \nu(1)$.*
2. *If $\nu^+ = 1$ and $\nu^- = -1$, then $g(\mu(1)) = 1$ and $g(\nu(1)) = 1$.*
3. *If $\nu^+ = -1$ and $\nu^- = 1$, then $g(\mu(1)) = -\frac{C^+}{C^-}$ and $g(\nu(1)) = -\frac{C^+}{C^-}$.*
4. *If $\nu^+ = -1$ and $\nu^- = -1$ $g(\mu(1)) = \nu(1)$, then and $g(\nu(1)) = \mu(1)$.*

5. The rescaling map g_α. Let $(\mu(\alpha), \nu(\alpha))$ be defined as in (4.1). Let $a^\pm(\cdot)$ be the maps in Lemma 3.1. Define in a neighborhood of $(a^+(\alpha), a^-(\alpha))$ onto a neighborhood of $(\mu(\alpha), \nu(\alpha))$ the map T_α by

$$T_\alpha(a^+, a^-) = \left(\frac{a^+}{q(\alpha)}, \frac{a^-}{q(\alpha)} \right). \tag{5.1}$$

Define the *new parameters* (μ, ν) by

$$(\mu, \nu) = T_\alpha(a^+, a^-) \tag{5.2}$$

and

$$g_\alpha(\mu, \nu, x) = \frac{1}{q(\alpha)} f_{\eta(\alpha, q(\alpha)\mu, q(\alpha)\nu)}(q(\alpha)x). \tag{5.3}$$

For the next lemma we consider $p(\alpha)$ and $q(\alpha)$ as in Lemma 3.1. For all $\alpha \in O$ we define

$$x(\alpha) = \frac{p(\alpha)}{q(\alpha)} \quad \text{and} \quad y(\alpha) = 1. \tag{5.4}$$

Combining Lemmas 3.1 and 4.1 we get the following lemma.

LEMMA 5.1. *The limits*

$$x(1) = \lim_{\alpha \to 1^+} x(\alpha) \quad \textit{and} \quad y(1) = \lim_{\alpha \to 1^+} y(\alpha)$$

exist and satisfy

$$x(1) = -\frac{C^+}{C^-} \quad \textit{and} \quad y(1) = 1.$$

In particular, we have the following:

1. *If $\nu^+ = 1$ and $\nu^- = 1$, then $\mu(1) = x(1)$ and $\nu(1) = y(1)$.*
2. *If $\nu^+ = 1$ and $\nu^- = -1$, then $\mu(1) = x(1)$ and $\nu(1) = x(1)$.*
3. *If $\nu^+ = -1$ and $\nu^- = 1$, then $\mu(1) = y(1)$ and $\nu(1) = y(1)$.*
4. *If $\nu^+ = -1$ and $\nu^- = -1$, then $\mu(1) = y(1)$ and $\nu(1) = x(1)$.*

Replacing (5.4) in (5.3) and using Lemma 3.1 for the respective cases, one gets the following lemma.

LEMMA 5.2. *For all $\alpha \in O$ consider $x(\alpha)$ and $y(\alpha)$ as in (5.4). Then the following cases hold:*

1. *If $\nu^+ = 1$ and $\nu^- = 1$, then*

$$g_\alpha(\mu(\alpha), \nu(\alpha), x(\alpha)) = x(\alpha) \quad \textit{and} \quad g_\alpha(\mu(\alpha), \nu(\alpha), y(\alpha)) = y(\alpha).$$

2. *If* $\nu^+ = 1$ *and* $\nu^- = -1$, *then*

$$g_\alpha(\mu(\alpha), \nu(\alpha), x(\alpha)) = y(\alpha) \quad \text{and} \quad g_\alpha(\mu(\alpha), \nu(\alpha), y(\alpha)) = y(\alpha).$$

3. *If* $\nu^+ = -1$ *and* $\nu^- = 1$, *then*

$$g_\alpha(\mu(\alpha), \nu(\alpha), x(\alpha)) = x(\alpha) \quad \text{and} \quad g_\alpha(\mu(\alpha), \nu(\alpha), y(\alpha)) = x(\alpha).$$

4. *If* $\nu^+ = -1$ *and* $\nu^- = -1$, *then*

$$g_\alpha(\mu(\alpha), \nu(\alpha), x(\alpha)) = y(\alpha) \quad \text{and} \quad g_\alpha(\mu(\alpha), \nu(\alpha), y(\alpha)) = x(\alpha).$$

The proof of the next lemma is straightforward.

LEMMA 5.3. *For all* $\alpha \in O$ *consider* $x(\alpha)$ *and* $y(\alpha)$ *as in* (5.4). *Then the following cases hold:*

1. *If* $\nu^+ = 1$ *and* $\nu^- = 1$, *then* $\mu(\alpha) = x(\alpha)$ *and* $\nu(\alpha) = y(\alpha)$.
2. *If* $\nu^+ = 1$ *and* $\nu^- = -1$, *then* $\mu(\alpha) = x(\alpha)$ *and* $\nu(\alpha) = x(\alpha)$.
3. *If* $\nu^+ = -1$ *and* $\nu^- = 1$, *then* $\mu(\alpha) = y(\alpha)$ *and* $\nu(\alpha) = y(\alpha)$.
4. *If* $\nu^+ = -1$ *and* $\nu^- = -1$, *then* $\mu(\alpha) = y(\alpha)$ *and* $\nu(\alpha) = x(\alpha)$.

The next definitions are the same as in [MPS] except that the differentiability is up to order C^2.

DEFINITION 5.4. *Let* $g : \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}$. *We say that* $g_\alpha \to g$ *in the* C^0 *topology in compact sets of* $\mathbb{R}^3$ *as* $\alpha \to 1^+$ *if*

(a) $Dom(g_\alpha) \to \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *as* $\alpha \to 1^+$, *that is, for all* $R > 0$ *there is* $\delta_0 > 0$ *such that if* $1 < \alpha < 1 + \delta_0$, *then* $B_R(0) \cap (\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}) \subset Dom(g_\alpha)$, *where* $B_R(0)$ *is the ball of radius* R *centered at* $(0, 0, 0)$.

(b) *for every compact set* $K \subset \mathbb{R}^3$ *and every* $\varepsilon > 0$ *there is* $\delta_0 > 0$ *such that if* $1 < \alpha < 1 + \delta_0$, *then*

$$\sup_{y \in K \cap (\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}))} |g_\alpha(y) - g(y)| < \varepsilon.$$

DEFINITION 5.5. *Let* $g : \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}$. *We say that* $g_\alpha \to g$ *in the* C^1 *topology in compact sets of* $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *if*

(a) $Dom(g_\alpha) \to \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *as* $\alpha \to 1^+$, *and*

(b) *for every compact set* $K \subset \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *and every* $\varepsilon > 0$ *there is* $\delta_0 > 0$ *such that if* $1 < \alpha < 1 + \delta_0$, *then*

$$\sup_{i \in \{0,1\}, y \in K} \left| D^i g_\alpha(y) - D^i g(y) \right| < \varepsilon.$$

DEFINITION 5.6. *We say that* $g_\alpha \to g$ *in the* C^2 *topology in x-compact sets of* $\mathbb{R} \setminus \{0\}$ *uniformly in compact sets of* $\mathbb{R}^2$ *if*

(a) $Dom(g_\alpha) \to \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *as* $\alpha \to 1^+$, *and*

(b) *for every compact* $K \subset \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *and every* $\varepsilon > 0$ *there is* $\delta_0 > 0$ *such that if* $1 < \alpha < 1 + \delta_0$, *then*

$$\sup_{i \in \{0,1,2\}, (\rho, x) \in K} \left| \partial_x^i g_\alpha(\rho, x) - \partial_x^i g(\rho, x) \right| < \varepsilon.$$

The following lemma corresponds to Lemma 3.3 in [MPS]. The difference is that in [MPS] we care about C^1 convergence and here we care about C^2 convergence.

LEMMA 5.7. *Let* g_α *be as in* (5.3). *Define*

$$g(\mu, \nu, x) = \begin{cases} \mu + \nu^+ C^+ B x & \text{if} \quad x > 0, \\ \nu + \nu^- C^- B x & \text{if} \quad x < 0. \end{cases}$$

Then the following hold:

(i) $g_\alpha \to g$ *in the* C^0 *topology in compact sets of* $\mathbb{R}^3$ *as* $\alpha \to 1^+$, $\alpha \in O$.
(ii) $g_\alpha \to g$ *in the* C^1 *topology in compact sets of* $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *as* $\alpha \to 1^+$, $\alpha \in O$.
(iii) $g_\alpha \to g$ *in the* C^2 *topology in compact sets of* $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *as* $\alpha \to 1^+$, $\alpha \in O$.

Moreover, for any $c > \max\{1, C^+/C^-\}$, *there are constants* $\alpha_0 > 1$, $0 < K_1 < K_2$ *such that for* $\alpha \in O \cap (1, \alpha_0)$ *we have the following:*

(a) $[-c, c]^2 \times ([-c, c] \setminus \{0\}) \subset Dom(g_\alpha)$.
(b) *If* $(\mu, \nu, x) \in [-c, c]^2 \times ([-c, c] \setminus \{0\})$, *then*

$$K_1 \mid x \mid^{\alpha-1} \leq \left| \frac{\partial}{\partial x} g_\alpha(\mu, \nu; x) \right| \leq K_2 \mid x \mid^{\alpha-1} . \tag{5.5}$$

(c) *If* $(\mu, \nu, x) \in [-c, c]^2 \times ([-c, c] \setminus \{0\})$, *then*

$$K_1 \mid x \mid^{\alpha-2} \leq \left| \frac{\partial^2}{\partial x^2} g_\alpha(\mu, \nu; x) \right| \leq K_2 \mid x \mid^{\alpha-2} . \tag{5.6}$$

Proof. The proofs of (i), (ii), (a), and (b) are similar to the proofs in [MPS, Lemma 3.3]. To prove (iii), put $k = q(\alpha)$, where $q(\alpha)$ is given by Lemma 3.1. Recall $k \to 0$ as $\alpha \to 1^+$. Now note that

$$\partial_x^2 g_\alpha(\mu, \nu, x) = \begin{cases} \nu^+ C_\eta^+ \alpha(\alpha-1) k^{\alpha-1} |x|^{\alpha-2} + O''_{\eta,1}(|kx|^\alpha) \alpha^2 k^{2\alpha} |x|^{2(\alpha-1)} \\ \qquad + O'_{\eta,1}(|kx|^\alpha) \alpha(\alpha-1) |kx|^{\alpha-2} k \quad \text{for} \quad x > 0, \\ -\nu^- C_\alpha^- \alpha(\alpha-1) k^{\alpha-1} |x|^{\alpha-2} - O''_{\eta,2}(|kx|^\alpha) \alpha^2 k |kx|^{2(\alpha-1)} \\ \qquad - O'_{\eta,2}(|kx|^\alpha) \alpha(\alpha-1) |kx|^{\alpha-2} k \quad \text{for} \quad x < 0. \end{cases}$$

By Lemma 2.1 $\mathcal{O}_{\eta,i}$ is C^2 and so $\mathcal{O}_{\eta,i}(|kx|^\alpha)$ is uniformly bounded in K. By Lemma 3.1 we have that $k^{2\alpha} \to 0$ and $k^{\alpha-1} \to B$ as $\alpha \to 1^+$. Since $|x|^{2\alpha-1}$ is uniformly bounded in K we finally obtain that $\partial_x^2 g_\alpha(\mu, \nu, x) \to 0$ as $\alpha \to 1^+$ in compact sets of $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$. Now note that the expression above for $\partial_x^2 g_\alpha(\mu, \nu, x)$ together with the bounds for k, $k^{2\alpha}$, $k^{\alpha-1}$ imply (c). All together conclude the proof of Lemma 5.7. □

Now we prove the existence of trapping regions for the maps $g_\alpha(\mu, \nu, \cdot)$. A *trapping region* for $g_\alpha(\mu, \nu, .)$ is a closed interval J such that $g_\alpha(\mu, \nu, J) \subset \mathrm{Int}(J)$. The following theorem corresponds to Theorem 4.1 in [MPS] with a similar proof.

Given $(\mu_0, \nu_0) \in \mathbb{R}^2$, $r > 0$ and angles $0 \leq \theta_1 < \theta_2 \leq 2\pi$ we define the cone $C_r(\mu_0, \nu_0; \theta_1, \theta_2) = \{(\mu, \nu) \in \mathbb{R}^2 : || (\mu_0, \nu_0) - (\mu, \nu) || < r; \mathrm{Arg}((\mu_0, \nu_0) - (\mu, \nu)) \in [\theta_1, \theta_2]\}$.

THEOREM 5.8. *Let* O *be the open set given in Lemma* 3.1 *and let* $(\mu(1), \nu(1))$ *be as in Lemma* 4.1. *For every neighborhood* V *of* $(\mu(1), \nu(1))$ *there is an open set* $\mathcal{O} \subset \mathbb{R}^3$ *and* $\alpha_1 > 1$ *such that if* $\alpha \in O \cap (1, \alpha_1)$, *then the set* $\mathcal{O}_\alpha$ *defined by*

$$\mathcal{O}_\alpha = \{(\mu, \nu) : (\alpha, \mu, \nu) \in \mathcal{O}\}$$

is contained in V. *In addition, for every* $\alpha \in O \cap (1, \alpha_1)$, *there is a cone field* $C_r(\mu(\alpha), \nu(\alpha); \theta_1, \theta_2) \subset \mathcal{O}_\alpha$ *such that if* $(\mu, \alpha) \in C_r(\mu(\alpha), \nu(\alpha); \theta_1, \theta_2)$, *then there exists a closed interval*

$$I_{(\alpha,\mu,\nu)} \subset [x(\alpha) - 1, y(\alpha) + 1]$$

with $0 \in \mathrm{Int}(I_{(\alpha,\mu,\nu)})$ *such that*

$$g_\alpha(\mu,\nu,x) \subset \mathrm{Int}(I_{(\alpha,\mu,\nu)}) \qquad \forall\, x \in I_{(\alpha,\mu,\nu)} \setminus \{0\}.$$

Proof. Let $(\mu(1),\nu(1))$ be as in Lemma 4.1. Fix a neighborhood V of $(\mu(1),\nu(1))$. We break the proof into four cases depending on the values of ν^+,ν^-. We work only with the case $\nu^+ = 1$, $\nu^- = 1$ since the remaining cases are similar.

Fix $\alpha \in O$. Let $x(\alpha)$ and $y(\alpha)$ be as in Lemma 5.2. By item 1 of Lemma 5.1 we have that the limits

$$x(1) = \lim_{\alpha\to 1^+} x(\alpha) \quad \text{and} \quad y(1) = \lim_{\alpha\to 1^+} y(\alpha)$$

exist and satisfy $\mu(1) = x(1)$ and $\nu(1) = y(1)$. Then, by item 1 of Lemma 4.2, $x(1)$ and $y(1)$ are fixed points of the map g defined in Lemma 4.2. We have that these fixed points are hyperbolic because $C^{\pm}B > 1$. Then, as $g_\alpha \to g$ in the C^1 topology by Lemma 5.7(ii), there are $\delta_1 > 0$ and a neighborhood $V' \subset V$ of $(\mu(1),\nu(1))$ such that the continuations $x(\alpha,\mu,\nu)$ and $y(\alpha,\mu,\nu)$ of $x(1)$ and $y(1)$ are defined for all (α,μ,ν) in the product neighborhood $(1-\delta_1, 1+\delta_1)\times V'$ with $\alpha \in O$.

Now define

$$\begin{cases} \mathcal{O} = (1-\delta_1, 1+\delta_1)\times V', \\ \alpha_1 = 1+\delta_1. \end{cases}$$

As $V' \subset V$ one has

$$\mathcal{O}_\alpha \subset V$$

for all $\alpha \in O \cap (1,\alpha_1)$.

We claim that

$$x(\alpha) = x(\alpha,\mu(\alpha),\nu(\alpha)) \quad \text{and} \quad y(\alpha) = y(\alpha,\mu(\alpha),\nu(\alpha))$$

for all $\alpha \in O \cap (1,\alpha_1)$. Indeed, by item 1 of Lemma 5.2 $x(\alpha)$ and $y(\alpha)$ are fixed points of the one-dimensional map $g_\alpha(\mu(\alpha),\nu(\alpha),\cdot)$. But $x(1) = \lim_{\alpha\to 1^+} x(\alpha)$ and $y(1) = \lim_{\alpha\to 1^+} y(\alpha)$. So, $x(\alpha) = x(\alpha,\mu(\alpha),\nu(\alpha))$ and $y(\alpha) = y(\alpha,\mu(\alpha),\nu(\alpha))$ since the continuation is unique. The claim is proved.

For all $\alpha \in O \cap (1,\alpha_1)$ we define the C^1 map $F_\alpha : V' \to \mathbb{R}^2$ by

$$F_\alpha(\mu,\nu) = (\mu - x(\alpha,\mu,\nu), \nu - y(\alpha,\mu,\nu)).$$

It follows from Lemmas 5.2 and 5.3 that $F_\alpha(\mu(\alpha),\nu(\alpha)) = (0,0)$.

Shrinking α_1 if necessary we can repeat the proof of Theorem 4.2 in [MPS] to obtain

$$\det DF_\alpha(\mu(\alpha),\nu(\alpha)) \neq 0$$

for all $\alpha \in O \cap (1,\alpha_1)$. The proof uses the above claim, the limits in Lemma 3.1, and $\alpha \to 0^+$ instead of $\alpha \to 0^-$. From this we obtain that F_α is a diffeomorphism in a neighborhood V_α of $(\mu(\alpha),\nu(\alpha))$ onto a neighborhood of $F_\alpha(\mu(\alpha),\nu(\alpha)) = (0,0)$.

Fix $\alpha \in O \cap (1,\alpha_1)$. Define $W_\alpha = F_\alpha(V_\alpha) \cap \{(x,y), x > 0, y < 0\}$. As $\{(x,y), x > 0, y < 0\}$ is a cone and F_α/V_α is a diffeomorphism there are $r > 0$ and angles $0 \le \theta_1 < \theta_2 \le 2\pi$ such that

$$C_r(\mu(\alpha),\nu(\alpha);\theta_1,\theta_2) \subset F_\alpha^{-1}(W_\alpha).$$

But $F_\alpha^{-1}(W_\alpha) \subset V_\alpha \subset V'$. Then, $C_r(\mu(\alpha), \nu(\alpha); \theta_1, \theta_2) \subset V'$ and so

$$C_r(\mu(\alpha), \nu(\alpha); \theta_1, \theta_2) \subset \mathcal{O}_\alpha$$

by the definition of $\mathcal{O}$. This proves the first part of the theorem.

Choose $\alpha \in O \cap (1, \alpha_1)$ and $(\mu, \nu) \in C_r(\mu(\alpha), \nu(\alpha); \theta_1, \theta_2)$. Repeating the argument of the proof of Theorem 4.1-(2) in [MPS] we can find the closed interval $I_{(\alpha,\mu,\nu)} \subset [x(\alpha) - 1, y(\alpha) + 1]$ with $0 \in \mathrm{Int}(I_{(\alpha,\mu,\nu)})$ required in the second part. This finishes the proof. □

COROLLARY 5.9. *Let $(\mu(1), \nu(1))$, V, $\mathcal{O}$, and α_1 be as in Theorem 5.8. Let $\alpha \in O \cap (1, \alpha_1)$ and $C_r(\mu(\alpha), \nu(\alpha); \theta_1, \theta_2)$ be as in the same theorem. Define $\mathcal{I}_\alpha = [\theta_1, \theta_2]$. For each $\theta \in \mathcal{I}_\alpha$ consider the straight line $L_{\alpha,\theta}$ in the (μ, ν)-plane given by*

$$L_{\alpha,\theta} = \{(\mu, \nu) \,:\, \nu = \theta(\mu - \mu(\alpha)) + \nu(\alpha)\}.$$

Then, there is an open segment $\mathcal{I}_{\alpha,\theta} \subset L_{\alpha,\theta} \cap \mathcal{O}_\alpha$ such that

$$\mathcal{I}_{\alpha,\theta} \subset C_r(\mu(\alpha), \nu(\alpha); \theta_1, \theta_2) \quad \textit{and} \quad (\mu(\alpha), \nu(\alpha)) \in \partial \mathcal{I}_{\alpha,\theta}.$$

6. Rovella's theorem. The following definition is similar to Definition 5.6.

DEFINITION 6.1. *Given $a < 0 < b$ let $\varphi : [a, b] \setminus \{0\} \to [a, b]$ be a C^r map, $r \geq 0$. A one-parameter family $\varphi_t : [a, b] \setminus \{0\} \to [a, b]$ of C^r maps converges to φ in the C^r topology as $t \to 0^+$ if for every $\eta > 0$ and $\epsilon > 0$ there is $\delta > 0$ such that*

$$\sup_{i \in \{0,1,\ldots,r\}, x \in [a,b] \setminus (-\eta, \eta)} \left| \varphi_t^{(i)}(x) - \varphi^{(i)}(x) \right| < \epsilon \quad \forall\, 0 < t < \delta.$$

Let φ be a map with domain $Dom(\varphi) \subset \mathbb{R}$. A point $p \in Dom(\varphi)$ is *fixed*, *prefixed*, or *periodic with period* 2 depending on whether $\varphi(p) = p$ or $\varphi(p) \in Dom(\varphi)$ and $\varphi(\varphi(p)) = \varphi(p)$ or $\varphi(p) \in Dom(\varphi)$ and $\varphi(\varphi(p)) = p$.

Now we define a *Rovella map*.

DEFINITION 6.2. *Let $a < 0 < b$ be fixed. A C^1 map $\varphi : [a, b] \setminus \{0\} \to [a, b]$ is a* Rovella map *if it satisfies the following properties.*

1. *φ is monotone in each connected component of $[a, b] \setminus \{0\}$.*
2. *The lateral limits*

$$\varphi(0+) = \lim_{x \to 0^+} \varphi(x), \qquad \varphi(0-) = \lim_{x \to 0^-} \varphi(x)$$

   exist.
3. *The points a and b are fixed, prefixed, or periodic with period 2 and*

$$\varphi(0+), \varphi(0-) \in \{a, b\}.$$

4. *There is $\lambda > 1$ such that for every $\epsilon > 0$ there is $c > 0$ such that if $x \in [a, b]$, $n \in \mathbb{N}$, and $|\varphi^j(x)| \geq \epsilon$ for all integers $0 \leq j \leq n - 1$, then*

$$|(\varphi^n)'(x)| \geq c\lambda^n.$$

As usual we denote $\omega_\varphi(x)$ the omega-limit of the point x with respect to a map φ. The name "Rovella map" above comes from the following result [Rov]. We use the notation

$$CH\{x, y, z, t\}$$

for the convex hull of $x, y, z, t \in \mathbb{R}$.

THEOREM 6.3 (Rovella's theorem). *Let a, a_0, b_0, b be real numbers such that*

$$a < a_0 < 0 < b_0 < b.$$

Let $\varphi : [a, b] \setminus \{0\} \to \mathbb{R}$ be a C^2 map such that $\varphi([a_0, b_0] \setminus \{0\}) \subset [a_0, b_0]$ and $\varphi/_{[a_0,b_0]}$ is a Rovella map. Let $\varphi_t : [a, b] \setminus \{0\} \to \mathbb{R}$, $t \in (0, d]$, be a one-parameter family of C^2 maps satisfying the properties below:

1. *$\varphi_t \to \varphi$ in the C^2 topology as $t \to 0^+$.*
2. *The lateral limits*

$$\varphi_t(0+) = \lim_{x \to 0^+} \varphi_t(x), \qquad \varphi_t(0-) = \lim_{x \to 0^-} \varphi_t(x)$$

exist and satisfy

$$\frac{\partial \varphi_t(0+)}{\partial t}(0) \neq 0, \qquad \frac{\partial \varphi_t(0-)}{\partial t}(0) \neq 0,$$

and

$$\omega_{\varphi_t}(\varphi_t(0+)) \cup \omega_{\varphi_t}(\varphi_t(0-)) \subset [a, b].$$

3. *There are $\alpha > 1$, $t_0 > 0$, and $0 < K_1 < K_2$ such that the following hold:*
   (a) *If $x \in [a, b] \setminus \{0\}$ and $t \in (0, t_0]$, then*

$$K_1 \mid x \mid^{\alpha-1} \leq \left| \frac{\partial}{\partial x} \varphi_t(x) \right| \leq K_2 \mid x \mid^{\alpha-1} . \tag{6.1}$$

   (b) *If $x \in [a, b] \setminus \{0\}$ and $t \in (0, t_0]$, then*

$$K_1 \mid x \mid^{\alpha-2} \leq \left| \frac{\partial^2}{\partial x^2} \varphi_t(x) \right| \leq K_2 \mid x \mid^{\alpha-2} . \tag{6.2}$$

Then

$$\lim_{\delta \to 0^+} \frac{m(C \cap (0, \delta])}{\delta} = 1,$$

where C is the set of $t \in (0, d]$ such that the interval

$$I_t = CH\{\varphi_t(0+), \varphi_t(0-), \varphi_t(\varphi_t(0+)), \varphi_t(\varphi_t(0-))\}$$

contains 0 in its interior and is φ_t-invariant and the map $\varphi_t/_{I_t}$ is transitive. Moreover there is a compact interval W_t with $I_t \subset \mathrm{Int}(W_t)$ such that $\mathrm{Cl}(\varphi_t(W_t \setminus \{0\})) \subset \mathrm{Int}(W_t)$ and $\varphi_t^2(W_t \setminus \{0\}) \subset I_t$.

7. The map h_α. Let $\Lambda > 0$ and $O \subset (1, 1 + \Lambda)$ be as in Lemma 3.1. Let $p(\alpha)$ and $q(\alpha)$ be as in Lemma 3.1. If $\alpha \in O$, we consider $x(\alpha)$ and $y(\alpha)$ as in (5.4), namely,

$$x(\alpha) = \frac{p(\alpha)}{q(\alpha)} \quad \text{and} \quad y(\alpha) = 1.$$

Since $p < 0 < q$ by Lemma 3.1 we have $x(\alpha) < 0 < y(\alpha)$. Let $x(1)$ and $y(1)$ be the limits in Lemma 5.1. Define

$$c_1 = 1 + \max\{-x(1), y(1)\}.$$

By Lemma 5.1 one has

$$c_1 > \max\{1, C^+/C^-\}.$$

Then, by Lemma 5.7, we can fix $\alpha_0 > 1$, $0 < K_1 < K_2$ such that the conclusions (a)–(c) of that lemma hold (with $c = c_1$).

Given $\alpha \in O \cap (1, \alpha_0)$ let $(\mu(\alpha), \nu(\alpha))$ and $g_\alpha(\mu, \nu, x)$ be as in (4.1) and (5.3), respectively.

Define

$$h_\alpha(x) = g_\alpha(\mu(\alpha), \nu(\alpha), x).$$

The first result of this section is the following.

THEOREM 7.1. *There is $\alpha_* > 1$ such that if $1 < \alpha < \alpha_*$, then*

1. *$h_\alpha([x(\alpha), y(\alpha)] \setminus \{0\}) \subset [x(\alpha), y(\alpha)]$, and*
2. *$h_\alpha/_{[x(\alpha), y(\alpha)]\setminus\{0\}}$ is a Rovella map.*

Proof. We shall work only with the cases $\nu^+ = 1$ and $\nu^- = 1$; the remaining cases are similar.

Recall the definition of $x(\alpha)$ and $y(\alpha) = 1$ in the beginning of this section. Let $1 < \alpha < \alpha_0$ and $\delta > 0$ be small (less than $1/2 \cdot \min\{-x(1), 1\}$, say). Define the numbers

$$M_\delta^+(\alpha) = \sup\left\{\left|\frac{\partial}{\partial x} g_\alpha(\mu(\alpha), \nu(\alpha), x)\right| : |x - y(\alpha)| \le \delta\right\},$$

$$m_\delta^+(\alpha) = \inf\left\{\left|\frac{\partial}{\partial x} g_\alpha(\mu(\alpha), \nu(\alpha), x)\right| : |x - y(\alpha)| \le \delta\right\},$$

$$M_\delta^-(\alpha) = \sup\left\{\left|\frac{\partial}{\partial x} g_\alpha(\mu(\alpha), \nu(\alpha), x)\right| : |x - x(\alpha)| \le \delta\right\},$$

and

$$m_\delta^-(\alpha) = \inf\left\{\left|\frac{\partial}{\partial x} g_\alpha(\mu(\alpha), \nu(\alpha), x)\right| : |x - x(\alpha)| \le \delta\right\}.$$

Let g be as in Lemma 5.7. Hence g is piecewise linear with slope C^-B (for $x < 0$) and C^+B (for $x > 0$). By Lemma 5.7(ii) we have that $g_\alpha \to g$ in the C^1 topology in compact sets of $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ as $\alpha \to 1$, $\alpha \in O$. This implies

$$\lim_{(\alpha,\delta)\to(1^+,0)} M_\delta^-(\alpha) = \lim_{(\alpha,\delta)\to(1^+,0)} m_\delta^-(\alpha) = C^-B$$

and

$$\lim_{(\alpha,\delta)\to(1^+,0)} M_\delta^+(\alpha) = \lim_{(\alpha,\delta)\to(1^+,0)} m_\delta^+(\alpha) = C^+B$$

because of the above-mentioned slopes of g. These limits imply

$$\lim_{(\alpha,\delta)\to(1^+,0)} \frac{m_\delta^-(\alpha)}{\left(M_\delta^-(\alpha)\right)^{\frac{\alpha-1}{\alpha}}} = C^-B$$

and

$$\lim_{(\alpha,\delta)\to(1^+,0)} \frac{m_\delta^+(\alpha)}{\left(M_\delta^+(\alpha)\right)^{\frac{\alpha-1}{\alpha}}} = C^+B$$

because $C^\pm B > 0$.

Define

$$\lambda_0 = \min\{C^+B, C^-B\}.$$

The definition of $B = B(X_{\eta_0})$ implies $\lambda_0 > 1$. Then we can fix $1 < \lambda_1 < \lambda_0$. The last limits imply that there are $\delta_1 > 0$ and $1 < \alpha_1 < \alpha_0$ such that

$$\min\left\{\frac{m_\delta^-(\alpha)}{\left(M_\delta^-(\alpha)\right)^{\frac{\alpha-1}{\alpha}}}, \frac{m_\delta^+(\alpha)}{\left(M_\delta^+(\alpha)\right)^{\frac{\alpha-1}{\alpha}}}\right\} > \lambda_1 \tag{7.1}$$

for all $0 < \delta < \delta_1$ and $1 < \alpha < \alpha_1$.

For fixed $\epsilon > 0$, $\delta_0 > 0$, and $\alpha > 1$ we define

$$l_\epsilon^-(x,\alpha) = \min\{l \geq 1 : |h_\alpha^l(x) - x(\alpha)| \geq \delta_0,\, x \in (0,\epsilon)\}$$

and

$$l_\epsilon^+(x,\alpha) = \min\{l \geq 1 : |h_\alpha^l(x) - y(\alpha)| \geq \delta_0,\, x \in (0,\epsilon)\}.$$

Because $h_\alpha(0+) = x(\alpha)$, $h_\alpha(0-) = y(\alpha)$, $h_\alpha(x(\alpha)) = x(\alpha)$, and $h_\alpha(y(\alpha)) = y(\alpha)$ one gets

$$\lim_{(\epsilon,\alpha)\to(0^+,1^+)} \inf_{x\in(0,\epsilon]} \{l_\epsilon^-(x,\alpha)\} = \infty \quad \text{and} \quad \lim_{(\epsilon,\alpha)\to(0^-,1^+)} \inf_{x\in[-\epsilon,0)} \{l_\epsilon^+(x,\alpha)\} = \infty.$$

Fix $1 < \lambda_2 < \lambda_1$ and define

$$N = \frac{\log(\frac{K_1}{\lambda_1}) + \left(\frac{\alpha-1}{\alpha}\right)\cdot\log\left(\frac{\alpha\cdot\delta_0}{K_2}\right)}{-\log\lambda_2}.$$

By the above limits we can fix $\epsilon_0 > 0$ and $1 < \alpha_2 < \alpha_1$ such that

$$l_\epsilon^-(x,\alpha) > N \quad \forall\,(\epsilon,x,\alpha) \in (0,\epsilon_0]\times(0,\epsilon]\times(1,\alpha_2] \tag{7.2}$$

and

$$l_\epsilon^+(x,\alpha) > N \quad \forall\,(\epsilon,x,\alpha) \in (0,\epsilon_0]\times[-\epsilon,0)\times(1,\alpha_2]. \tag{7.3}$$

Hereafter we fix $0 < \delta_0 < \delta_1$ and use the notation

$$m^\pm = m_{\delta_0}^\pm(\alpha), \quad M^\pm = M_{\delta_0}^\pm(\alpha), \quad l^\pm(x) = l_\epsilon^\pm(x,\alpha)$$

for the sake of simplicity.

By Lemma 5.7(ii) we have that $g_\alpha \to g$ in the C^1 topology as $\alpha \to 1^+$. More precisely, the following property holds (see Definition 5.5): For every compact set

$K \subset \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ and every $\varepsilon > 0$ there is $1 < \alpha_3 < \alpha_2$ such that if $1 < \alpha < \alpha_3$, then

$$\sup_{i\in\{0,1\},y\in K} \left|D^i g_\alpha(y) - D^i g(y)\right| < \varepsilon.$$

Applying this property to $\varepsilon = \epsilon_0$ and $K = [-c_1, c_1]^2 \times ([-c_1, c_1] \setminus (-\epsilon_0, \epsilon_0))$ we obtain

$$\left|\frac{\partial}{\partial x} g_\alpha(\mu(\alpha), \nu(\alpha), x)\right| \geq \lambda_1 \quad \forall |x| \geq \epsilon_0. \tag{7.4}$$

(For this we use $(\mu(\alpha), \nu(\alpha), x) \in K$ (as $\alpha \to 1^+$), $\lambda_1 < \lambda_0$, and λ_0 is about the slope of the piecewise linear maps $x \to g(\mu(\alpha), \nu(\alpha), x)$.)

Define

$$\lambda = \frac{\lambda_1}{\lambda_2}.$$

Then, $\lambda > 1$. We use λ in the following lemma.

LEMMA 7.2. *Let $0 < \epsilon < \epsilon_0$ and $1 < \alpha < \alpha_3$ be fixed. If $\epsilon \leq x < \epsilon_0$, then*

$$|(h_\alpha^{l^-(x)})'(x)| \geq \lambda^{l^-(x)}$$

and if $-\epsilon_0 < x \leq -\epsilon$, then

$$|(h_\alpha^{l^+(x)})'(x)| \geq \lambda^{l^+(x)}.$$

Proof. We prove only the first inequality since the second one is analogous. Fix $0 < \epsilon < \epsilon_0$, $1 < \alpha < \alpha_3$, and $\epsilon \leq x < \epsilon_0$. The chain rule and the definitions of m^- and h_α imply

$$|(h_\alpha^{l^-(x)-1})'(h_\alpha(x))| = \Pi_{i=1}^{l^-(x)-1} \left|\frac{\partial}{\partial x} g_\alpha(\mu(\alpha), \nu(\alpha), h_\alpha^i(x))\right| \geq (m^-)^{l^-(x)-1}.$$

As above we have $(\mu(\alpha), \nu(\alpha), x) \in [-c_1, c_1]^2 \times ([-c_1, c_1] \setminus \{0\})$ because $\alpha \to 1^+$. The chain rule, (5.5) in Lemma 5.7(b), and the definitions of m^- and $l^-(x)$ imply

$$|(h_\alpha^{l^-(x)})'(x)| = |h_\alpha'(x)| \cdot |(h_\alpha^{l^-(x)-1})'(h_\alpha(x))| \geq K_1 \cdot |x|^{\alpha-1} \cdot (m^-)^{l^-(x)-1}. \tag{7.5}$$

On the other hand, Lemma 3.1(a) and the definition of g_α imply

$$h_\alpha(x(\alpha)) = g_\alpha(\mu(\alpha), \nu(\alpha), x(\alpha)) = x(\alpha).$$

So, $|h_\alpha^{l^-(x)-1}(x(\alpha)) - h_\alpha^{l^-(x)-1}(h_\alpha(x))| = |x(\alpha) - h_\alpha^{l^-(x)}(x)|$. Then, the definition of $l^-(x)$ implies

$$|h_\alpha^{l^-(x)-1}(x(\alpha)) - h_\alpha^{l^-(x)-1}(h_\alpha(x))| \geq \delta_0.$$

Consequently, for some intermediate point ξ in between $x(\alpha)$ and $h_\alpha(x)$ we obtain

$$\begin{aligned} |x(\alpha) - h_\alpha^{l^-(x)}(x)| &= |h_\alpha^{l^-(x)-1}(x(\alpha)) - h_\alpha^{l^-(x)-1}(h_\alpha(x))| \\ &= |(h_\alpha^{l^-(x)-1})'(\xi)| \cdot |x(\alpha) - h_\alpha(x)|. \end{aligned}$$

However, $|(h_\alpha^{l^-(x)-1})'(\xi)| \le (M^-)^{l^-(x)-1}$ because $|x(\alpha)-\xi| \le |x(\alpha)-h_\alpha(x)|$ and so

$$|x(\alpha)-h_\alpha^i(\xi)| \le |x(\alpha)-h_\alpha^{i+1}(x)| \le \delta_0$$

for all $0 \le i \le l^-(x)-2$ because h_α is monotone in $[x(\alpha),0)$. Then,

$$|(h_\alpha^{l^-(x)-1})'(\xi)| \cdot |x(\alpha)-h_\alpha(x)| \le (M^-)^{l^-(x)-1} \cdot |x(\alpha)-h_\alpha(x)|$$

and so

$$|x(\alpha)-h_\alpha^{l^-(x)}(x)| \le (M^-)^{l^-(x)-1} \cdot |x(\alpha)-h_\alpha(x)|.$$

As $|x(\alpha)-h_\alpha^{l^-(x)}(x)| \ge \delta_0$ by the definition of $l^-(x)$ we obtain

$$|x(\alpha)-h_\alpha(x)| \cdot (M^-)^{l^-(x)-1} \ge \delta_0$$

and so

$$|x(\alpha)-h_\alpha(x)| \ge (M^-)^{1-l^-(x)} \cdot \delta_0.$$

On the other hand,

$$|h_\alpha(x)-x(\alpha)| = \left|\int_0^x h'_\alpha(\tau)d\tau\right| \le \int_0^x K_2 \cdot \tau^{\alpha-1}d\tau = \frac{K_2}{\alpha}|x|^\alpha,$$

where K_2 comes from inequality (5.5) in Lemma 5.7(b). With this we obtain

$$|x|^{\alpha-1} \ge \left(\frac{\alpha\cdot\delta_0}{K_2}\cdot(M^-)^{1-l^-(x)}\right)^{\frac{\alpha-1}{\alpha}}. \tag{7.6}$$

Combining (7.5) and (7.6) we obtain

$$|(h_\alpha^{l^-(x)})'(x)| \ge K_1 \cdot \left(\frac{\alpha\cdot\delta_0}{K_2}\right)^{\frac{\alpha-1}{\alpha}} \cdot \left(\frac{m^-}{(M^-)^{\frac{\alpha-1}{\alpha}}}\right)^{l^-(x)-1} \ge K_1 \cdot \left(\frac{\alpha\cdot\delta_0}{K_2}\right)^{\frac{\alpha-1}{\alpha}} \cdot \lambda_1^{l^-(x)-1}$$

because $x \in (0,\epsilon_0)$ and inequality (7.1) holds. Finally using inequality (7.2) we obtain

$$|(h_\alpha^{l^-(x)})'(x)| \ge \lambda^{l^-(x)}.$$

The proof follows. □

Choose $\alpha_3 > 1$ as in Lemma 7.2 and define

$$\alpha_* = \alpha_3.$$

We shall prove that α_* satisfies 1 and 2 of Theorem 7.1. For this we fix $1 < \alpha < \alpha_*$.

Proof of item 1 *of Theorem* 7.1. By item 1 of Lemma 5.2 we have

$$h_\alpha(x(\alpha)) = x(\alpha) \quad \text{and} \quad h_\alpha(y(\alpha)) = y(\alpha).$$

By (5.3) we have

$$h_\alpha(0+) = g_\alpha(\mu(\alpha),\nu(\alpha),0+) = \frac{f_\eta(0+)}{q(\alpha)} = \frac{a^+(\alpha)}{q(\alpha)} = \mu(\alpha)$$

and analogously $h_\alpha(0-) = \nu(\alpha)$. Moreover, h_α is monotone in $[x(\alpha), 0)$ and $(0, y(\alpha)]$ by Lemma 2.1 and the definition of $g_\alpha(\mu, \nu, x)$. Then the result follows.

Proof of item 2 *of Theorem* 7.1. Items 1 and 2 of Definition 6.2 follow from the previous calculations. Item 3 of Definition 6.2 follows from item 1 of Lemma 5.2 and item 1 of Lemma 5.3.

Now we prove item 4 of Definition 6.2. Recall $\lambda = \frac{\lambda_1}{\lambda_2}$ (thus $\lambda > 1$) and $1 < \alpha < \alpha_*$. Fix $\epsilon > 0$. We have to find $c > 0$ such that if for all $x \in [x(\alpha), y(\alpha)]$, $n \in \mathbb{N}$ and $|h_\alpha^j(x)| \geq \epsilon$ for all integers $0 \leq j \leq n-1$, then

$$|(h_\alpha^n)'(x)| \geq c\lambda^n. \tag{7.7}$$

If $\epsilon \geq \epsilon_0$, then (7.7) holds by inequality (7.4) choosing $c = 1$. Then, we can assume

$$0 < \epsilon < \epsilon_0.$$

For simplicity we write h instead of h_α.

Define

$$c = \inf_{|y| \geq \epsilon} \{|h'(y)|\}.$$

Then, $c > 0$ because of the choice of K_1.

Next we break the orbit $\{x, h(x), \ldots, h^{n-1}(x)\}$ of x in the following way:

$$0 \leq r_0 < r_1 < \cdots < r_s \leq n-1,$$

such that

$$\epsilon \leq |h^{r_i}(x)| < \epsilon_0$$

for all $0 \leq i \leq s$. For all i we define the integers

$$l_i = l^*(h^{r_i}(x)),$$

where $*$ is $+$ or $-$ depending on the signal of $h^{r_i}(x)$. The chain rule implies

$$|(h^n)'(x)| = |(h^{r_0})'(x)| \cdot (\Pi_{i=1}^s |(h^{r_i})'(h^{r_{i-1}}(x))|) \cdot |(h^{n-r_s})'(h^{r_s}(x))|. \tag{7.8}$$

This product consists of three terms to be bounded in what follows. The first one is bounded by

$$|(h^{r_0})'(x)| > \lambda^{r_0} \tag{7.9}$$

because of (7.4) since $|h^j(x)| \geq \epsilon_0$ for $0 \leq j < r_0$. The last one is bounded by

$$|(h^{n-r_s})'(h^{r_s}(x))| \geq c \cdot \lambda^{n-r_s-1} \tag{7.10}$$

by the definition of c because $|h^{r_s}(x)| \geq \epsilon$ and $|h^j(x)| \geq \epsilon_0$ for $r_s + 1 \leq j < n$.

Next we bound the middle term $\Pi_{i=1}^s |(h^{r_i})'(h^{r_{i-1}}(x))|$. For all $1 \leq i \leq s$ we have that

$$|(h^{r_i})'(h^{r_{i-1}}(x))| = |(h^{l_{i-1}})'(h^{r_{i-1}}(x))| \cdot |(h^{r_i - l_{i-1}})'(h^{r_i + l_i}(x))|.$$

On the other hand

$$|(h^{l_{i-1}})'(h^{r_{i-1}}(x))| \geq \lambda^{l_{i-1}}$$

by Lemma 7.2 applied to $h^{r_{i-1}}(x)$ (instead of x). And also

$$|(h^{r_i-l_{i-1}})'(h^{r_i+l_i}(x))| \geq \lambda^{r_i+l_{i-1}}$$

by inequality (7.2) because $|h^j(x)| \geq \epsilon_0$ for all $r_{i-1}+1 \leq j < r_i$. All together yield

$$|(h^{r_i})'(h^{r_{i-1}}(x))| \geq \lambda^{l_{i-1}} \cdot \lambda^{r_i+l_{i-1}} = \lambda^{r_i}. \tag{7.11}$$

Replacing (7.11), (7.10), and (7.9) in (7.8) we obtain (7.7). This finishes the proof of Theorem 7.1. □

The second result of this section is the following.

THEOREM 7.3. *Let $(\mu(1),\nu(1))$, V, $\mathcal{O}$, α_1, and O be as in Corollary 5.9. Fix $\alpha \in O \cap (1,\alpha_1)$ and consider $\mathcal{I}_\alpha$ as in that corollary. For each $\theta \in \mathcal{I}_\alpha$ consider the open segment $\mathcal{I}_{\alpha,\theta} \subset L_{\alpha,\theta} \cap \mathcal{O}_\alpha$ as in that corollary. Let $s : [0, d_{\alpha,\theta}] \subset \mathbb{R} \to \mathcal{I}_{\alpha,\theta}$ be the parametrization of $\mathcal{I}_{\alpha,\theta}$ given by*

$$s(t) = t \cdot v_\theta + (\mu(\alpha),\nu(\alpha)),$$

where $v_\theta = (\cos(\theta),\sin(\theta))$ is the unitary vector with argument θ. Define the numbers

$$a = -3\max\{1, C^+/C^-\}, \qquad b = 3\max\{1, C^+/C^-\},$$

and the map $\varphi : [a,b] \setminus \{0\} \to \mathbb{R}$ by

$$\varphi(x) = g_\alpha(\mu(\alpha),\nu(\alpha),x).$$

There is $\alpha_{**} > 1$ such that if $1 < \alpha < \alpha_{**}$ and $\theta \in \mathcal{I}_\alpha \setminus \{0,\pi/2,\pi,3\pi/2,2\pi\}$, then φ and the family $\varphi_t : [a,b] \setminus \{0\} \to \mathbb{R}$, $t \in (0, d_{\alpha,\theta}]$ defined by

$$\varphi_t(x) = g_\alpha(s(t),x)$$

satisfy the hypotheses (1)–(3) of Rovella's theorem (Theorem 6.3).

Proof. Let $x(\alpha)$ and $y(\alpha)$ be as in (5.4), let $\alpha_* > 1$ be as in Theorem 7.1, and define

$$a_0 = x(\alpha) \quad \text{and} \quad b_0 = y(\alpha).$$

The definition of $x(\alpha), y(\alpha)$ and Lemma 5.1 imply

$$\lim_{\alpha\to 1^+} x(\alpha) = -C^+/C^-,$$

and obviously

$$\lim_{\alpha\to 1^+} y(\alpha) = 1.$$

So, there is $1 < \alpha_{**} < \alpha_*$ close to 1 such that if $1 < \alpha < \alpha_{**}$, then

$$a < x(\alpha) - 1 < a_0 < 0 < b_0 < y(\alpha) + 1 < b. \tag{7.12}$$

Let us prove that such an α_{**} works. Hereafter we fix

$$1 < \alpha < \alpha_{**} \quad \text{and} \quad \theta \in \mathcal{I}_\alpha \setminus \{0,\pi/2,\pi,3\pi/2,2\pi\}.$$

By Theorem 7.1 we have $\varphi([a_0,b_0] \setminus \{0\}) \subset [a_0,b_0]$ and $\varphi/_{[a_0,b_0]}$ is a Rovella map.

Let us verify hypothesis 1 of Rovella's theorem. By (5.3) one has

$$\varphi_t(x) = g_\alpha(s(t), x) = \frac{1}{q(\alpha)} \cdot f_{\eta(\alpha, q(\alpha)\cdot s(t))}(q(\alpha)\cdot x).$$

As $t \to 0^+$ we have

$$\varphi_t(x) \to \frac{1}{q(\alpha)} \cdot f_{\eta(\alpha, q(\alpha)\cdot s(0))}(q(\alpha)\cdot x) = \varphi(x)$$

because of (5.3) and the definition of φ. The C^2 convergence $\varphi_t \to \varphi$ required in hypothesis 1 follows from the C^2 variation mentioned in Lemma 2.1.

Next we verify hypothesis 2 of Rovella's theorem. Recalling the definition of v_θ we have

$$s(t) = (\mu_t(\alpha), \nu_t(\alpha)),$$

where

$$\begin{cases} \mu_t(\alpha) = t\cdot\cos\theta + \mu(\alpha), \\ \nu_t(\alpha) = t\cdot\sin\theta + \mu(\alpha). \end{cases}$$

By Lemma 2.1 we have

$$f_\eta(0+) = a_\eta^+ \quad \text{and} \quad f_\eta(0-) = a_\eta^-$$

for all η, where $a_\eta^\pm$ are the leading terms in the Taylor expansion of f_η in Lemma 2.1. Then,

$$f_{\eta(\alpha, q(\alpha)\cdot s(t))}(0+) = a^+_{\eta(\alpha, q(\alpha)\cdot s(t))} = q(\alpha)\cdot\mu_t(\alpha)$$

by (3.2) applied to $a^+ = q(\alpha)\cdot\mu_t(\alpha)$ and $a^- = q(\alpha)\cdot\nu_t(\alpha)$. Analogously

$$f_{\eta(\alpha, q(\alpha)\cdot s(t))}(0-) = q(\alpha)\cdot\nu_t(\alpha).$$

Thus,

$$\begin{aligned} \lim_{x\to 0^+} \varphi_t(x) &= \lim_{x\to 0^+} g_\alpha(s(t), x) = \lim_{x\to 0^+} \frac{1}{q(\alpha)} \cdot f_{\eta(\alpha, q(\alpha)\cdot s(t))}(q(\alpha)\cdot x) \\ &= \frac{1}{q(\alpha)} \cdot f_{\eta(\alpha, q(\alpha)\cdot s(t))}(0+) = \frac{1}{q(\alpha)} \cdot q(\alpha)\cdot\mu_t(\alpha) = \mu_t(\alpha), \end{aligned}$$

and so,

$$\varphi_t(0+) = \mu_t(\alpha).$$

Analogously

$$\varphi_t(0-) = \nu_t(\alpha).$$

This proves the existence of the lateral limits in hypothesis 2. These equalities also imply

$$\frac{\partial\varphi_t(0+)}{\partial t}(0) = \frac{\partial\mu_t(\alpha)}{\partial t}(0) = \frac{\partial(t\cdot\cos\theta + \mu(\alpha))}{\partial t}(0) = \cos\theta \neq 0$$

because $\theta \neq \pi/2, 3\pi/2$. Analogously

$$\frac{\partial \varphi_t(0+)}{\partial t}(0) \neq 0$$

since $\theta \neq 0, \pi$ (for in such cases $\sin\theta \neq 0$). These facts together imply

$$\frac{\partial \varphi_t(0+)}{\partial t}(0) \neq 0, \qquad \frac{\partial \varphi_t(0-)}{\partial t}(0) \neq 0.$$

To see

$$\omega_{\varphi_t}(\varphi_t(0+)) \cup \omega_{\varphi_t}(\varphi_t(0-)) \subset [a, b]$$

we proceed as follows. By Corollary 5.9 we have $\mathcal{I}_{\alpha,\theta} \subset C_r(s(t); \theta_1, \theta_2)$. In particular, $s(t) \in C_r(s(t); \theta_1, \theta_2)$ for all $t \in (0, d_{\alpha,\theta}]$, and so, by the second part of Theorem 5.8, there is a closed interval $I_{(\alpha,s(t))} \subset [x(\alpha)-1, y(\alpha)+1]$ with $0 \in \mathrm{Int}(I_{(\alpha,s(t))})$ such that

$$\varphi_t(x) = g_\alpha(s(t), x) \subset \mathrm{Int}(I_{(\alpha,s(t))}) \quad \forall\, x \in I_{(\alpha,s(t))}.$$

This already implies $\omega_{\varphi_t}(\varphi_t(0+)) \cup \omega_{\varphi_t}(\varphi_t(0-)) \subset I_{(\alpha,s(t))}$ and then we are done because (7.12) implies $[x(\alpha) - 1, y(\alpha) + 1] \subset [a, b]$.

To finish we prove hypothesis 3 of Rovella's theorem. Shrinking α_{**} if necessary we can assume $1 < \alpha_{**} < \alpha_0$, where α_0 comes from Lemma 5.7. Choosing $c = -a$ (or b) we have $c > \max\{1, C^+/C^-\}$. Then, the result follows from (a), inequality (5.5) in (b), and inequality (5.6) in (c) of Lemma 5.7. □

We finish this section with the following corollary. Recall that $CH\{x, y, z, t\}$ denote the convex hull of $x, y, z, t \in \mathbb{R}$.

COROLLARY 7.4. *Let O be as in Lemma* 3.1 *and $(\mu(1), \nu(1))$ be the limit in Lemma* 4.1. *Let V be a neighborhood of $(\mu(1), \nu(1))$. Let $(\mu(\alpha), \nu(\alpha))$ be as in* (4.1). *There is $\alpha_{***} > 1$ such that for all $\alpha \in O \cap (1, \alpha_{***})$ there is a positive Lebesgue measure subset $\tilde{E}_\alpha \subset V$ with $(\mu(\alpha), \nu(\alpha)) \in \mathrm{Cl}(\tilde{E}_\alpha)$ such that the following property holds: If $(\mu, \nu) \in \tilde{E}_\alpha$, then the interval*

$$I = CH\{g_\alpha(\mu, \nu, 0+), g_\alpha(\mu, \nu, 0-), g_\alpha(\mu, \nu, g_\alpha(\mu, \nu, 0+)), g_\alpha(\mu, \nu, g_\alpha(\mu, \nu, 0-))\}$$

contains 0 in its interior and is $g_\alpha(\mu, \nu, \cdot)$-invariant and the map $g_\alpha(\mu, \nu, \cdot)/_I$ is transitive. Moreover there is a compact interval W with $\mathrm{Int}(W) \supset I$ such that $\mathrm{Cl}(g_\alpha(W \setminus \{0\})) \subset W$ and $g_\alpha^2(W \setminus \{0\}) \subset I$.

Proof. Let $\mathcal{O}$ as in Theorem 5.8 and $\alpha_{**} > 1$ be as in Theorem 7.3.

We claim that $\alpha_{***} = \alpha_{**}$ works. Indeed, by Theorem 7.3, for each $\theta \in \mathcal{I}_\alpha \setminus \{o, \pi/2, \pi, 3\pi/2, 2\pi\}$ the family $\varphi_t : [a, b] \setminus \{0\} \to \mathbb{R}$, $t \in (0, d_{\alpha,\theta}]$, defined by

$$\varphi_t(x) = g_\alpha(s(t), x)$$

satisfies (1)–(3) of Rovella's theorem. So, for those θ's we have

$$\lim_{\delta \to 0^+} \frac{m(C(\alpha, \theta) \cap [0, \delta])}{\delta} = 1, \tag{7.13}$$

where $C(\alpha, \theta)$ is the set of $t \in (0, d_{\alpha,\theta}]$ such that the interval

$$I_t = CH\{\varphi_t(0+), \varphi_t(0-), \varphi_t(\varphi_t(0+)), \varphi_t(\varphi_t(0-))\}$$

contains 0 in its interior and is φ_t-invariant and the map $\varphi_t/_{I_t}$ is transitive. Moreover there is a compact interval W_t with $I_t \subset \mathrm{Int}(W_t)$ such that $\mathrm{Cl}(\varphi_t(W_t \setminus \{0\})) \subset \mathrm{Int}(W_t)$ and $\varphi_t^2(W_t \setminus \{0\}) \subset I_t$. Observe that $0 \in \mathrm{Cl}(C(\alpha,\theta))$ because of (7.13). Again by (7.13) we have that $C(\alpha,\theta)$ has positive Lebesgue measure in $\mathbb{R}$.

Define

$$\tilde{E}_{\alpha,\theta} = s(C(\alpha,\theta)).$$

Then, $\tilde{E}_{\alpha,\theta}$ has positive Lebesgue measure in $\mathcal{I}_{\alpha,\theta}$.

Let us prove $\tilde{E}_{\alpha,\theta} \subset V$. Indeed, Corollary 5.9 implies $\mathcal{I}_{\alpha,\theta} \subset C_r(\mu(\alpha),\nu(\alpha);\theta_1,\theta_2)$ where the last cone comes from Theorem 5.8. By the first part of this theorem we have $C_r(\mu(\alpha),\nu(\alpha);\theta_1,\theta_2) \subset \mathcal{O}_\alpha$ which in turn is contained in V. This proves the result.

Now define

$$\tilde{E}_\alpha = \bigcup_{\theta \in \mathcal{I}_\alpha \setminus \{0,\pi/2,\pi,3\pi/2,2\pi\}} \tilde{E}_{\alpha,\theta}.$$

As $\tilde{E}_{\alpha,\theta} \subset V$ for all α,θ we have

$$\tilde{E}_\alpha \subset V.$$

We also have that $\tilde{E}_\alpha$ has positive Lebesgue measure in $\mathbb{R}^2$ by Fubini's theorem since $\tilde{E}_{\alpha,\theta}$ has positive Lebesgue measure in $\mathbb{R}$. We have $(\mu(\alpha),\nu(\alpha)) \in \mathrm{Cl}(\tilde{E}_{\alpha,\theta})$ because of $0 \in \mathrm{Cl}(C(\alpha,\theta))$. Now, the proof follows from the definition of φ_t. □

8. Proof of Theorem 2.2. Let $\{X_\eta\}$ be a C^k three-parametrized family of C^r-vector fields transverse at X_{η_0} to the submanifold $\mathcal{N}$ given in Definition 1.1, $k, r \geq 3$. Hence X_{η_0} satisfies (A1)–(A6) in that definition. Let $f_\eta : J \subset [-1,1] \to [-1,1]$ be the corresponding three-parametrized family of one-dimensional maps as in Lemma 2.1. Fix a neighborhood $\mathcal{V}$ of η_0 in the parameter space. We have to prove that there is a positive Lebesgue measure set $\mathcal{L}' \subset \mathcal{V}$ with $\eta_0 \in \mathrm{Cl}(\mathcal{L}')$ such that for every $\eta \in \mathcal{L}'$ one has $\alpha(\eta) > 1$ and further there is a closed nontrivial interval Θ_η with the following properties:

1. $0 \in \mathrm{Int}(\Theta_\eta)$.
2. There is a compact interval $W \subset J$ with $\Theta_\eta \subset \mathrm{Int}(W)$ such that $\mathrm{Cl}(f_\eta(W \setminus \{0\})) \subset \mathrm{Int}(W)$ and $f_\eta^2(W \setminus \{0\}) \subset \Theta_\eta$.
3. $f_\eta/_{\Theta_\eta \setminus \{0\}}$ is transitive.

We proceed as follows. Let η be the diffeomorphism in (3.1). Fix $\Lambda > 0$ and $O \subset (1, 1+\Lambda)$ as in Lemma 3.1. Let $(\mu(1),\nu(1))$ be the limits in Lemma 4.1. Let T_α be as in (5.1). Then, the inverse of T_α is given by

$$T_\alpha^{-1}(\mu,\nu) = q(\alpha) \cdot (\mu,\nu). \tag{8.1}$$

For all neighborhoods V of $(\mu(1),\nu(1))$ and $\beta > 1$ we define

$$\tilde{V}(\beta) = \bigcup_{\alpha \in O \cap (1,\beta)} \{\alpha\} \times T_\alpha^{-1}(V).$$

Lemma 3.1 implies $\lim_{\alpha \to 1^+} q(\alpha) = 0$. Thus, there are V and $\beta > 1$ such that

$$\eta(\tilde{V}(\beta)) \subset \mathcal{V}.$$

In what follows we fix such V and β.

Let $\alpha_{***} > 1$ be as in Corollary 7.4 for such a V. Shrinking α_{***} if necessary we can assume

$$1 < \alpha_{***} < \beta.$$

For all $\alpha \in O \cap (1, \alpha_{***})$ we define

$$E_\alpha = (T_\alpha)^{-1}(\tilde{E}_\alpha),$$

where $\tilde{E}_\alpha \subset V$ comes from Corollary 7.4. It follows that

$$E_\alpha \subset T_\alpha^{-1}(V).$$

Since T_α is linear (and $\tilde{E}_\alpha$ has positive Lebesgue measure by Corollary 7.4) we have that E_α has positive Lebesgue measure for all $\alpha \in O \cap (1, \alpha_{***})$.

Define

$$\mathcal{L}' = \{\eta(\alpha, a^+, a^-) : \alpha \in O \cap (1, \alpha_{***}), (a^+, a^-) \in E_\alpha\}.$$

As $E_\alpha \subset T_\alpha^{-1}(V)$, the definition of $\tilde{V}(\beta)$ implies $\mathcal{L}' \subset \eta(\tilde{V}(\beta))$, and so,

$$\mathcal{L}' \subset \mathcal{V}.$$

We have that $\mathcal{L}'$ has positive Lebesgue measure because E_α has positive Lebesgue measure (for all $\alpha \in O \cap (1, \alpha_{***})$) and O has full Lebesgue measure in $(1, 1 + \Lambda)$.

LEMMA 8.1. *If η_0 is the parameter in Theorem* 1.2*, then $\eta_0 \in \mathrm{Cl}(\mathcal{L})$.*

Proof. Note that $\eta_0 = \eta(1, 0, 0)$, where η is the map given in (3.1). Pick a sequence $\alpha_n \in O$ converging to 1^+. It follows that $\alpha_n \in O \cap (1, \alpha_{***})$ for n large.

We have $E_{\alpha_n} = T_{\alpha_n}^{-1}(\tilde{E}_{\alpha_n})$ by definition.

Recalling (8.1) and (4.1) we get

$$(\mu(\alpha), \nu(\alpha)) = \left(\frac{a^+(\alpha)}{q(\alpha)}, \frac{a^-(\alpha)}{q(\alpha)}\right);$$

see Lemma 3.1. Then, because $(\mu(\alpha_n), \nu(\alpha_n)) \in \mathrm{Cl}(\tilde{E}_{\alpha_n})$ by Corollary 7.4 we have

$$(a^+(\alpha_n), a^-(\alpha_n)) \in \mathrm{Cl}(E_{\alpha_n})$$

by the definition of E_α. Then, there is a sequence $(a_n^+, a_n^-) \in E_{\alpha_n}$ arbitrarily close to $(a^+(\alpha_n), a^-(\alpha_n))$.

On the other hand,

$$(a^+(\alpha_n), a^-(\alpha_n)) \to (0, 0)$$

as $n \to \infty$ because $q(\alpha_n) \to 0$ (see Lemma 3.1) and (8.1) (note that the limit $(\mu(\alpha_n), \nu(\alpha_n)) \to (\mu(1), \nu(1))$ by Lemma 4.1). Then,

$$(a_n^+, a_n^-) \to (0, 0)$$

and so

$$\eta(\alpha_n, a_n^+, a_n^-) \to \eta(1, 0, 0) = \eta_0.$$

Then the result follows since $\eta(\alpha_n, a_n^+, a_n^-) \in \mathcal{L}$ because $(a_n^+, a_n^-) \in E_{\alpha_n}$. This proves the lemma. □

Define

$$\Theta_\eta = CH\{f_\eta(a^+), f_\eta(a^-), a^+, a^-\}.$$

Recalling (5.3) one has

$$g_\alpha(\mu, \nu, x) = \frac{1}{q(\alpha)} f_\eta(q(\alpha)x).$$

Then, the properties (1)–(3) in Theorem 2.2 follow from the conclusion of Corollary 7.4. The proof follows. □

REFERENCES

[GW] J. Guckenheimer and A. Williams, *Structural stability of Lorenz attractors*, Inst. Hautes Études Sci. Publ. Math., 50 (1979), pp. 59–72.

[HPS] M. Hirsch, C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.

[dMP] W. de Melo and J. Palis, *Geometric Theory of Dynamical Systems. An Introduction*, Springer-Verlag, New York, 1982.

[MPS] C. A. Morales, M. J. Pacifico, and B. San Martin, *Expanding Lorenz attractors through resonant double homoclinic loops*, SIAM J. Math. Anal., 36 (2005), pp. 1836–1861.

[Rob1] C. Robinson, *Nonsymmetric Lorenz attractors from a homoclinic bifurcation*, SIAM J. Math. Anal., 32 (2000), pp. 119–141.

[Rob2] C. Robinson, *Homoclinic bifurcation to a transitive attractor of Lorenz type,* II, SIAM J. Math. Anal., 23 (1992), pp. 1255–1268.

[Rob3] C. Robinson, *Dynamical Systems, Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, FL, 1995.

[Rov] A. Rovella, *The dynamics of perturbations of the contracting Lorenz attractor*, Bol. Soc. Brasil. Mat. (N.S.), 24 (1993), pp. 233–259.

[Sh] M. Shub, *Global Stability of Dynamical Systems*, Springer-Verlag, New York, 1987.

SIAM J. MATH. ANAL.
Vol. 38, No. 1, pp. 333–345

AN OPTIMAL EXAMPLE FOR THE BALIAN–LOW UNCERTAINTY PRINCIPLE*

JOHN J. BENEDETTO†, WOJCIECH CZAJA‡, AND ALEXANDER M. POWELL§

Abstract. We analyze the time-frequency concentration of the Gabor orthonormal basis $\mathcal{G}(f,1,1)$ constructed by Høholdt, Jensen, and Justesen. We prove that their window function f has near optimal time and frequency localization with respect to a nonsymmetric version of the Balian–Low theorem. In particular, we show that if $(p,q) = (3/2,3)$, then $\int |t|^{p-\epsilon}|f(t)|^2dt < \infty$ and $\int |\gamma|^{q-\epsilon}|\widehat{f}(\gamma)|^2d\gamma < \infty$, for $0 < \epsilon \leq 3/2$, but that both integrals are infinite if $\epsilon = 0$.

Key words. Gabor analysis, Balian–Low theorem, time-frequency analysis

AMS subject classifications. Primary, 42C15, 42C25; Secondary, 46C15

DOI. 10.1137/050634104

1. Introduction.

Given a square integrable function $g \in L^2(\mathbb{R})$ and constants $a, b > 0$, the associated *Gabor system*, $\mathcal{G}(g,a,b) = \{g_{m,n}\}_{m,n\in\mathbb{Z}}$, is defined by

$$g_{m,n}(t) = e^{2\pi iamt}g(t-bn).$$

Gabor systems are of considerable interest for their ability to give frame decompositions for many function spaces [18], [13], [14], [4]. A collection $\{e_n\}_{n\in\mathbb{Z}} \subseteq L^2(\mathbb{R})$ is a *frame* for $L^2(\mathbb{R})$ if there exist constants $0 < A \leq B < \infty$ such that

$$\forall f \in L^2(\mathbb{R}), \quad A||f||^2_{L^2(\mathbb{R})} \leq \sum_{n\in\mathbb{Z}} |\langle f, e_n\rangle|^2 \leq B||f||^2_{L^2(\mathbb{R})}.$$

If $\mathcal{G}(g,a,b)$ is a frame for $L^2(\mathbb{R})$, we shall refer to it as a *Gabor frame* for $L^2(\mathbb{R})$; if it is an orthonormal basis for $L^2(\mathbb{R})$, we refer to it as a *Gabor orthonormal basis* for $L^2(\mathbb{R})$.

A key property of Gabor systems is the fact that one can construct Gabor frames, $\mathcal{G}(g,a,b)$, for $L^2(\mathbb{R})$ such that the *window function* g has excellent time and frequency localization. For example, if $0 < ab < 1$ and $g(t) = e^{-t^2}$, then $\mathcal{G}(g,a,b)$ is an oversampled Gabor frame for $L^2(\mathbb{R})$; see, e.g., [18, Chapter 7]. Overcompleteness is a very important part of such well-localized constructions and can provide robustness and numerical stability in applied settings. On the other hand, if $g \in L^2(\mathbb{R})$, and $\mathcal{G}(g,a,b)$ is an orthonormal basis for $L^2(\mathbb{R})$, then one must have $ab = 1$; see, e.g., [18, Corollary 7.5.2]. If one wishes to construct Gabor orthonormal bases, i.e., nonredundant frames,

*Received by the editors June 20, 2005; accepted for publication (in revised form) January 10, 2006; published electronically May 5, 2006.
http://www.siam.org/journals/sima/38-1/63410.html

†Department of Mathematics, University of Maryland, College Park, MD 20742 (jjb@math.umd.edu). The research of this author was supported in part by NSF DMS grant 0139759 and ONR grant N000140210398.

‡Institute of Mathematics, University of Wrocław, Pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland. Current address: Department of Mathematics, University of Vienna, Nordbergstrasse 15, 1090 Vienna, Austria (czaja@math.uni.wroc.pl). The research of this author was supported by European Commission grant MEIF-CT-2003-500685.

§Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (alexander.m.powell@vanderbilt.edu). The research of this author was supported in part by NSF DMS grants 0139759, 0504924, and 0219233 and by an Erwin Schrödinger Institute (ESI) Junior Research Fellowship.

then there are severe restrictions on the window function's time and frequency localization. The Balian–Low theorem makes this precise. We use the Fourier transform defined by $\widehat{g}(\gamma) = \int g(t)e^{-2\pi i\gamma t}dt$, where our convention is that the integral without specific limits denotes the integral over $\mathbb{R}$.

THEOREM 1.1 (Balian–Low). *Let $g \in L^2(\mathbb{R})$. If*

$$\int |t|^2|g(t)|^2dt < \infty \quad \textit{and} \quad \int |\gamma|^2|\widehat{g}(\gamma)|^2d\gamma < \infty,$$

then $\mathcal{G}(g,1,1)$ is not an orthonormal basis for $L^2(\mathbb{R})$.

The Balian–Low theorem has undergone numerous extensions and generalizations since the early references [2], [23], [3], [8]. For example, it holds in higher dimensions for rather general time-frequency lattices, and also holds if one replaces "orthonormal basis" with "Riesz basis." For recent work related to the Balian–Low theorem, see [1], [5], [6], [7], [9], [10], [16], [19], [11]. The issue of sharpness or optimality in the Balian–Low theorem was investigated in [6]. There, it was shown that the following result holds true.

THEOREM 1.2. *If $\frac{1}{p} + \frac{1}{q} = 1$, where $1 < p, q < \infty$, and $d > 2$, then there exists a function $g \in L^2(\mathbb{R})$ such that $\mathcal{G}(g,1,1)$ is an orthonormal basis for $L^2(\mathbb{R})$ and*

$$\int \frac{1+|t|^p}{\log^d(2+|t|)}|g(t)|^2dt < \infty \quad \textit{and} \quad \int \frac{1+|\gamma|^q}{\log^d(2+|\gamma|)}|\widehat{g}(\gamma)|^2d\gamma < \infty.$$

Letting $(p,q) = (2,2)$ in Theorem 1.2 shows how to construct Gabor orthonormal bases which are essentially optimally localized with respect to the Balian–Low theorem. In particular, the bases constructed come within a logarithmic factor of satisfying the forbidden localization hypotheses of the Balian–Low theorem.

Since Theorem 1.2 also constructs Gabor orthonormal bases for values of (p,q) other than $(2,2)$, it is natural to ask whether there are versions of the Balian–Low theorem for the weights (t^p, γ^q). The best that is known is the following.

THEOREM 1.3. *Suppose $\frac{1}{p} + \frac{1}{q} = 1$ with $1 < p < \infty$ and let $\epsilon > 0$. If*

$$\int |t|^{(p+\epsilon)}|g(t)|^2dt < \infty \quad \textit{and} \quad \int |\gamma|^{(q+\epsilon)}|\widehat{g}(\gamma)|^2d\gamma < \infty,$$

then $\mathcal{G}(g,1,1)$ is not an orthonormal basis for $L^2(\mathbb{R})$.

The above theorem follows by combining Theorem 4.4 of [12] and Theorem 1 of [17]. By the Balian–Low theorem, one may set $\epsilon = 0$ if $(p,q) = (2,2)$. A version of the Balian–Low theorem for the case $(p,q) = (1,\infty)$ is given in [7].

2. Overview. Theorem 1.2 constructively produces Gabor orthonormal bases which are almost optimally localized with respect to the Balian–Low theorem and Theorem 1.3. However, these bases do not have simple expressions. The main aim of this paper is to study the elegant Gabor orthonormal basis constructed by Høholdt, Jensen, and Justesen in [22] and to show that it is almost optimally localized with respect to Theorem 1.3 for a certain choice of (p,q). Their basis has a simpler, more explicit form than those in [6] and gives insight into other components needed for constructing well-localized Gabor bases. The key ingredients in the constructions in [22] and [6] are functions which possess unimodular Zak transforms with small singular supports. For perspective, we remark that [21] provides several examples of functions with Zak transforms with few zeros which are used to construct tight Gabor

frames. These examples could provide further insight into the study of optimality in the Balian–Low theorem and merit future investigation.

The remainder of the paper is organized as follows. In section 3, we recall the basis of Høholdt, Jensen, and Justesen, and we state our main result, Theorem 3.3. In section 4 we prove the time localization estimates for the basis, and in section 5 we prove the frequency localization estimates. We conclude with some relevant remarks in section 6.

3. The Gabor basis of Høholdt, Jensen, and Justesen. The Zak transform is an important tool in the analysis and construction of Gabor systems; see, e.g., [18, Chapter 8]. Given $g \in L^2(\mathbb{R})$, the *Zak transform* is formally defined by

$$\forall (t,\gamma) \in \mathbb{R}\times\mathbb{R}, \quad Zg(t,\gamma) = \sum_{n\in\mathbb{Z}} g(t-n)e^{2\pi i n\gamma}.$$

With the above definition, the Zak transform satisfies the *quasi-periodicity relations*

$$\forall k \in \mathbb{Z}, \quad Zf(x,\gamma+k) = Zf(x,\gamma)$$

and

$$\forall k \in \mathbb{Z}, \quad Zf(x+k,\gamma) = Zf(x,\gamma)\; e^{2\pi i k\gamma};$$

see, e.g., [18, section 8.2]. Thus, the Zak transform Zf of a function $f \in L^2(\mathbb{R})$ is a locally square integrable function defined on all of $\mathbb{R}^2$ and is uniquely determined by its values on $Q \equiv [0,1)^2$. Therefore, Z defines a unitary operator from $L^2(\mathbb{R})$ to $L^2(Q)$, and its inverse $Z^{-1} : L^2(Q) \to L^2(\mathbb{R})$ is formally given by

$$\forall t \in \mathbb{R}, \quad (Z^{-1}F)(t) = \int_0^1 F(t,\gamma)d\gamma.$$

The utility of the Zak transform for constructing Gabor bases stems from the following result (see, e.g., [18, Corollary 8.3.2]), which forms the foundation for the constructions in both [6] and [22]; cf. [21].

THEOREM 3.1. *Let $g \in L^2(\mathbb{R})$. Then $\mathcal{G}(g,1,1)$ is an orthonormal basis for $L^2(\mathbb{R})$ if and only if $|Zg(t,\gamma)| = 1$ for a.e. $(t,\gamma) \in Q$.*

This shows that constructing Gabor orthonormal bases is equivalent to constructing unimodular functions on $L^2(Q)$. Høholdt, Jensen, and Justesen consider the function $F \in L^2(Q)$ defined by

$$\forall (t,\gamma) \in Q, \quad F(t,\gamma) = \frac{1+\alpha(t)e^{2\pi i\gamma}}{1+\alpha(t)e^{-2\pi i\gamma}}, \tag{3.1}$$

where $\alpha : [0,1] \to [0,1]$ is a measurable function. In [22], the function α was chosen as $\alpha(t) = \sin(\frac{\pi}{2}t)$, since this was shown to minimize $\int |\gamma|^2 |\widehat{(Z^{-1}F)}(\gamma)|^2 d\gamma$.

DEFINITION 3.2. *Let $f \in L^2(\mathbb{R})$ be the function defined by* (3.1)*, where*

$$f = Z^{-1}F \quad \text{and} \quad \alpha(t) = \sin\left(\frac{\pi}{2}t\right). \tag{3.2}$$

It was proven in [22] *that $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and that f (see Figure* 3*) is explicitly defined by*

$$f(t) = \begin{cases} 0 & \text{if } t \in (-\infty,-1], \\ \sin(\frac{\pi}{2}(t+1)) & \text{if } t \in (-1,0], \\ (-1)^n \cos^2(\frac{\pi}{2}(t-n)) \sin^n(\frac{\pi}{2}(t-n)) & \text{if } t \in (n,n+1], n = 0,1,2,\ldots. \end{cases}$$

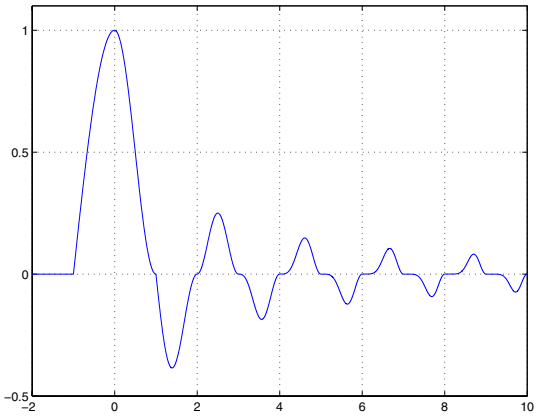


FIG. 3.1. *Graph of the function f.*

It is easy to verify that $|Zf(t,\gamma)| = |F(t,\gamma)| = 1$ for a.e. $(t,\gamma) \in Q$, and hence $\mathcal{G}(f,1,1)$ is an orthonormal basis for $L^2(\mathbb{R})$. We may now state our main result as follows.

THEOREM 3.3. *Let $f \in L^2(\mathbb{R})$ be the window function defined by Definition* 3.2. *For every $0 < \epsilon \leq 3/2$, f satisfies*

$$\int |t|^{3/2-\epsilon}|f(t)|^2 dt < \infty \quad \text{and} \quad \int |\gamma|^{3-\epsilon}|\widehat{f}(\gamma)|^2 d\gamma < \infty. \tag{3.3}$$

Moreover,

$$\int |t|^{3/2}|f(t)|^2 dt = \infty \quad \text{and} \quad \int |\gamma|^{3}|\widehat{f}(\gamma)|^2 d\gamma = \infty. \tag{3.4}$$

In particular, the Gabor orthonormal basis $\mathcal{G}(f,1,1)$ is almost optimally localized with respect to Theorem 1.3 with $(p,q) = (3/2,3)$.

4. Time localization estimates. In this section we derive the time localization estimates in Theorem 3.3.

THEOREM 4.1. *Let $f \in L^2(\mathbb{R})$ be the function defined in Definition* 3.2 *and let $a > 0$. Then*

$$\int |t|^a |f(t)|^2 dt < \infty \quad \text{if and only if} \quad a < 3/2.$$

Proof. A direct calculation shows that for $n = 0, 1, 2, \ldots$

$$\begin{aligned} \int_n^{n+1} |f(t)|^2 dt &= \int_0^1 \cos^4\left(\frac{\pi}{2}t\right) \sin^{2n}\left(\frac{\pi}{2}t\right) dt \\ &= \frac{2}{\pi}\left(\frac{3}{4n^2+12n+8}\right) \int_0^{\pi/2} \sin^{2n} u \, du. \end{aligned} \tag{4.1}$$

One can also calculate that

$$\frac{2}{\pi}\int_0^{\pi/2} \sin^{2n} u \, du = \frac{(1)(3)(5)(7)\cdots(2n-1)}{(2)(4)(6)(8)\cdots(2n)} \equiv P_n. \tag{4.2}$$

By taking the natural log of P_n and using Taylor approximations for $\ln(1-x)$ near $x=0$ to estimate the resulting sum, it is straightforward to show that

$$P_n \sim \frac{1}{\sqrt{n}}. \tag{4.3}$$

Equivalently, we could use Stirling's formula for the Gamma function to show (4.3). Here and subsequently $A \sim B$ means that $A \lesssim B \lesssim A$, where $A \lesssim B$, in turn, means that there exists an absolute constant C such that $A \leq CB$. When necessary, we shall point out any dependence of the implicit constants on other parameters. Therefore,

$$\begin{aligned}\int_1^\infty |t|^a |f(t)|^2 dt &\geq \sum_{n=1}^\infty n^a \int_n^{n+1} |f(t)|^2 dt \\ &= \sum_{n=1}^\infty n^a \left(\frac{3}{4n^2+12n+8}\right) P_n \gtrsim \sum_{n=1}^\infty n^{a-5/2}.\end{aligned}$$

In particular,

$$a \geq 3/2 \Longrightarrow \int |t|^a |f(t)|^2 dt = \infty.$$

Also, using (4.1), (4.2), and (4.3), we obtain the estimate

$$\int_1^\infty |t|^a |f(t)|^2 dt \lesssim \sum_{n=1}^\infty \frac{(n+1)^a}{n^{5/2}}.$$

Since f is bounded on $[-1,1]$, and $f=0$ on $(-\infty,-1)$, it follows that

$$0 < a < 3/2 \Longrightarrow \int |t|^a |f(t)|^2 dt < \infty. \qquad \square$$

5. Frequency localization estimates. In this section we derive the frequency localization estimates in Theorem 3.3.

THEOREM 5.1. *Let $f \in L^2(\mathbb{R})$ be the function defined in Definition* 3.2 *and let $0 < a$. Then*

$$\int |\gamma|^a |\widehat{f}(\gamma)|^2 d\gamma < \infty \quad \text{if and only if} \quad a < 3.$$

It will be convenient to view Theorem 5.1 in terms of Sobolev spaces. Given $s > 0$, the *homogeneous Sobolev space* of order s, denoted by $\dot{H}^s(\mathbb{R})$, consists of all $g \in L^2(\mathbb{R})$ such that $||g||^2_{\dot{H}^s(\mathbb{R})} \equiv \int |\gamma|^{2s} |\widehat{g}(\gamma)|^2 d\gamma < \infty$. For later convenience, we also define $\langle f, g\rangle_{\dot{H}^s(\mathbb{R})} = \int |\gamma|^{2s} \widehat{f}(\gamma)\overline{\widehat{g}(\gamma)} d\gamma$. Theorem 5.1 now says that $0 < s < 3/2$ implies $f \in \dot{H}^s(\mathbb{R})$, and that $s \geq 3/2$ implies $f \notin \dot{H}^s(\mathbb{R})$. The following result (see, e.g., [24, Chapter 8]) gives a useful alternate characterization of $\dot{H}^s(\mathbb{R})$. It is used in the proof of Lemma 5.6.

LEMMA 5.2. *If $0 < s < 2$ and $f \in \dot{H}^s(\mathbb{R})$, then there exists $C_s > 0$ such that*

$$||f||^2_{\dot{H}^s(\mathbb{R})} = C_s \int\int \frac{|f(x+t) + f(x-t) - 2f(x)|^2}{|t|^{1+2s}} dx dt.$$

Lemma 5.2 can be proven by applying Parseval's theorem to the inner integral. A similar calculation gives the following result used in the proof of Lemma 5.8.

LEMMA 5.3. *If $0 < s < 2$ and $f, g \in \dot{H}^s(\mathbb{R})$, then there exists $C_s > 0$ such that*

$$\langle f, g\rangle_{\dot{H}^s(\mathbb{R})} = C_s \iint \frac{(f(x+t)+f(x-t)-2f(x))\overline{(g(x+t)+g(x-t)-2g(x))}}{|t|^{2s+1}} dxdt.$$

We shall use the following lemma directly in Theorem 5.1.

LEMMA 5.4. *If $3 \le a$, then $f \notin \dot{H}^{a/2}(\mathbb{R})$.*

Proof. If $3 \le a < 4$, then for $0 < \eta$ small

$$||f||^2_{\dot{H}^{a/2}(\mathbb{R})} \sim \iint \frac{|f(x+t)+f(x-t)-2f(x)|^2}{|t|^{1+a}} dxdt \ge \int_0^\eta \int_{-t}^0 \frac{\sin^2(\frac{\pi}{2}(x+t))}{|t|^{1+a}} dxdt$$
$$\gtrsim \int_0^\eta \int_{-t}^0 \frac{(x+t)^2}{|t|^{1+a}} dxdt \gtrsim \int_0^\eta \frac{t^3}{|t|^{1+a}} dt = \infty.$$

Since $f \in L^1(\mathbb{R})$, it now also follows that $f \notin \dot{H}^{a/2}(\mathbb{R})$ for all $3 \le a$. ∎

We now prove that $0 < a < 3$ implies $f \in \dot{H}^{a/2}$. Since this is more involved than our prior estimates, we split it up into several lemmas (Lemmas 5.9 and 5.10). Lemma 5.5 is used in Lemma 5.6, which, in turn, is used in the proof of Lemma 5.7. Lemmas 5.7 and 5.8 allow us to prove Lemma 5.9.

LEMMA 5.5. *For $n \ge 3$, let $f_n(t) = \mathbf{1}_{(n,n+1]}(t) f(t)$, where $\mathbf{1}_S(t)$ denotes the characteristic function of a set $S \subseteq \mathbb{R}$. The functions f_n have the following properties:*

1. *f_n is continuous and differentiable on $\mathbb{R}$.*
2. *$f_n''(t)$ exists for all $t \in \mathbb{R}\backslash\{n+1\}$.*
3. *$||f_n||^2_{L^2(\mathbb{R})} \lesssim 1/n^{5/2}$ and $||f_n||^{1/2}_{L^{1/2}(\mathbb{R})} \lesssim 1/n$.*
4. *If $0 < \delta < 3/2$, then $||f_n||^{1/2+\delta}_{L^{1/2+\delta}(\mathbb{R})} \lesssim 1/n^{1+\delta}$.*
5. *For all $t \in \mathbb{R}\backslash\{n+1\}$, $|f_n''(t)| \lesssim 1$.*

The implicit constants in (3) *and* (4) *are independent of n, and the implicit constant in* (5) *is independent of t and n.*

Proof. The first two items can be verified by direct calculations. The estimate for $||f_n||^2_{L^2(\mathbb{R})}$ in (3) has already been done in the proof of Theorem 4.1. In fact,

$$||f_n||^2_{L^2(\mathbb{R})} = \left(\frac{3}{4n^2+12n+8}\right) P_n \lesssim \frac{1}{n^{2.5}}.$$

The estimate for $||f_n||^{1/2}_{L^{1/2}(\mathbb{R})}$ in (3) holds since

$$||f_n||^{1/2}_{L^{1/2}(\mathbb{R})} = \int_n^{n+1} \cos\left(\frac{\pi}{2}(t-n)\right) \sin^{\frac{n}{2}}\left(\frac{\pi}{2}(t-n)\right) dt = \frac{4}{\pi(n+2)}.$$

The fourth item follows from (3) and the following standard interpolation formula (see, e.g., [15, Proposition 6.10]):

$$||f||_{L^q(\mathbb{R})} \le ||f||^\lambda_{L^p(\mathbb{R})} ||f||^{1-\lambda}_{L^r(\mathbb{R})},$$

where

$$0 < p < q < r \le \infty \quad \text{and} \quad \lambda = \frac{1/q - 1/r}{1/p - 1/r}.$$

To prove (5) first note that, for $n < t < n+1$,

$$f_n''(t-n) = (-1)^n \left(-(5n+2)\frac{\pi^2}{4} \sin^n\left(\frac{\pi}{2}t\right)\cos^2\left(\frac{\pi}{2}t\right) + 2\frac{\pi^2}{4}\sin^{n+2}\left(\frac{\pi}{2}t\right) + \frac{\pi^2}{4}n(n-1)\cos^4\left(\frac{\pi}{2}t\right)\sin^{n-2}\left(\frac{\pi}{2}t\right)\right).$$

Therefore, $f_n''(t-n) = (-1)^n h_n(u)$, where $u = \sin^2(\frac{\pi}{2}(t-n))$, and

$$h_n(u) = \frac{\pi^2}{4} u^{n/2-1}\left[2u^2 - (5n+2)u(1-u) + n(n-1)(1-u)^2\right].$$

Straightforward, but tedious, calculations show that $|h_n(t)| \le C$ on $(n, n+1)$ for some constant C independent of t and n. Since $f_n''(t) = 0$ on $\mathbb{R}\backslash(n, n+1]$, we conclude that $|f_n''(t)| \lesssim 1$ on $\mathbb{R}\backslash\{n+1\}$. □

LEMMA 5.6. *Assume $0 < a < 3$, and let $\epsilon = 3 - a$. Then*

$$\forall n \ge 3, \quad ||f_n||^2_{\dot{H}^{a/2}(\mathbb{R})} \lesssim \frac{1}{n^{1+\epsilon/4}}.$$

The implicit constant is independent of n.

Proof. We shall estimate $||f_n||^2_{\dot{H}^{a/2}(\mathbb{R})}$ by using the double integral in Lemma 5.2. Let $B = \{t \in \mathbb{R} : |t| < 1\}$, and note that

$$\int_{\mathbb{R}\backslash B}\int_{\mathbb{R}} \frac{|f_n(x+t) + f_n(x-t) - 2f_n(x)|^2}{|t|^{a+1}} dxdt \lesssim ||f_n||^2_{L^2(\mathbb{R})} \int_{\mathbb{R}\backslash B} \frac{1}{|t|^{a+1}} dt \lesssim \frac{1}{n^{5/2}}.$$

It remains for us to estimate

$$\int_B \int_{\mathbb{R}} \frac{|f_n(x+t) + f_n(x-t) - 2f_n(x)|^2}{|t|^{a+1}} dxdt.$$

We write this as the sum of two integrals, over $[0,1]\times\mathbb{R}$ and $[-1,0]\times\mathbb{R}$, respectively. Since the estimates for both integrals are similar, it suffices to consider the first, which, in turn, is estimated by breaking it up into the following four integrals:

$$I_1 = \int_0^1 \int_{-\infty}^{n-t}, \quad I_2 = \int_0^1 \int_{n-t}^{n+1-t}, \quad I_3 = \int_0^1 \int_{n+1-t}^{n+1+t}, \quad I_4 = \int_0^1 \int_{n+1+t}^{\infty}.$$

First, note that the support properties of f_n imply that $I_1 = 0$ and $I_4 = 0$.

Next note that if $x+t, x-t, x$ are all less than $n+1$, then Lemma 5.5 and the mean value theorem imply

$$|f_n(x+t) + f_n(x-t) - 2f_n(x)| \lesssim |t|^2, \tag{5.1}$$

where the implicit constant is independent of x, t, and n. To estimate I_2 note that, by (5.1) and Lemma 5.5,

$$\begin{aligned} W_2(t) &\equiv \int_{n-t}^{n+1-t} |f_n(x+t) + f_n(x-t) - 2f_n(x)|^2 dx \\ &\lesssim |t|^{3-\epsilon/2} \int_{n-t}^{n+1-t} |f_n(x+t) + f_n(x-t) - 2f_n(x)|^{1/2+\epsilon/4} dx \\ &\lesssim |t|^{3-\epsilon/2}\, ||f_n||^{1/2+\epsilon/4}_{L^{1/2+\epsilon/4}(\mathbb{R})} \lesssim \frac{|t|^{3-\epsilon/2}}{n^{1+\epsilon/4}}. \end{aligned}$$

It now follows that

$$I_2 = \int_0^1 \frac{W_2(t)}{|t|^{a+1}}dt = \int_0^1 \frac{W_2(t)}{|t|^{4-\epsilon}}dt \lesssim 1/n^{1+\epsilon/4}\int_0^1 \frac{1}{|t|^{1-\epsilon/2}}dt \lesssim 1/n^{1+\epsilon/4}.$$

To estimate I_3, define

$$W_3(t) = \int_{n+1-t}^{n+1+t} |f_n(x+t) + f_n(x-t) - 2f_n(x)|^2 dx.$$

Note that by the definition of f_n and its support properties,

$$\begin{aligned} W_3(t) &= \int_{n+1-t}^{n+1+t} |f_n(x-t) - 2f_n(x)|^2 dx \lesssim \int_{n+1-2t}^{n+1} |f_n(x)|^2 dx \\ &\lesssim \int_{n+1-2t}^{n+1} \cos^4\left(\frac{\pi}{2}(x-n)\right) dx \lesssim |t|^5. \end{aligned}$$

Moreover, we also have

$$W_3(t) \lesssim ||f_n||^2_{L^2(\mathbb{R})} \lesssim \frac{1}{n^{5/2}}.$$

Thus, in order to estimate W_3 we may use the fact that for $x, y > 0$ and $\alpha \in [0,1]$, $\min\{x,y\} \le x^\alpha y^{1-\alpha}$. When $\alpha = (6-\epsilon)/10$, we obtain

$$W_3(t) \lesssim \frac{|t|^{3-\epsilon/2}}{n^{1+\epsilon/4}}.$$

Thus

$$I_3 = \int_0^1 \frac{W_3(t)}{|t|^{a+1}}dt = \int_0^1 \frac{W_3(t)}{|t|^{4-\epsilon}}dt \lesssim 1/n^{1+\epsilon/4}. \qquad \square$$

LEMMA 5.7. *Assume $0 < a < 3$ and let $\epsilon = 3 - a$. If $3 \le m, n$, and $|m-n| = 1$, then*

$$|\langle f_n, f_m\rangle_{\dot{H}^{a/2}(\mathbb{R})}| \lesssim \frac{1}{n^{1+\epsilon/4}}.$$

The implicit constant is independent of n and m.

Proof. Without loss of generality assume $m = n+1$. It follows from Lemma 5.6 that

$$\begin{aligned} |\langle f_n, f_m\rangle_{\dot{H}^{a/2}(\mathbb{R})}| &\le ||f_n||_{\dot{H}^{a/2}(\mathbb{R})}||f_m||_{\dot{H}^{a/2}(\mathbb{R})} \\ &\lesssim \left(\frac{1}{n^{1+\epsilon/4}}\right)^{\frac{1}{2}}\left(\frac{1}{m^{1+\epsilon/4}}\right)^{\frac{1}{2}} \le \frac{1}{n^{1+\epsilon/4}}. \qquad \square \end{aligned}$$

LEMMA 5.8. *Let $0 < a$. If $3 \le m, n$ and $1 < |m-n|$, then*

$$|\langle f_n, f_m\rangle_{\dot{H}^{a/2}(\mathbb{R})}| \lesssim \frac{1}{|m-n|^a |n|^{5/4}|m|^{5/4}}.$$

The implicit constant is independent of m and n.

Proof. Without loss of generality assume $0 < n < m - 1$. Let $S_{m,n} = \{t \in \mathbb{R} : |t| > (m - n - 1)/2\}$ and let

$$F_m(x,t) = f_m(x+t) + f_m(x-t) - 2f_m(x).$$

Note that

$$t \notin S_{m,n} \Longrightarrow F_m(x,t)F_n(x,t) = 0.$$

Also,

$$\int |F_m(x,t)F_n(x,t)|dx \lesssim ||f_n||_{L^2(\mathbb{R})}||f_m||_{L^2(\mathbb{R})}.$$

Therefore, by Lemma 5.3,

$$\begin{aligned}|\langle f_n, f_m\rangle_{\dot{H}^{a/2}(\mathbb{R})}| &\leq \int_{S_{m,n}}\int_{\mathbb{R}} \frac{|F_m(x,t)F_n(x,t)|}{|t|^{a+1}}dxdt \\ &\lesssim ||f_n||_{L^2(\mathbb{R})}||f_m||_{L^2(\mathbb{R})}\int_{S_{m,n}} \frac{1}{|t|^{a+1}}dt \\ &\lesssim \frac{||f_n||_{L^2(\mathbb{R})}||f_m||_{L^2(\mathbb{R})}}{|m-n|^a} \lesssim \frac{1}{n^{5/4}m^{5/4}|m-n|^a}. \qquad \square\end{aligned}$$

To estimate the norm $||f||_{\dot{H}^{a/2}(\mathbb{R})}$ we first break f up into the two parts $F_1(t) = f(t)\mathbf{1}_{(-1,3]}(t)$ and $F_2(t) = f(t)\mathbf{1}_{(3,\infty)}(t)$. Since $f = 0$ on $(-\infty, -1]$ we have $f = F_1 + F_2$. We have the following estimate for F_2.

LEMMA 5.9. *If $0 < a < 3$, then $||F_2||^2_{\dot{H}^{a/2}(\mathbb{R})} < \infty$.*

Proof. Let $\epsilon = 3 - a$, and note that $F_2 = \sum_{n=3}^{\infty} f_n$. Define

$$S_1 = \{(m,n) \in \mathbb{Z}^2 : m, n \geq 3 \text{ and } |m-n| = 1\}$$

and

$$S_2 = \{(m,n) \in \mathbb{Z}^2 : m, n \geq 3 \text{ and } |m-n| > 1\}.$$

By Lemmas 5.6, 5.7, and 5.8 we have

$$\begin{aligned}||F_2||^2_{\dot{H}^{a/2}(\mathbb{R})} &= \left|\left|\sum_{n=3}^{\infty} f_n\right|\right|^2_{\dot{H}^{a/2}(\mathbb{R})} \leq \sum_{m=3}^{\infty}\sum_{n=3}^{\infty} |\langle f_m, f_n\rangle_{\dot{H}^{a/2}(\mathbb{R})}| \\ &= \sum_{n=3}^{\infty} ||f_n||^2_{\dot{H}^{a/2}(\mathbb{R})} + \sum_{(m,n)\in S_1} |\langle f_n, f_m\rangle_{\dot{H}^{a/2}(\mathbb{R})}| + \sum_{(m,n)\in S_2} |\langle f_n, f_m\rangle_{\dot{H}^{a/2}(\mathbb{R})}| \\ &\lesssim \sum_{n=1}^{\infty} \frac{1}{n^{1+\epsilon/4}} + \sum_{n=1}^{\infty} \frac{1}{n^{1+\epsilon/4}} + \sum_{(m,n)\in S_2} \frac{1}{n^{5/4}m^{5/4}|m-n|^{2a}} < \infty. \qquad \square\end{aligned}$$

In view of Lemma 5.9, and since

$$||f||_{\dot{H}^{a/2}(\mathbb{R})} \leq ||F_1||_{\dot{H}^{a/2}(\mathbb{R})} + ||F_2||_{\dot{H}^{a/2}(\mathbb{R})},$$

it remains only to estimate $||F_1||_{\dot{H}^{a/2}(\mathbb{R})}$. Note that by the definition of f, F_1 is compactly supported, continuous on $\mathbb{R}$, and infinitely differentiable away from $x_1 =$

$-1, x_2 = 0, x_3 = 1, x_4 = 2$, and $x_5 = 3$. Moreover the first derivative of F_1 also exists at x_2, x_3, x_5. However, the second derivative of F_1 does not exist at any of the points x_1, x_2, x_3, x_4, x_5.

It therefore suffices to estimate $||\varphi_j F_1||_{\dot{H}^{a/2}(\mathbb{R})}$ for $j = 1, 2, 3, 4, 5$, where φ_j is an infinitely differentiable function satisfying

$$\varphi_j(x) = 1 \text{ for } |x - x_j| < 2\nu \quad \text{and} \quad \varphi_j(x) = 0 \text{ for } |x - x_j| > 4\nu,$$

with $0 < \nu$ sufficiently small.

We present a proof that is analogous to our previous estimates and uses Lemma 5.2. Alternately, one can proceed more directly and use an argument involving integration by parts.

LEMMA 5.10. *Let $0 < a < 3$ and let $\varphi_j F_1$ be as above for $j = 1, 2, 3, 4, 5$. Then*

$$||\varphi_j F_1||_{\dot{H}^{a/2}(\mathbb{R})} < \infty, \quad j = 1, 2, 3, 4, 5.$$

Consequently, $||F_1||_{\dot{H}^{a/2}(\mathbb{R})} < \infty$.

Proof. We shall show the estimate for $||\varphi_3 F_1||_{\dot{H}^{a/2}(\mathbb{R})}$ only, since the other four estimates proceed along similar lines.

Let $h(t) = (\varphi_3 F_1)(t - 1)$. We need to estimate the double integral

$$||\varphi_3 F_1||^2_{\dot{H}^{a/2}(\mathbb{R})} = ||h||^2_{\dot{H}^{a/2}(\mathbb{R})} \sim \int\int \frac{|h(x+t) + h(x-t) - 2h(x)|^2}{|t|^{1+a}} dxdt.$$

Let ν be as in the definition of φ_3 above, and note that if $B_\nu = \{t \in \mathbb{R} : |t| < \nu\}$, then

$$\int_{\mathbb{R}\setminus B_\nu} \int_{\mathbb{R}} \frac{|h(x+t) + h(x-t) - 2h(x)|^2}{|t|^{1+a}} dxdt \lesssim ||h||^2_{L^2(\mathbb{R})} \int_{\mathbb{R}\setminus B_\nu} \frac{1}{|t|^{1+a}} dt < \infty.$$

Next note that $h(t)$ is infinitely differentiable away from $t = 0$ and has bounded first and second derivatives on $\mathbb{R}\setminus\{0\}$. Therefore, if $x + t, x - t$ and x are all positive, or all negative, then it follows from the mean value theorem that

$$|h(x+t) + h(x-t) - 2h(x)| \lesssim |t|^2. \tag{5.2}$$

Likewise, if $x + t$ and x are both positive or both negative, then

$$|h(x+t) - h(x)| \lesssim |t|. \tag{5.3}$$

The implicit constants in (5.2) and (5.3) are independent of x and t.

To estimate the remaining integral

$$\int_{-\nu}^{\nu} \int_{\mathbb{R}} \frac{|h(x+t) + h(x-t) - 2h(x)|^2}{|t|^{1+a}} dxdt,$$

we break this integral up over the domains $[\nu, 0] \times \mathbb{R}$ and $[-\nu, 0] \times \mathbb{R}$. Since both integrals are similar we show estimates for the first only, which, in turn, we estimate by considering the integrals

$$J_1 = \int_0^\nu \int_t^\infty, \quad J_2 = \int_0^\nu \int_0^t, \quad J_3 = \int_0^\nu \int_{-t}^0, \quad J_4 = \int_0^\nu \int_{-\infty}^{-t}.$$

It follows from (5.2) and the compact support of h that

$$\begin{aligned} J_1 &= \int_0^\nu \int_t^\infty \frac{|h(x+t)+h(x-t)-2h(x)|^2}{|t|^{1+a}}dxdt \\ &= \int_0^\nu \int_t^{5\nu} \frac{|h(x+t)+h(x-t)-2h(x)|^2}{|t|^{1+a}}dxdt \lesssim \int_0^\nu \frac{|t|^4}{|t|^{1+a}}dt < \infty. \end{aligned}$$

The estimate for $J_4 < \infty$ is similar.

To estimate J_2, note that $J_2 \lesssim J_{2,1} + J_{2,2}$ where

$$J_{2,1} = \int_0^\nu \int_0^t \frac{|h(x+t)-h(x)|^2}{|t|^{a+1}}dxdt \quad \text{and} \quad J_{2,2} = \int_0^\nu \int_0^t \frac{|h(x-t)-h(x)|^2}{|t|^{a+1}}dxdt.$$

It follows from (5.3) that

$$J_{2,1}(t) \lesssim \int_0^\nu \int_0^t \frac{|t|^2}{|t|^{1+a}}dxdt = \int_0^\nu \frac{|t|^3}{|t|^{a+1}}dt < \infty.$$

Next, define

$$\begin{aligned} H(x,t) &= \cos^2\left(\frac{\pi}{2}(x-t-1)\right) + \cos^2\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}x\right) \\ &= \sin^2\left(\frac{\pi}{2}(x-t)\right) + \cos^2\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}x\right), \end{aligned}$$

and note that

$$\int_0^t |H(x,t)|^2 dx \lesssim \int_0^t |(x-t)^2|^2 dx + \int_0^t |x|^2 dx \lesssim |t|^3.$$

It now follows that

$$J_{2,2} = \int_0^\nu \int_0^t \frac{|H(x,t)|^2}{|t|^{a+1}}dxdt \lesssim \int_0^\nu \frac{|t|^3}{|t|^{a+1}}dt < \infty.$$

Therefore,

$$J_2 \lesssim J_{2,1} + J_{2,2} < \infty.$$

By using calculations similar to those used to deal with J_2, one can also show that $J_3 < \infty$. We can now conclude that $||\varphi_3 F_1||_{\dot{H}^{a/2}(\mathbb{R})} = ||h||_{\dot{H}^{a/2}(\mathbb{R})} < \infty$.

The estimates for $||\varphi_j F_2||_{\dot{H}^{a/2}(\mathbb{R})}$, $j = 1, 2, 4, 5$, proceed along similar lines to those above. □

Proof of Theorem 5.1. Combining Lemmas 5.9 and 5.10 shows that if $0 < a < 3$, then

$$\left(\int |\gamma|^a |\widehat{f}(\gamma)|^2 d\gamma\right)^{1/2} = ||f||_{\dot{H}^{a/2}(\mathbb{R})} \leq ||F_1||_{\dot{H}^{a/2}(\mathbb{R})} + ||F_2||_{\dot{H}^{a/2}(\mathbb{R})} < \infty.$$

Together with Lemma 5.4 this completes the proof of Theorem 5.1. □

Our main result, Theorem 3.3, now follows by combining Theorems 4.1 and 5.1.

6. Concluding remarks. 1. Throughout this remark we shall assume that Zak transforms have been quasi-periodically extended to $L^2_{\text{loc}}(\mathbb{R}^2)$. A key idea in the construction of [6] was to choose the Gabor window function g so that $|Zg| = 1$ almost everywhere and such that Zg has minimal singular support. In fact, the function Zg used in [6] was locally C^∞ on $\mathbb{R}^2$ except at one point in each square $S_{j,k} = (j, j+1] \times (k, k+1], j, k \in \mathbb{Z}$.

By comparison, one can show that the quasi-periodic extension of Høholdt, Jensen, and Justesen's function $F = Zf$ defined in (3.1) is continuous on $\mathbb{R}^2$ except at the set $\{(j, k+1/2) : j, k \in \mathbb{Z}\}$. However, F is nondifferentiable on the set $\{(t, j) : t \in \mathbb{R}, j \in \mathbb{Z}\}$. In this regard, the construction in [6] provides a Gabor orthonormal basis $\mathcal{G}(g, 1, 1)$ such that Zg has more smoothness than $F = Zf$ in (3.1).

2. We have shown that the basis of Høholdt, Jensen, and Justesen is almost optimally localized with respect to the (p, q) Balian–Low result in Theorem 1.3 when $(p, q) = (3/2, 3)$. It would be interesting to see whether Høholdt, Jensen, and Justesen's method of construction can be extended to provide optimality for other values of (p, q). With respect to further potentially optimal examples, Janssen in [20] and [21] provides several other families of functions which have the Zak transforms with minimal singular support. These include Gaussians, hyperbolic secants, and two-sided exponentials. The operation $Z^{-1}(Zg/|Zg|)$ applied to these functions yields examples of Gabor orthonormal bases for $L^2(\mathbb{R})$. The examples of Janssen are analogous in nature to the examples in [22] and [6], but they possess more symmetry in their decay properties. At the present, [6] provides the only construction which has been proven to be optimal for general values of the time and frequency localization parameters (p,q).

Acknowledgments. The authors thank Jacob Sterbenz for valuable discussions related to the material. They also thank the referees for many valuable suggestions, and in particular for pointing out the relevance of the work [21]. A portion of this work was done while the third author was a Junior Research Fellow at the Erwin Schrödinger Institute (ESI) in Vienna. He is grateful to ESI for their hospitality and financial support, and to Hans Feichtinger for having helped to arrange this visit.

REFERENCES

[1] R. Balan and I. Daubechies, *Optimal stochastic encoding and approximation schemes using Weyl–Heisenberg sets*, in Advances in Gabor Analysis, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 2003, pp. 259–320.

[2] R. Balian, *Un principe d'incertitude fort en théorie du signal en méchanique quantique*, C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 292 (1981), pp. 1357–1362.

[3] G. Battle, *Heisenberg proof of the Balian–Low theorem*, Lett. Math. Phys., 15 (1988), pp. 175–177.

[4] J. J. Benedetto and D. F. Walnut, *Gabor frames for L^2 and related spaces*, in Wavelets: Mathematics and Applications, CRC Press, Boca Raton, FL, 1994, pp. 97–162.

[5] J. J. Benedetto, W. Czaja, and A. Y. Maltsev, *The Balian–Low theorem for the symplectic form on $\mathbb{R}^{2d}$*, J. Math. Phys., 44 (2003), pp. 1735–1750.

[6] J. J. Benedetto, W. Czaja, P. Gadziński, and A. M. Powell, *The Balian–Low theorem and regularity of Gabor systems*, J. Geom. Anal., 13 (2003), pp. 239–254.

[7] J. J. Benedetto, W. Czaja, A. M. Powell, and J. Sterbenz, *An endpoint $(1, \infty)$ Balian–Low theorem*, Math. Res. Lett., to appear.

[8] J. J. Benedetto, C. Heil, and D. Walnut, *Differentiation and the Balian–Low theorem*, J. Fourier Anal. Appl., 1 (1995), pp. 355–402.

[9] J. J. Benedetto and A. M. Powell, *A (p, q) version of Bourgain's theorem*, Trans. Amer. Math. Soc., 358 (2006), pp. 2489–2505.

[10] J. Bourgain, *A remark on the uncertainty principle for Hilbertian basis*, J. Funct. Anal., 79 (1988), pp. 136–143.
[11] W. Czaja and A. M. Powell, *Recent developments in the Balian–Low theorem*, in Harmonic Analysis and Applications, C. Heil, ed., Birkäuser Boston, Boston, MA, 2005, to appear.
[12] H. G. Feichtinger and K. Gröchenig, *Gabor frames and time-frequency analysis of distributions*, J. Funct. Anal., 146 (1997), pp. 464–495.
[13] H. G. Feichtinger and T. Strohmer, eds., *Gabor Analysis and Algorithms. Theory and Applications*, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 1998.
[14] H. G. Feichtinger and T. Strohmer, eds., *Advances in Gabor Analysis*, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2003.
[15] G. F. Folland, *Real Analysis. Modern Techniques and Their Applications*, Pure Appl. Math. (N. Y.), Wiley-Interscience, New York, 1999.
[16] J.-P. Gabardo and D. Han, *Balian–Low phenomenon for subspace Gabor frames*, J. Math. Phys., 45 (2004), pp. 3362–3378.
[17] K. Gröchenig, *An uncertainty principle related to the Poisson summation formula*, Studia Math., 121 (1996), pp. 87–104.
[18] K. Gröchenig, *Foundations of Time-Frequency Analysis*, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2001.
[19] K. Gröchenig, D. Han, C. Heil, and G. Kutyniok, *The Balian–Low theorem for symplectic lattice in higher dimensions*, Appl. Comput. Harmon. Anal., 13 (2002), pp. 169–176.
[20] A. J. E. M. Janssen, *Zak transforms with few zeros and the tie*, in Advances in Gabor Analysis, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 2003, pp. 31–70.
[21] A. J. E. M. Janssen, *On generating tight Gabor frames at critical density*, J. Fourier Anal. Appl. 9 (2003), pp. 175–214.
[22] H. Jensen, T. Høholdt, and J. Justesen, *Double series representation of bounded signals*, IEEE Trans. Inform. Theory, 34 (1988), pp. 613–624.
[23] F. Low, *Complete sets of wave packets*, in A Passion for Physics—Essays in Honor of Geoffrey Chew, C. DeTar et al., eds., World Scientific, Singapore, 1985, pp. 17–22.
[24] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton Math. Ser. 30, Princeton University Press, Princeton, NJ, 1970.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 347–367

THE STRUCTURE OF C^1 SPLINE SPACES ON FREUDENTHAL PARTITIONS*

G. HECKLIN†, G. NÜRNBERGER†, AND F. ZEILFELDER†

Abstract. We analyze the structure of trivariate C^1 splines on uniform tetrahedral partitions Δ. The Freudenthal partitions Δ are obtained from uniform cube partitions by using three planes with a common line to subdivide every cube into six tetrahedra. This is a natural three-dimensional generalization of the well-known three-directional mesh in the plane. By using Bernstein–Bézier techniques, we construct minimal determining sets for C^1 spline spaces on Δ of arbitrary degree and give explicit formulae for the dimension of the spaces.

Key words. trivariate splines, Freudenthal partitions, C^1 smoothness conditions, minimal determining sets, Bernstein–Bézier techniques, dimension of spline spaces

AMS subject classifications. 65D07, 41A63, 65D17

DOI. 10.1137/040614980

1. Introduction.

Spline spaces play a fundamental role in approximation theory and computer aided geometric design. In particular, this concerns *multivariate splines* (see Chui [12], and the references therein), i.e., splines in more than one variable which are defined on appropriate partitions of a multidimensional domain. In contrast to the univariate case (cf. Nürnberger [24], Schumaker [34]), only few results on the structure of these spaces are known. Even the most basic problems such as determining the *dimension* (i.e., the number of degrees of freedom) of the spaces and *constructing interpolation sets* are often difficult to solve. Recently, there has been some progress in developing efficient interpolation and approximation methods for bivariate splines (cf. Alfeld and Schumaker [4, 5], Davydov and Zeilfelder [13], Lai and Schumaker [19, 20], Nürnberger, Rayevskaya, Schumaker and Zeilfelder [25], Nürnberger, Schumaker and Zeilfelder [27], Nürnberger and Zeilfelder [29, 30], Zeilfelder [42], and the references therein) which are useful for many applications. On the other hand, much less is known for *trivariate splines*, i.e., splines defined on *tetrahedral partitions*. One reason for this is that the smoothness conditions imply complex relations between the coefficients of the splines. For example, it has been observed (cf. Alfeld, Schumaker and Sirvent [6, Example 25], Alfeld, Schumaker and Whiteley [7, Remark 66]) that even for tetrahedral partitions of certain conic domains an analysis of these conditions would require the knowledge of the structure of bivariate spline spaces of any degree on general triangulations. It is known that even for *uniform (type) three-dimensional partitions* it is a complex task to analyze this structure since all the smoothness conditions in space directions have to be taken into account. In this paper, we investigate C^1 spline spaces of any polynomial degree defined on uniform tetrahedral partitions.

Different types of splines in three variables have been suggested in the literature (see the comparison and discussion in Remark 6.2). For trivariate splines as we consider here, there are basically two approaches. For the case of arbitrary tetrahedral partitions, there exist methods for *super spline spaces* (i.e., subspaces with a higher smoothness on lower-dimensional faces) with a relatively high degree (cf. Ženišek [43]).

*Received by the editors September 14, 2004; accepted for publication (in revised form) September 22, 2005; published electronically May 12, 2006.
http://www.siam.org/journals/sima/38-2/61498.html
†University of Mannheim, Institute for Mathematics, D-68131 Mannheim, Germany (hecklin@rumms.uni-mannheim.de, nuern@rumms.uni-mannheim.de, zeilfeld@rumms.uni-mannheim.de).

A C^1 quintic super spline construction, where all the tetrahedra are split into four *subtetrahedra*, was given by Alfeld [2], while the C^1 cubic and quadratic spline methods of Worsey and Farin [40] and Worsey and Piper [41] (see also Sorokina and Worsey [38]) are based on splitting all the tetrahedra into 12 and 24 subtetrahedra, respectively. In the second approach, classes of tetrahedral partitions are investigated, where no splits are used. Smooth quintic super spline constructions on such partitions were recently developed by Lai and Le Méhauté [18] and Schumaker and Sorokina [35]. One special class of tetrahedral partitions is obtained from a cube partition, where each cube is uniformly split into 24 tetrahedra. These partitions are sometimes called *type-6 tetrahedral partitions* and are natural three-dimensional generalizations of the well-known *four-direction mesh* in the plane. For these partitions, independently, Schumaker and Sorokina [36] constructed a generalization of the element of Fraejis de Veubeke and Sander (see the survey of Zeilfelder [42], for instance) by using C^1 trivariate super splines of degree six, while Hangelbroek, Nürnberger, Rössl, Seidel and Zeilfelder [16] determined the dimension of C^1 spline spaces (with no super smoothness) of arbitrary degree. Based on the latter results on the structure of spline spaces, Nürnberger, Rössl, Seidel and Zeilfelder [26, 31, 32] (see also Schlosser, Hesser, Zeilfelder, Rössl, Männer, Nürnberger, and Seidel [33]) developed methods for the *approximation and visualization of volumetric data* by using quadratic and cubic trivariate splines with appropriate smoothness properties. This shows that a structural analysis of the spline spaces is sometimes the first step to develop approaches useful for the applications, where volumetric models with advantageous properties such as high-quality visualization have to be reconstructed from discrete data. The numerical tests in [26, 31, 32, 33] indicate the potential of trivariate splines to approximate huge volumetric (gridded and scattered) data sets with up to millions of data points with a high efficiency.

In this paper, we determine the dimension of trivariate C^1 splines of arbitrary degree on a natural class of uniform type tetrahedral partitions Δ, where no tetrahedron is split. These partitions Δ generalize the well-known *three-direction mesh* in the plane and are called *Freudenthal partitions*. The Freudenthal partitions Δ are obtained from a uniform cube partition of a domain Ω in $\mathbb{R}^3$, where each cube is subdivided into six tetrahedra (see Figure 1). Hence the number of tetrahedra within each cube is smaller than for the above-mentioned type-6 tetrahedral partitions. We investigate the structure of the *spaces of trivariate C^1 polynomial splines on* Δ, i.e., the spaces defined by

$$\mathcal{S}_q^1(\Delta) = \{s \in C^1(\Omega) : s_{|_T} \in \mathcal{P}_q \text{ for all tetrahedra } T \in \Delta\}, \tag{1.1}$$

where

$$\mathcal{P}_q = \text{ span}\{x^i y^j z^k : \ i,j,k \geq 0, \ i+j+k \leq q\}$$

denotes the space of *trivariate polynomials of total degree* q. The main result of this paper (Theorem 3.1) is an explicit formula for the dimension of the spline spaces in (1.1). Its proof is complex. We construct a so-called *minimal determining set* (cf. Alfeld, Piper and Schumaker [3], Schumaker and Sorokina [35, 36]) for $\mathcal{S}_q^1(\Delta)$ based on the piecewise *Bernstein–Bézier representation* of the splines. Our inductive construction is based on two steps. In the first step, we construct minimal determining sets for C^1 splines defined on a tetrahedral partition of a single cube. Then, in the second step, we construct a minimal determining set for the whole C^1 spline space (Theorem 5.1), where in each step, we determine the remaining degrees of freedom. To do this, we pass through the cubes in an appropriate order. The cardinality of the minimal determining set gives the dimension of the spline spaces. As a main tool

for proving our results, we use a well-known *characterization of the C^1 smoothness* of two neighboring polynomial pieces in its Bernstein–Bézier representation (cf. de Boor [8], Farin [14]). We note that, although the results presented here are of independent interest, this paper can be understood as a basis for our recent local data interpolation methods (cf. Hecklin, Nürnberger, Schumaker and Zeilfelder [17]) for *cubic C^1* splines on *partially subdivided* Freudenthal partitions, which yield optimal approximation order of smooth functions. A first paper (cf. Nürnberger, Schumaker and Zeilfelder [28]) dealing with the difficult problem of constructing local Lagrange interpolation sets for smooth trivariate splines has recently appeared.

The paper is organized as follows. In section 2, we describe trivariate splines on Freudenthal partitions Δ, their piecewise Bernstein–Bézier representation and C^1 smoothness conditions. Section 3 contains our main result on the dimension of the spline spaces $\mathcal{S}_q^1(\Delta)$, and the definition of minimal determining sets. In section 4, we construct such sets for C^1 splines on a partition of a cube and give an explicit formula for the dimension of the corresponding spline spaces. These results are used in section 5, where we construct minimal determining sets for the whole space of C^1 splines and prove our main results. We close the paper with some remarks.

2. Splines on Freudenthal partitions and C^1 smoothness conditions. Let $\diamondsuit$ be the uniform cube partition of the domain $\Omega = [0,n] \times [0,n] \times [0,n] \subseteq \mathbb{R}^3$ which is obtained by intersecting Ω with $n+1$ parallel planes in each of the three space dimensions, i.e.,

$$\diamondsuit = \{Q_{(i,j,k)} : \ Q_{(i,j,k)} = [i-1,i] \times [j-1,j] \times [k-1,k],\ i,j,k = 1,\dots,n\}.$$

We cut each of the n^3 cubes $Q_{(i,j,k)},\ i,j,k = 1,\dots,n$, in $\diamondsuit$ with the three planes in $\mathbb{R}^3$ defined by

$$L_1 : \ x - y = i - j, \quad L_2 : \ x - z = i - k, \quad \text{and} \quad L_3 : \ y - z = j - k. \tag{2.1}$$

This leads to a natural, uniform tetrahedral partition Δ of Ω, where in correspondence to the six possible orderings of the variables x, y, and z each cube $Q \in \diamondsuit$ is split into six tetrahedra (see Figure 1, left and middle). More precisely, the construction is as follows. First, we subdivide each cube by slicing with L_1. Then, the two resulting prisms are further subdivided by slicing each of them with L_2 and L_3. Note that the planes L_1, L_2, and L_3 coincide along the main diagonal of $Q_{i,j,k}$, which connects the points $(i-1, j-1, k-1)$ and (i,j,k). It can be observed that the six tetrahedra

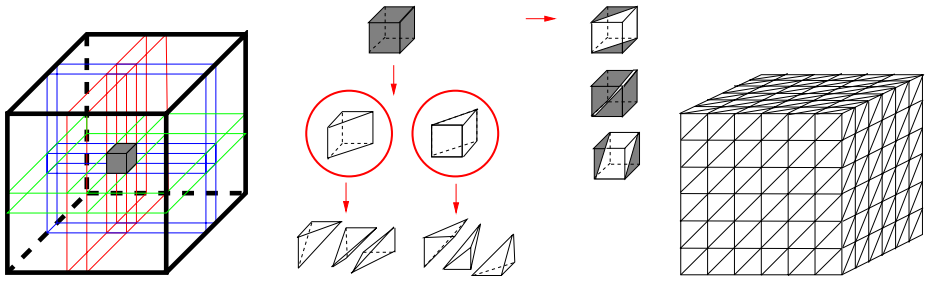


FIG. 1. *Freudenthal partitions Δ are obtained by splitting each cube of a uniform cube partition $\diamondsuit$ (left) into six tetrahedra by using three planes: first each cube is split into two prisms, then both prisms are further subdivided into three tetrahedra (middle). Each cutting plane of these partitions is a three-direction mesh known from bivariate spline theory (right).*

within each cube Q join the main diagonal of Q as a common edge. Moreover, the six tetrahedra can be obtained by taking the main diagonal of the cube as one edge, where the opposite edge is one of the six edges of the cube which are disjoint to the main diagonal. It is easy to see that the intersection of Δ with any plane parallel to the $xy-, xz-$, and $yz-$plane as well as with L_1, L_2, and L_3 is a three-direction mesh known from the bivariate setting (see Figure 1, right). Therefore, the tetrahedral partitions Δ are natural three-dimensional generalizations of these uniform triangulations. According to our knowledge, the partitions Δ were first considered by Freudenthal [15]; therefore we follow Carr, Möller and Snoeyink [11] and call them *Freudenthal (tetrahedral) partitions*.

Counting the number of tetrahedra N_Δ, triangular faces F_Δ, edges E_Δ, and vertices V_Δ of a Freudenthal partition Δ, respectively, it is easy to see that

$$N_\Delta = 6\ n^3, \quad F_\Delta = 12\ n^3 + 6n^2, \quad E_\Delta = 7\ n^3 + 9\ n^2 + 3\ n,$$
$$\text{and} \quad V_\Delta = n^3 + 3\ n^2 + 3\ n + 1. \tag{2.2}$$

In the following, we consider trivariate C^1 splines of arbitrary degree with respect to Δ . This means that we investigate the spaces $\mathcal{S}_q^1(\Delta)$ defined in (1.1). Since this is the trivial space of linear polynomials for $q = 1$, we are obviously interested in the choice $q \geq 2$. In order to analyze the structure of these spaces, the *smoothness conditions* across the interior triangular faces of Δ have to be taken into account. To do this, we use the piecewise Bernstein–Bézier representation of the splines from $\mathcal{S}_q^1(\Delta)$, which allows us to describe these conditions in a convenient form.

Given a (nondegenerate) tetrahedron $T = [v_0, v_1, v_2, v_3] \in \Delta$ with vertices v_0, v_1, v_2, and v_3, the linear polynomials $\Phi_\nu \in \mathcal{P}_1,\ \nu = 0, \ldots, 3$, with the interpolation property $\Phi_\nu(v_\mu) = \delta_{\nu,\mu},\ \mu = 0, \ldots, 3$, are called the *barycentric coordinates w.r.t. T*. Every polynomial piece $p = s_{|T} \in \mathcal{P}_q$ of a continuous spline s on Δ can be written in its *Bernstein–Bézier representation*

$$s_{|T}(z) = \sum_{\omega+\rho+\sigma+\tau=q} b_{\omega,\rho,\sigma,\tau}\ B^{q,T}_{\omega,\rho,\sigma,\tau}(z), \qquad z \in T, \tag{2.3}$$

where $B^{q,T}_{\omega,\rho,\sigma,\tau} = q!/(\omega!\rho!\sigma!\tau!)\ \Phi_0^\omega \Phi_1^\rho \Phi_2^\sigma \Phi_3^\tau \in \mathcal{P}_q,\ \omega+\rho+\sigma+\tau = q$, are the *Bernstein polynomials of degree q w.r.t. T* and $b_{\omega,\rho,\sigma,\tau} \in \mathbb{R},\ \omega+\rho+\sigma+\tau = q$, are the *Bernstein–Bézier coefficients* of p. It is well known (cf. Alfeld, Piper and Schumaker [3], Alfeld, Schumaker and Whiteley [7]) that each coefficient $b_{\omega,\rho,\sigma,\tau}$ in the Bernstein–Bézier representation (2.3) of $s_{|T}$ can be uniquely associated with the *domain point* $\xi_{\omega,\rho,\sigma,\tau} = (\omega\ v_0 + \rho\ v_1 + \sigma\ v_2 + \tau\ v_3)/q$, i.e., we define $b_{\omega,\rho,\sigma,\tau}(s) := b_{\omega,\rho,\sigma,\tau}(s_{|T}) = b_{\omega,\rho,\sigma,\tau},\ \omega + \rho + \sigma + \tau = q$. We note that the coefficients associated with the domain points on common triangular faces of the tetrahedra coincide, since the splines are continuous. Moreover, we set $\mathcal{D}_q(\Delta)$ as the union of all sets of domain points $\mathcal{D}_{q,T}$ w.r.t. the tetrahedra T of Δ.

Due to the special structure of Freudenthal partitions Δ the domain points can be described in terms of Euclidian coordinates. In the following, we set

$$\xi^{Q_{(i,j,k)}}_{\alpha,\beta,\gamma} = (i - 1 + \alpha/q, j - 1 + \beta/q, k - 1 + \gamma/q), \qquad \alpha, \beta, \gamma = 0, \ldots, q,$$

for the domain points $\mathcal{D}_{q,Q_{(i,j,k)}}$ within the cube $Q_{(i,j,k)}$ and indicate for $i, j, k \in \{1, \ldots, n\}$,

$$b_\xi(s) = b_\xi, \quad \xi = \xi^{Q_{(i,j,k)}}_{\alpha,\beta,\gamma}, \quad \alpha, \beta, \gamma = 0, \ldots, q.$$

We use this natural notation to keep the proofs in the subsequent sections short.

It is known that for any tetrahedral partition, the number of domain points coincides with the dimension of continuous spline spaces of arbitrary degree. For Freudenthal partitions Δ, a straightforward computation shows that this number is equal to $(q+1+(n-1)\ q)^3 = (n\ q+1)^3 = q^3\ n^3 + 3q^2\ n^2 + 3q\ n + 1$, since the continuity of a spline s on Δ is equivalent to $b_\xi(s) = b_\xi = b_{\xi'} = b_{\xi'}(s)$ for all ξ, ξ', where

$$\left(\xi = \xi^{Q_{(i,j,k)}}_{q,\alpha,\beta} \text{ and } \xi' = \xi^{Q_{(i+1,j,k)}}_{0,\alpha,\beta}\right), \qquad \left(\xi = \xi^{Q_{(j,i,k)}}_{\alpha,q,\beta} \text{ and } \xi' = \xi^{Q_{(j,i+1,k)}}_{\alpha,0,\beta}\right),$$
$$\left(\xi = \xi^{Q_{(j,k,i)}}_{\alpha,\beta,q} \text{ and } \xi' = \xi^{Q_{(j,k,i+1)}}_{\alpha,\beta,0}\right), \qquad \alpha,\beta = 0,\dots,q,$$

and $i,j,k = 1,\dots,n,\ i \neq n$. The situation is much more complex if we consider differentiable splines s on Δ, i.e., $s \in \mathcal{S}^1_q(\Delta)$, where $\mathcal{S}^1_q(\Delta)$ is the space defined in (1.1). As noted above, in this case, it is necessary to describe the *smoothness conditions* of the polynomial pieces of the splines on *neighboring tetrahedra* of Δ (i.e., tetrahedra with a common triangular face) in a convenient form. It is well known that this can be done by using the next result of Farin [14] (see also de Boor [8], and Chui [12]) which characterizes smoothness conditions for splines on two neighboring tetrahedra in terms of the coefficients of its piecewise Bernstein–Bézier representation.

THEOREM 2.1. *Let two neighboring tetrahedra $T = [v_0, v_1, v_2, v_3]$, $\tilde{T} = [v_0, v_1, v_2, \tilde{v}_3]$, and a continuous spline s on $T \cup \tilde{T}$ with*

$$s_{|_T} = p = \sum_{\omega+\rho+\sigma+\tau=q} b_{\omega,\rho,\sigma,\tau}\ B^{q,T}_{\omega,\rho,\sigma,\tau} \quad \text{and} \quad s_{|_{\tilde{T}}} = \tilde{p} = \sum_{\omega+\rho+\sigma+\tau=q} \tilde{b}_{\omega,\rho,\sigma,\tau}\ B^{q,\tilde{T}}_{\omega,\rho,\sigma,\tau}$$

be given. Then, s is differentiable across the common triangular face $T \cap \tilde{T} = [v_0, v_1, v_2]$, i.e., $s \in C^1(T \cup \tilde{T})$ if and only if for all $\omega+\rho+\sigma = q-1$,

$$\begin{aligned}\tilde{b}_{\omega,\rho,\sigma,1} = {} & b_{\omega+1,\rho,\sigma,0}\ \Phi_0(\tilde{v}_3) + b_{\omega,\rho+1,\sigma,0}\ \Phi_1(\tilde{v}_3) \\ & + b_{\omega,\rho,\sigma+1,0}\ \Phi_2(\tilde{v}_3) + b_{\omega,\rho,\sigma,1}\ \Phi_3(\tilde{v}_3),\end{aligned} \tag{2.4}$$

where $\Phi_\nu(\tilde{v}_3),\ \nu = 0,\dots,3$, are the barycentric coordinates with respect to T, evaluated at the point $\tilde{v}_3$.

The relations in (2.4) show that in order to guarantee the C^1 smoothness across $T \cap \tilde{T}$ exactly $q(q+1)/2$ conditions have to be satisfied. In particular, the relations imply that if coefficients of the form $b_{\omega,\rho,\sigma,\ell},\ \omega+\rho+\sigma = q-\ell,\ \ell = 0,1$, are determined, then the coefficients $\tilde{b}_{\omega,\rho,\sigma,1},\ \omega+\rho+\sigma = q-1$ (i.e., the coefficients *in distance one to* $T \cap \tilde{T}$) are uniquely determined. In general, there are five coefficients involved in each of these conditions (see Figure 2, left). Analogous to the univariate and bivariate case, each of these conditions has the geometric interpretation that there are five corresponding points in $\mathbb{R}^4$ which lie in the same (three-dimensional) hyperplane. The fourth component of these points is a Bernstein–Bézier coefficient which appears in (2.4) while the first three components are the associated domain points. If one or even two of the barycentric coordinates at the point $\tilde{v}_3$ vanish, then the number of involved coefficients is four and three, respectively (see Figure 2, middle and right). In these cases, the smoothness conditions degenerate to lower dimensional conditions which are similar as in the bivariate and univariate setting, i.e., four points of the above form lie in a plane and three points lie on a line, respectively. Therefore, these cases are sometimes called *degenerate cases*.

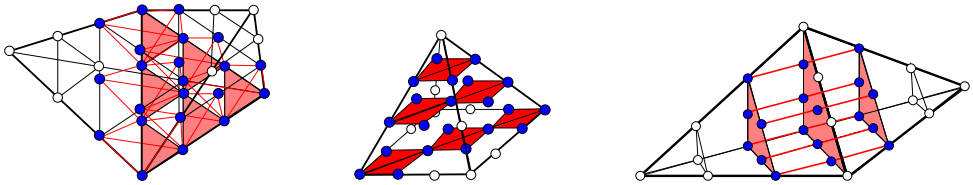


FIG. 2. *Illustration of C^1 smoothness conditions across the common triangular face of two neighboring tetrahedra (in the case $q = 3$). The domain points associated with coefficients involved are shown as dark dots. In general, there are five coefficients involved in every condition (left). In the special case when three of the five vertices of the neighboring tetrahedra lie on one line, the smoothness conditions degenerate to conditions of univariate type, which means that only three points are affected by every condition (right). For Freudenthal partitions Δ only four coefficients are relevant for each C^1 smoothness condition (middle), since for every pair of neighboring tetrahedra in Δ four of the five vertices lie in a plane.*

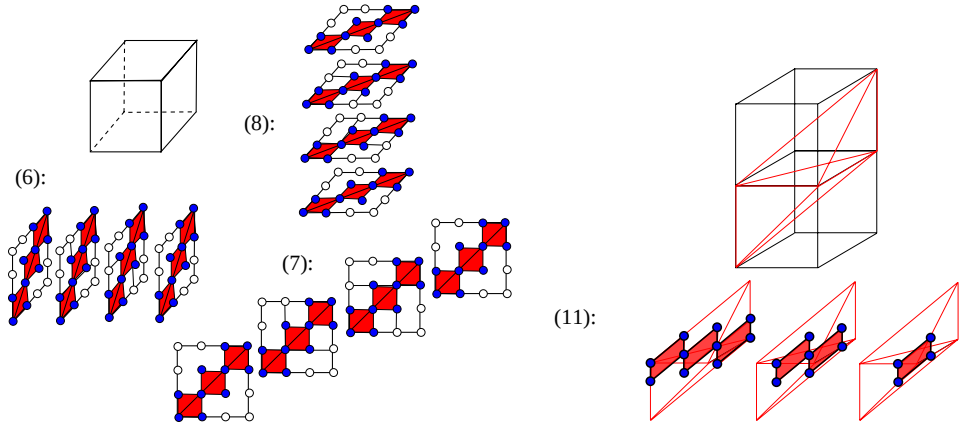


FIG. 3. *Each smoothness condition satisfied by the C^1 splines on Freudenthal partitions Δ degenerates to a C^1 condition of bivariate type. The example shows the case $q = 3$, where the domain points associated with the four coefficients involved in each smoothness condition are shown as dark dots. Conditions on neighboring tetrahedra inside a cube of $\diamondsuit$ (i.e., the conditions in* (6)–(8)*) are illustrated on the left, while an example illustrating the smoothness conditions on two neighboring tetrahedra from different cubes (i.e., the conditions in* (11)*) is shown on the right.*

While on general tetrahedral partitions these three types of smoothness conditions can appear, we observe that *all* the conditions on neighboring tetrahedra satisfied by the splines from $\mathcal{S}_q^1(\Delta)$, where Δ is a Freudenthal partition, are of the *same type*. More precisely, for every pair of neighboring tetrahedra in Δ exactly four Bernstein–Bézier coefficients are involved in each C^1 condition and hence only degenerate smoothness conditions of bivariate type appear for these splines. The following lemma shows that the smoothness conditions satisfied by these splines can easily be described by one single formula. However, the arguments given in the subsequent sections indicate that even for Δ the analysis of the overall smooth spline spaces is a complex task since the conditions have to be satisfied simultaneously across all the interior triangular faces of Δ. Figure 3 further illustrates the lemma.

LEMMA 2.2. *Let s be a continuous spline on a Freudenthal partition Δ with BB-coefficients $b_\xi(s) = b_\xi$, $\xi \in \mathcal{D}_q(\Delta)$ in its piecewise Bernstein–Bézier representation*

(2.3). *Then, $s \in \mathcal{S}_q^1(\Delta)$ if and only if $b_\xi + b_{\xi'} = b_{\xi^*} + b_{\xi''}$ for all ξ, ξ', ξ^*, ξ'', where*

$$\left(\xi = \xi^{Q_{(i,j,k)}}_{\beta,\alpha,\alpha},\ \xi' = \xi^{Q_{(i,j,k)}}_{\beta,\alpha+1,\alpha+1},\ \xi^* = \xi^{Q_{(i,j,k)}}_{\beta,\alpha+1,\alpha},\ \textit{and}\ \ \xi'' = \xi^{Q_{(i,j,k)}}_{\beta,\alpha,\alpha+1}\right), \tag{2.5}$$

$$\left(\xi = \xi^{Q_{(i,j,k)}}_{\alpha,\beta,\alpha},\ \xi' = \xi^{Q_{(i,j,k)}}_{\alpha+1,\beta,\alpha+1},\ \xi^* = \xi^{Q_{(i,j,k)}}_{\alpha+1,\beta,\alpha},\ \textit{and}\ \ \xi'' = \xi^{Q_{(i,j,k)}}_{\alpha,\beta,\alpha+1}\right), \tag{2.6}$$

$$\left(\xi = \xi^{Q_{(i,j,k)}}_{\alpha,\alpha,\beta},\ \xi' = \xi^{Q_{(i,j,k)}}_{\alpha+1,\alpha+1,\beta},\ \xi^* = \xi^{Q_{(i,j,k)}}_{\alpha+1,\alpha,\beta},\ \textit{and}\ \ \xi'' = \xi^{Q_{(i,j,k)}}_{\alpha,\alpha+1,\beta}\right), \tag{2.7}$$
$$\alpha, \beta = 0, \dots, q,\ \alpha \neq q,\ i, j, k = 1, \dots, n,$$

and

$$\left(\xi = \xi^{Q_{(i,j,k)}}_{q,\alpha,\beta},\ \xi' = \xi^{Q_{(i,j,k)}}_{q,\alpha+1,\beta+1},\ \xi^* = \xi^{Q_{(i,j,k)}}_{q-1,\alpha,\beta},\ \textit{and}\ \ \xi'' = \xi^{Q_{(i+1,j,k)}}_{1,\alpha+1,\beta+1}\right), \tag{2.8}$$

$$\left(\xi = \xi^{Q_{(j,i,k)}}_{\alpha,q,\beta},\ \xi' = \xi^{Q_{(j,i,k)}}_{\alpha+1,q,\beta+1},\ \xi^* = \xi^{Q_{(j,i,k)}}_{\alpha,q-1,\beta},\ \textit{and}\ \ \xi'' = \xi^{Q_{(j,i+1,k)}}_{\alpha+1,1,\beta+1}\right), \tag{2.9}$$

$$\left(\xi = \xi^{Q_{(j,k,i)}}_{\alpha,\beta,q},\ \xi' = \xi^{Q_{(j,k,i)}}_{\alpha+1,\beta+1,q},\ \xi^* = \xi^{Q_{(j,k,i)}}_{\alpha,\beta,q-1},\ \textit{and}\ \ \xi'' = \xi^{Q_{(j,k,i+1)}}_{\alpha+1,\beta+1,1}\right), \tag{2.10}$$
$$\alpha, \beta = 0, \dots, q-1,\ i, j, k = 1, \dots, n,\ i \neq n.$$

Proof of Lemma 2.2. Let $T = [v_0, v_1, v_2, v_3]$ and $\tilde{T} = [v_0, v_1, v_2, \tilde{v}_3]$ be two tetrahedra from Δ with a common triangular face $F = T \cap \tilde{T} = [v_0, v_1, v_2]$, where $v_3 \neq \tilde{v}_3$. It follows from the special structure of Δ that in all cases the vertices v_0, v_1, and v_2 of F can be arranged such that v_0, v_1, v_3, and $\tilde{v}_3$ lie in the same plane and form a parallelogram P with diagonal $[v_0, v_1]$. If T and $\tilde{T}$ are contained in two different cubes, then P lies within a plane of the form (2.1). Otherwise, P lies in a plane parallel to one of the three coordinate planes. In both cases, we compute

$$\Phi_0(\tilde{v}_3) = 1, \quad \Phi_1(\tilde{v}_3) = 1, \quad \Phi_2(\tilde{v}_3) = 0, \quad \text{and} \quad \Phi_3(\tilde{v}_3) = -1,$$

where $\Phi_\nu,\ \nu = 0, \dots, 3$, denote the barycentric coordinates w.r.t. T. Therefore, the assertion follows from Theorem 2.1. This proves the lemma. □

Remark 2.3. Combining the conditions in (2.5)–(2.7) which involve the coefficients along the main diagonal of the cube $Q_{(i,j,k)}$, we observe that for any C^1 spline on Δ the hidden conditions $b_\xi + 2\, b_{\xi'} = b_{\xi''} + b_{\xi^*} + b_{\tilde{\xi}}$ are satisfied, where

$$\begin{aligned}&\left(\xi = \xi^{Q_{(i,j,k)}}_{\alpha+1,\alpha+1,\alpha+1},\ \xi' = \xi^{Q_{(i,j,k)}}_{\alpha,\alpha,\alpha},\ \xi'' = \xi^{Q_{(i,j,k)}}_{\alpha+1,\alpha,\alpha},\ \xi^* = \xi^{Q_{(i,j,k)}}_{\alpha,\alpha+1,\alpha},\right.\\ &\left.\text{and}\ \ \tilde{\xi} = \xi^{Q_{(i,j,k)}}_{\alpha,\alpha,\alpha+1}\right), \qquad \alpha = 0, \dots, q-1, \quad i, j, k = 1, \dots, n.\end{aligned} \tag{2.11}$$

3. Minimal determining sets and main result. An analysis of the structure of smooth multivariate splines is frequently done by constructing minimal determining sets for the spaces. Following Alfeld, Piper and Schumaker [3] (see also Schumaker and Sorokina [35, 36]), we call $\mathcal{M} \subseteq \mathcal{D}_q(\Delta)$ a *determining set (DS) for* $\mathcal{S}_q^1(\Delta)$, if setting the coefficients $b_\xi(s),\ \xi \in \mathcal{M}$, of a spline $s \in \mathcal{S}_q^1(\Delta)$ to zero, implies that $s = 0$. A determining set $\mathcal{M}$ is called a *minimal determining set (MDS) for* $\mathcal{S}_q^1(\Delta)$ if no determining set for $\mathcal{S}_q^1(\Delta)$ with fewer elements than in $\mathcal{M}$ exists. Hence, it is obvious that $\mathcal{M}$ is an MDS for $\mathcal{S}_q^1(\Delta)$, if setting the coefficients $b_\xi(s),\ \xi \in \mathcal{M}$, of a spline $s \in \mathcal{S}_q^1(\Delta)$ to arbitrary real values, all its coefficients $b_\xi(s),\ \xi \in \mathcal{D}_q(\Delta)$, are uniquely determined (whereby all smoothness conditions described in (2.5)–(2.10) are satisfied). The cardinality of an MDS for $\mathcal{S}_q^1(\Delta)$ coincides with the dimension (i.e., the

TABLE 1
Dimensions of C^1 and C^0 splines on Freudenthal partitions Δ.

q	Dimension of $\mathcal{S}_q^1(\Delta)$	$n=1$	Dimension of $\mathcal{S}_q^0(\Delta)$	$n=1$
1	4	4	$n^3+\mathcal{O}(n^2)$	8
2	$9\,n+4$	13	$8\,n^3+\mathcal{O}(n^2)$	27
3	$12\,n^2+18\,n+4$	34	$27\,n^3+\mathcal{O}(n^2)$	64
4	$6\,n^3+36\,n^2+27\,n+4$	73	$64\,n^3+\mathcal{O}(n^2)$	125
5	$24\,n^3+72\,n^2+36\,n+4$	136	$125\,n^3+\mathcal{O}(n^2)$	216
6	$60\,n^3+120\,n^2+45\,n+4$	229	$216\,n^3+\mathcal{O}(n^2)$	343
7	$120\,n^3+180\,n^2+54\,n+4$	358	$343\,n^3+\mathcal{O}(n^2)$	512
8	$210\,n^3+252\,n^2+63\,n+4$	529	$512\,n^3+\mathcal{O}(n^2)$	729
9	$336\,n^3+336\,n^2+72\,n+4$	748	$729\,n^3+\mathcal{O}(n^2)$	1 000

number of degrees of freedom) of the spline space. Moreover, we note that it is well known in multivariate spline theory that the explicit construction of MDS gives insight on the structure of spline spaces. In section 5, we construct an MDS for C^1 splines of any degree on Freudenthal partitions Δ. From this construction (see Theorem 5.1), we obtain the following theorem, which we state here as our main result.

THEOREM 3.1. *Let Δ be a Freudenthal partition. Then the dimension of $\mathcal{S}_q^1(\Delta)$ is given by*

$$(q-3)(q-2)(q-1)\,n^3+6\,(q-2)(q-1)\,n^2+9\,(q-1)\,n+4. \tag{3.1}$$

Table 1 lists the dimension of $\mathcal{S}_q^1(\Delta)$ for low degrees, i.e., $q\in\{1,\dots,9\}$, for arbitrary n and $n=1$ (i.e., the case of one cube), and compares them with the asymptotic number of degrees of freedom as well as with the case $n=1$ for the continuous spline spaces on Δ of the same degree.

In the next corollary, we give some alternative formulae for the dimension of $\mathcal{S}_q^1(\Delta)$. Recall that for Freudenthal partitions Δ the quantities $N_\Delta, F_\Delta, E_\Delta$, and V_Δ are given as in (2.2). In addition, let V_I be the number of interior vertices of Δ, V_B the number of boundary vertices of Δ, F_I the number of interior triangular faces of Δ, and E_I the number of interior edges of Δ. An elementary computation shows that we have

$$V_I=n^3-3\,n^2+3\,n-1,\quad V_B=6\,n^2+2,\quad F_I=12\,n^3-6\,n^2,$$

and

$$E_I=7\,n^3-9\,n^2+3\,n.$$

Since Δ can be completely constructed by starting with an appropriate tetrahedron and adding successively tetrahedra which intersect the union of the tetrahedra considered before along one, two or three triangular faces, the *Euler-type formulae*

$$V_B\;=\;2\,N_\Delta-F_I+2,\qquad N_\Delta\;=\;V_I-E_I+F_I+1,$$

are satisfied (see the discussion on *shellable tetrahedral partitions* in Lai and Schumaker [21]). By using these relations, the formula in (3.1), and some elementary computations, we obtain the following result.

COROLLARY 3.2. *Let Δ be a Freudenthal partition. Then, the dimension of $\mathcal{S}_q^1(\Delta)$ is given by*

$$\begin{aligned}
&\tfrac{1}{6}\,\Big(24\;V_\Delta + 6\;(3\;q-7)\;E_\Delta + 3\;(2\;q^2-15\;q+21)\;F_\Delta \\
&\qquad + (q^3-18\;q^2+80\;q-87)\;N_\Delta\Big) \\
&= \tfrac{1}{6}\,\Big(18\;(q-1)\;V_I + 3\;(2\;q^2-3\;q+1)\;V_B + (q^3-6\;q^2+8\;q-3)\;N_\Delta \\
&\qquad - 12\;(q^2-3\;q)\Big) \\
&= \tfrac{1}{6}\,\Big((q^3+6\;q^2+8\;q-15)\;V_I - (q^3+6\;q^2-10\;q+3)\;E_I + \\
&\qquad (q^3-q)\;F_I + (q^3+6\;q^2+8\;q+9)\Big).
\end{aligned}$$

4. MDS for C^1 splines on a tetrahedrized cube. In this section, we deal with C^1 splines of arbitrary degree on the Freudenthal partition Δ_Q which is obtained by subdividing the single cube $Q = Q_{(1,1,1)}$ into six tetrahedra, i.e., the case $n = 1$. This can be considered as the starting point of our inductive construction of an MDS for $\mathcal{S}_q^1(\Delta)$, which is given in the next section. In the following, we construct two different MDSs for the spaces $\mathcal{S}_q^1(\Delta_Q)$, which we denote by $\widetilde{\mathcal{M}}_Q$ and $\mathcal{M}_Q$, respectively. We first use the set $\widetilde{\mathcal{M}}_Q$ to determine the number of degrees of freedom of the smooth splines on Δ_Q. For defining this set, we consider the tetrahedra T_i of Δ_Q and choose appropriate subsets of $\mathcal{D}_{q,T_i},\ i = 1,\dots,6$. The definition of the second set $\mathcal{M}_Q$ requires the notation introduced in section 2 since its points are not chosen tetrahedron by tetrahedron. (We want to keep the number of symbols as small as possible, and therefore we describe both sets by using the above notation.) The set $\mathcal{M}_Q$ is more complex than $\widetilde{\mathcal{M}}_Q$, but at that point we only have to show that $\mathcal{M}_Q$ is a DS, since the cardinalities of $\mathcal{M}_Q$ and $\widetilde{\mathcal{M}}_Q$ coincide. Note that the construction of $\mathcal{M}_Q$ is the key of the proof of our main result (see section 5).

In the following, we set $\xi_{\alpha,\beta,\gamma} = \xi^Q_{\alpha,\beta,\gamma},\ \alpha,\beta,\gamma = 0,\dots,q$, and define $\widetilde{\mathcal{M}}_Q \subseteq \mathcal{D}_q(\Delta_Q)$. To do this, we consider six auxiliarly sets $\widetilde{\mathcal{M}}_i,\ i = 1,\dots,6$, which are given as follows:

$$\begin{aligned}
\widetilde{\mathcal{M}}_1 &= \{\xi_{\alpha,\beta,\gamma} :\ \gamma \le \beta \le \alpha\},\\
\widetilde{\mathcal{M}}_2 &= \{\xi_{\alpha,\beta,\gamma} :\ \beta+2 \le \gamma \le \alpha\},\\
\widetilde{\mathcal{M}}_3 &= \{\xi_{\alpha,\beta,\gamma} :\ \beta \le \alpha \le \gamma-2\},\\
\widetilde{\mathcal{M}}_4 &= \{\xi_{\alpha,\beta,\gamma} :\ \gamma \le \alpha \le \beta-2\},\\
\widetilde{\mathcal{M}}_5 &= \{\xi_{\alpha,\beta,\gamma} :\ \alpha+2 \le \gamma < \beta\} \cup \{\xi_{\alpha,\beta,\beta} :\ \alpha+3 \le \beta\},\\
\widetilde{\mathcal{M}}_6 &= \{\xi_{\alpha,\beta,\gamma} :\ \alpha+2 \le \beta \le \gamma-2\}\ .
\end{aligned}$$

In these definitions as well as in what follows, the indices which are not further specified run over all possible choices from $\{0,\dots,q\}$. It is easy to see that $\widetilde{\mathcal{M}}_1$ is equal to $\mathcal{D}_{q,T_1}$, where T_1 is the tetrahedron of Δ_Q with the vertices $(0,0,0)$, $(1,0,0)$, $(1,1,0)$, and $(1,1,1)$. The set $\widetilde{\mathcal{M}}_2$ is obtained from $\mathcal{D}_{q,T_2} = \{\xi_{\alpha,\beta,\gamma} :\ \beta \le \gamma \le \alpha\}$, where T_2 is the tetrahedron of Δ_Q with the vertices $(0,0,0)$, $(1,0,0)$, $(1,0,1)$, and $(1,1,1)$, by removing the domain points in $\mathcal{D}_{q,T_2}$ on the common triangular face of T_2 and T_1 as well as the points in $\mathcal{D}_{q,T_2}$ associated with coefficients in distance one to

this face. The set $\widetilde{\mathcal{M}}_4$ is similar to $\widetilde{\mathcal{M}}_2$: here we remove the analogous points from $\mathcal{D}_{q,T_4} = \{\xi_{\alpha,\beta,\gamma} : \gamma \le \alpha \le \beta\}$, where T_4 is the tetrahedron of Δ_Q with vertex $(0,1,0)$, which has the triangular face with vertices $(0,0,0)$, $(1,1,0)$, and $(1,1,1)$ in common with T_1. Similarly, the set $\widetilde{\mathcal{M}}_3$ is obtained from $\mathcal{D}_{q,T_3} = \{\xi_{\alpha,\beta,\gamma} : \beta \le \alpha \le \gamma\}$, where T_3 is the tetrahedron of Δ_Q with the vertices $(0,0,0)$, $(0,0,1)$, $(1,0,1)$, and $(1,1,1)$, by removing the domain points in T_3 on the common triangular face of T_3 and T_2 as well as the points in T_3 associated with coefficients in distance one to this face. The sets $\widetilde{\mathcal{M}}_5$ and $\widetilde{\mathcal{M}}_6$ are subsets of $\mathcal{D}_{q,T_5} = \{\xi_{\alpha,\beta,\gamma} : \alpha \le \gamma \le \beta\}$ and $\mathcal{D}_{q,T_6} = \{\xi_{\alpha,\beta,\gamma} : \alpha \le \beta \le \gamma\}$, where T_5 is the tetrahedron of Δ_Q with the vertices $(0,0,0)$, $(0,1,0)$, $(0,1,1)$, and $(1,1,1)$, and T_6 is the tetrahedron of Δ_Q with the vertices $(0,0,0)$, $(0,0,1)$, $(0,1,1)$, and $(1,1,1)$, respectively: here, we remove similar points as above, and some additional points on their common triangular face with vertices $(0,0,0)$, $(0,1,1)$, and $(1,1,1)$, i.e., the points in distance two to the main diagonal in Q.

We set

$$\widetilde{\mathcal{M}}_Q \;=\; \bigcup_{i=1}^{6} \widetilde{\mathcal{M}}_i.$$

LEMMA 4.1. *Let Δ_Q be the Freudenthal partition of $Q = Q_{(1,1,1)}$. Then, $\widetilde{\mathcal{M}}_Q$ is an MDS for $\mathcal{S}_q^1(\Delta_Q)$ and the dimension of $\mathcal{S}_q^1(\Delta_Q)$ is given by $q^3 + 2\,q + 1$.*

Proof. Let arbitrary coefficients $b_\xi = b_\xi(s),\ \xi \in \widetilde{\mathcal{M}}_Q$ of a spline $s \in \mathcal{S}_q^1(\Delta_Q)$ be given. We have to show that the remaining coefficients of s, i.e., the coefficients b_ξ, where $\xi \in \mathcal{D}_q(\Delta_Q) \setminus \widetilde{\mathcal{M}}_Q$, are uniquely determined, while the C^1 smoothness conditions involving the coefficients associated with each choice of four points in (2.5)–(2.7) are satisfied. In the following, we use the above notation for the tetrahedra $T_1, \ldots, T_6$ in Δ_Q. First, it is obvious that the choice of $\widetilde{\mathcal{M}}_1 \subseteq \widetilde{\mathcal{M}}_Q$ implies that the polynomial piece $s_{|_{T_1}}$ is uniquely determined. In particular, the coefficients b_ξ, where $\xi = \xi_{\beta,\alpha,\alpha} \in T_1,\ \alpha \le \beta$, and $\xi = \xi_{\beta,\alpha+1,\alpha} \in T_1,\ \alpha + 1 \le \beta$, are uniquely determined. In view of (2.5), it follows that the coefficients b_ξ, where $\xi = \xi_{\beta,\alpha,\alpha+1} \in T_2,\ \alpha + 1 \le \beta$, are uniquely determined. The choice of $\widetilde{\mathcal{M}}_2$ therefore implies that $s_{|_{T_2}}$ is uniquely determined. Similarly, the smoothness conditions in (2.6) for the neighboring polynomial pieces on T_2 and T_3 determine the coefficients b_ξ, where $\xi = \xi_{\alpha,\beta,\alpha+1} \in T_3,\ \beta \le \alpha$, and it follows that the choice of $\widetilde{\mathcal{M}}_3$ uniquely determines $s_{|_{T_3}}$. Independently, the smoothness conditions in (2.7) for the neighboring polynomial pieces on T_1 and T_4 determine the coefficients b_ξ, where $\xi = \xi_{\alpha,\alpha+1,\beta} \in T_4,\ \beta \le \alpha$, and hence, the choice of $\widetilde{\mathcal{M}}_4$ shows that $s_{|_{T_4}}$ is uniquely determined. Moreover, the remaining smoothness conditions in (2.6) for the neighboring polynomial pieces on T_4 and T_5 determine the coefficients b_ξ, where $\xi = \xi_{\alpha,\beta,\alpha+1} \in T_5,\ \alpha + 1 \le \beta$, and the remaining smoothness conditions in (2.7) for the neighboring polynomial pieces of T_3 and T_6 determine the coefficients b_ξ, where $\xi = \xi_{\alpha,\alpha+1,\beta} \in T_6,\ \alpha + 1 \le \beta$. Now, we consider the remaining smoothness conditions in (2.5) for the neighboring polynomial pieces on T_5 and T_6 in the case $\alpha = \beta + 1$, i.e.,

$$b_{\xi_{\beta,\beta+2,\beta+2}} = b_{\xi_{\beta,\beta+2,\beta+1}} + b_{\xi_{\beta,\beta+1,\beta+2}} - b_{\xi_{\beta,\beta+1,\beta+1}}\ .$$

It follows from above that the coefficients on the right of this equation are already uniquely determined, and hence $b_{\xi_{\beta,\beta+2,\beta+2}}$ is uniquely determined. The choice of $\widetilde{\mathcal{M}}_5$ now uniquely determines $s_{|_{T_5}}$. Finally, we get that the remaining smoothness

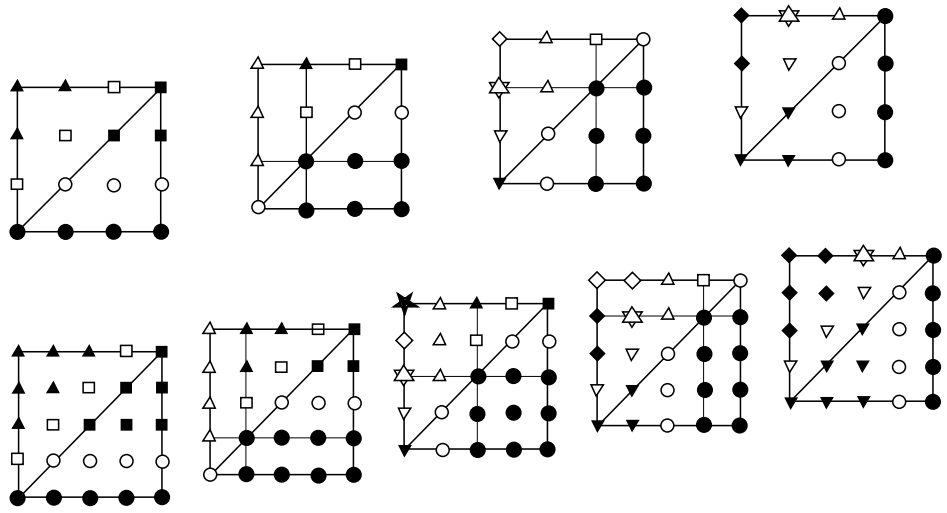


FIG. 4. *Illustration of the MDS $\widetilde{\mathcal{M}}_Q$ for $\mathcal{S}_q^1(\Delta_Q)$ in the cases $q = 3$ (top) and $q = 4$ (bottom). The points of $\widetilde{\mathcal{M}}_Q$ are marked by black symbols, while the white symbols indicate the order of determining the remaining coefficients in the proof of Lemma* 4.1.

conditions in (2.5) for the neighboring polynomial pieces on T_5 and T_6 and the choice of $\widetilde{\mathcal{M}}_6$ imply that $s_{|_{T_6}}$ is uniquely determined, and therefore all the coefficients b_ξ of s are uniquely determined. It is easy to see that the number of points $\widetilde{\mathcal{M}}_i$ is equal to $\binom{q+3}{3} = (q^3 + 6q^2 + 11q + 6)/6$ if $i = 1$, $\binom{q+1}{3} = (q^3 - q)/6$ if $i \in \{2, 3, 4\}$, $\binom{q+1}{3} - (q-1) = (q^3 - 7q + 6)/6$ if $i = 5$, and $\binom{q-1}{3} = (q^3 - 6q^2 + 11q - 6)/6$ if $i = 6$. (Note that we set here and in the following $\binom{j}{k} := 0$ whenever $j < k$.) Therefore, it follows from some elementary computations that the number of points in $\widetilde{\mathcal{M}}_Q$ is equal to $q^3 + 2\,q + 1$. This completes the proof of the lemma. ∎

Example 4.2. In Figure 4, we show the MDS $\widetilde{\mathcal{M}}_Q$ in the cases $q = 3$ (top) and $q = 4$ (bottom). The figure illustrates the different layers of domain points within the cube Q in a front to back manner, where the thin lines indicate the intersections with triangular faces in the interior of Q. In these examples the set $\widetilde{\mathcal{M}}_Q$ contains 34 and 73 points, respectively, which we illustrate by black symbols, i.e., we mark the points from the sets $\widetilde{\mathcal{M}}_i$, $i = 1, \ldots, 6$, by ●, ■, ▲, ▼, ♦ and ★. In the proof of Lemma 4.1, the coefficients associated with the remaining domain points are determined in the order of the white symbols, i.e., ○, □, △, ▽, ⯐ , and ◇. Note that the coefficients indicated by ⯐ are determined from △ and ▽ by using conditions described in (2.5).

We proceed by defining a subset $\mathcal{M}_Q$ of $\mathcal{D}_q(\Delta_Q)$ different from $\widetilde{\mathcal{M}}_Q$ which is also an MDS for $\mathcal{S}_q^1(\Delta_Q)$, but more symmetric. Again, we use the above abbreviations and consider three auxiliary sets $\mathcal{M}_i$, $i = 1, \ldots, 3$, which are defined as follows:

$$\begin{aligned}
\mathcal{M}_1 &= \{\xi_{0,0,0}\} \cup \{\xi_{0,\alpha,\beta},\ \xi_{\alpha,0,\beta},\ \xi_{\alpha,\beta,0} :\ \alpha \neq \beta\},\\
\mathcal{M}_2 &= \{\xi_{\alpha,\beta,\gamma},\ \xi_{\beta,\alpha,\gamma},\ \xi_{\beta,\gamma,\alpha} :\ 1 \leq \alpha \leq q-3,\ \alpha+1 \leq \beta, \gamma \leq q-1,\ \beta \neq \gamma\},\\
\mathcal{M}_3 &= \{\xi_{q,\alpha,\beta},\ \xi_{\alpha,q,\beta},\ \xi_{\alpha,\beta,q} :\ 1 \leq \alpha, \beta \leq q-1,\ \alpha \neq \beta\}.
\end{aligned}$$

The set $\mathcal{M}_1$ consists of all domain points contained in the left, front, and bottom face of Q except for those which differ from $\xi_{0,0,0}$ and lie on a diagonal of such a face. The set $\mathcal{M}_2$ represents a symmetric constellation of points in the interior of Q. Moreover,

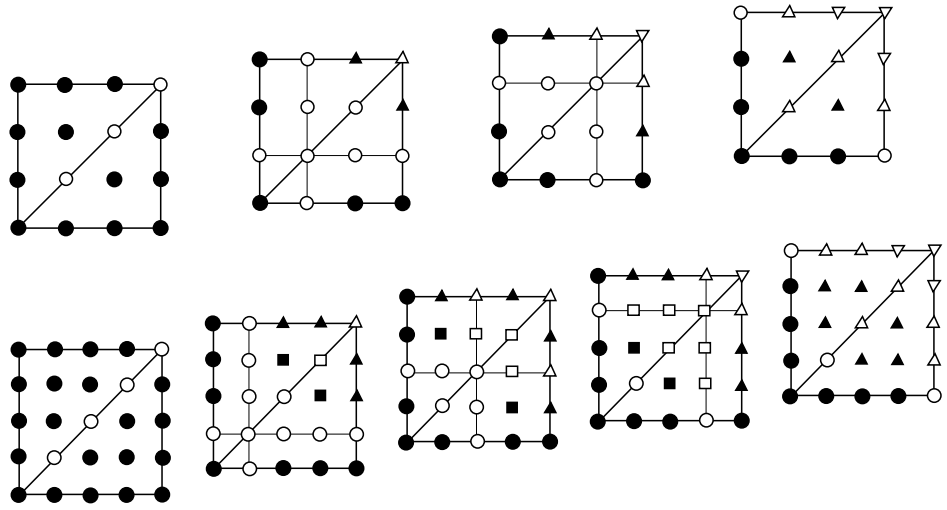


FIG. 5. *Illustration of the MDS $\mathcal{M}_Q$ for $\mathcal{S}_q^1(\Delta_Q)$ in the cases $q = 3$ (top) and $q = 4$ (bottom). The points from $\mathcal{M}_Q$ are marked by black symbols, while the white symbols indicate the order of determining the remaining coefficients in the proof of Theorem* 4.3.

$\mathcal{M}_3$ includes certain domain points contained in the right, back, and top face of Q. This choice of points is illustrated in Figure 5 for the cases $q = 3$ (top) and $q = 4$ (bottom). Here, we use the symbols ●, ■, and ▲ for the points in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, respectively. The next theorem shows that

$$\mathcal{M}_Q \;=\; \bigcup_{i=1}^{3} \mathcal{M}_i$$

is an MDS for $\mathcal{S}_q^1(\Delta_Q)$. We note that the construction of $\mathcal{M}_Q$ is the key for showing our main result (see next section). By using Lemma 4.1 its proof is essentially simplified, since it now suffices to show that $\mathcal{M}_Q$ contains the right number of points and is a determining set. Following the proof of the next theorem, it can be seen that the coefficients of the splines associated with the remaining domain points indicated as white symbols in Figure 5 are determined in the order ○, □, △, and ▽.

THEOREM 4.3. *Let Δ_Q be the Freudenthal partition of $Q = Q_{(1,1,1)}$. Then, $\mathcal{M}_Q$ is an MDS for $\mathcal{S}_q^1(\Delta_Q)$.*

Proof. The number of points in $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{M}_3$ is equal to $3q^2 + 1$, $q^3 - 6q^2 + 11q - 6$, and $3(q^2 - 3q + 2)$, respectively. Hence, some elementary computations show that the number of points in $\mathcal{M}_Q$ is equal to $q^3 + 2\ q + 1$. Therefore, in view of Lemma 4.1, it suffices to show that $\mathcal{M}_Q$ is a DS. We may assume that $q \geq 2$ and have to show that for any spline $s \in \mathcal{S}_q^1(\Delta_Q)$, with coefficients $b_\xi = b_\xi(s) = 0$, where $\xi \in \mathcal{M}_Q$, the remaining coefficients $b_\xi(s)$, $\xi \in \mathcal{D}_q(\Delta_Q) \setminus \mathcal{M}_Q$, are also zero, i.e., $s = 0$.

First, we note that it follows from an inductive argument involving the smoothness conditions in (2.5) for the case $\beta = 0$ that $b_\xi = 0$, where $\xi = \xi_{0,\alpha,\alpha},\ \alpha = 1, \ldots, q$. Analogously, we get $b_\xi = 0$, where $\xi = \xi_{\alpha,0,\alpha}$ and $\xi = \xi_{\alpha,\alpha,0},\ \alpha = 1, \ldots, q$, by using (2.6) and (2.7), respectively. The choice of $\mathcal{M}_1$ therefore implies that $b_\xi = 0$, where $\xi = \xi_{0,\alpha,\beta},\ \xi = \xi_{\alpha,0,\beta}$, and $\xi = \xi_{\alpha,\beta,0},\ \alpha, \beta = 0, \ldots, q$. We now show inductively that for any $m \in \{1, \ldots, q-1\}$, we have $b_\xi = 0$, where $\xi = \xi_{m,\alpha,\beta}, \xi = \xi_{\alpha,m,\beta}$, and $\xi = \xi_{\alpha,\beta,m},\ \alpha, \beta = m, \ldots, q-1$. Let us assume that this has already been shown

for $m-1 \le q-2$. Then, it follows from induction hypothesis that $b_\xi = 0$, where $\xi = \xi_{\alpha,\beta,\gamma}$ and $\alpha = m-1$ or $\beta = m-1$ or $\gamma = m-1$. The smoothness conditions in (2.5) for the cases $\alpha = m-1$ and $\beta = m, \ldots, q-1$, now read as follows:

$$b_{\xi_{\beta,m,m}} = b_{\xi_{\beta,m,m-1}} + b_{\xi_{\beta,m-1,m}} - b_{\xi_{\beta,m-1,m-1}},$$

and we therefore get $b_\xi = 0$, where $\xi = \xi_{\beta,m,m},\ \beta = m, \ldots, q-1$. Analogously, we get $b_\xi = 0$, where $\xi = \xi_{m,\beta,m}$ and $\xi = \xi_{m,m,\beta},\ \beta = m+1, \ldots, q-1$, by using (2.6) and (2.7), respectively. In particular, $b_\xi = 0$, where $\xi = \xi_{m,m,m}$, $\xi = \xi_{m,m+1,m}$, and $\xi = \xi_{m,m,m+1}$. Thus, (2.5) in the case $\alpha = m+1$ and $\beta = m$ gives $b_\xi = 0$, where $\xi = \xi_{m,m+1,m+1}$. Similarly, $b_\xi = 0$, where $\xi = \xi_{m+1,m,m+1}$ and $\xi = \xi_{m+1,m+1,m}$. The points $\xi_{m,\alpha,\alpha+1}$ and $\xi_{m,\alpha+1,\alpha},\ \alpha = m+1, \ldots, q-2$, are contained in $\mathcal{M}_2$, and therefore a similar inductive argument as above involving the smoothness conditions in (2.5) for the case $\beta = m$ now shows that $b_\xi = 0$, where $\xi = \xi_{m,\alpha,\alpha},\ \alpha = m+2, \ldots, q-1$. Similarly, $b_\xi = 0$, where $\xi = \xi_{\alpha,m,\alpha}$ and $\xi = \xi_{\alpha,\alpha,m},\ \alpha = m+2, \ldots, q-1$. The choice of the remaining points in $\mathcal{M}_2$ now shows that the assertion holds true for m.

We conclude that $b_\xi = 0$ for all domain points ξ which lie in the interior of Q.

It remains to consider the coefficients associated with points on the right, back, and top face of Q. By using the conditions in (2.5)–(2.7) in the case $\alpha = 0$ and $\beta = q$, we can see that $b_\xi = 0$, where $\xi = \xi_{q,1,1},\ \xi = \xi_{1,q,1}$, and $\xi = \xi_{1,1,q}$. The points $\xi_{q,\alpha,\alpha+1}$ and $\xi_{q,\alpha+1,\alpha},\ \alpha = 1, \ldots, q-2$, are contained in $\mathcal{M}_3$, and therefore a similar inductive argument as above involving the smoothness conditions in (2.5) for the case $\beta = q$ now shows that $b_\xi = 0$, where $\xi = \xi_{q,\alpha,\alpha},\ \alpha = 1, \ldots, q-1$. Similarly, $b_\xi = 0$, where $\xi = \xi_{\alpha,q,\alpha}$ and $\xi = \xi_{\alpha,\alpha,q},\ \alpha = 1, \ldots, q-1$. Now, we consider (2.5) in the cases $\alpha = q-1$ and $\beta = 1, \ldots, q-1$, i.e.,

$$b_{\xi_{\beta,q,q}} = b_{\xi_{\beta,q,q-1}} + b_{\xi_{\beta,q-1,q}} - b_{\xi_{\beta,q-1,q-1}}.$$

As one can see from the above, the coefficients on the right of this equation are zero, and hence we have $b_\xi = 0$, where $\xi = \xi_{\beta,q,q},\ \beta = 1, \ldots, q-1$. Similarly, $b_\xi = 0$, where $\xi = \xi_{q,\beta,q}$ and $\xi = \xi_{q,q,\beta},\ \beta = 1, \ldots, q-1$. Finally, (2.11) gives

$$b_{\xi_{q,q,q}} = b_{\xi_{q,q-1,q-1}} + b_{\xi_{q-1,q,q-1}} + b_{\xi_{q-1,q-1,q}} - 2\, b_{\xi_{q-1,q-1,q-1}} = 0,$$

and we conclude that $s = 0$. This proves the theorem. □

We close this section with an explicit example.

Example 4.4. The space $\mathcal{S}_2^1(\Delta_Q)$ is 13-dimensional. In Figure 6 we show the MDS $\mathcal{M}_Q$ with coefficients $a_i,\ i = 0, \ldots, 12$. The remaining coefficients $b_i,\ i = 0, \ldots, 13$, are uniquely determined as follows:

$$\begin{aligned}
b_0 &= -a_0 + a_1 + a_5, & b_1 &= a_0 - a_1 - a_5 + a_7 + a_8,\\
b_2 &= -a_0 + a_3 + a_5, & b_3 &= a_0 - a_3 - a_5 + a_9 + a_{10},\\
b_4 &= -a_0 + a_1 + a_3, & b_5 &= a_0 - a_1 - a_3 + a_{11} + a_{12},\\
b_6 &= -2a_0 + a_1 + a_3 + a_5, & b_7 &= -a_2 + a_7 + a_{11},\\
b_8 &= -a_4 + a_{10} + a_{12}, & b_9 &= -a_6 + a_8 + a_9,\\
b_{10} &= 2a_0 - a_1 - a_2 - a_3 - a_4 - a_5 + a_7 + a_{10} + a_{11} + a_{12},\\
b_{11} &= 2a_0 - a_1 - a_3 - a_4 - a_5 - a_6 + a_8 + a_9 + a_{10} + a_{12},\\
b_{12} &= 2a_0 - a_1 - a_2 - a_3 - a_5 - a_6 + a_7 + a_8 + a_9 + a_{11},\\
b_{13} &= 4a_0 - 2a_1 - a_2 - 2a_3 - a_4 - 2a_5 - a_6 + a_7 + a_8 + a_9 + a_{10} + a_{11} + a_{12}.
\end{aligned}$$

5. An MDS for $\mathcal{S}_q^1(\Delta)$ and proof of main result. We construct an MDS $\mathcal{M}$ for $\mathcal{S}_q^1(\Delta)$, where Δ is a Freudenthal partition obtained from n^3 cubes $Q_{(i,j,k)},\ i,j,k =$

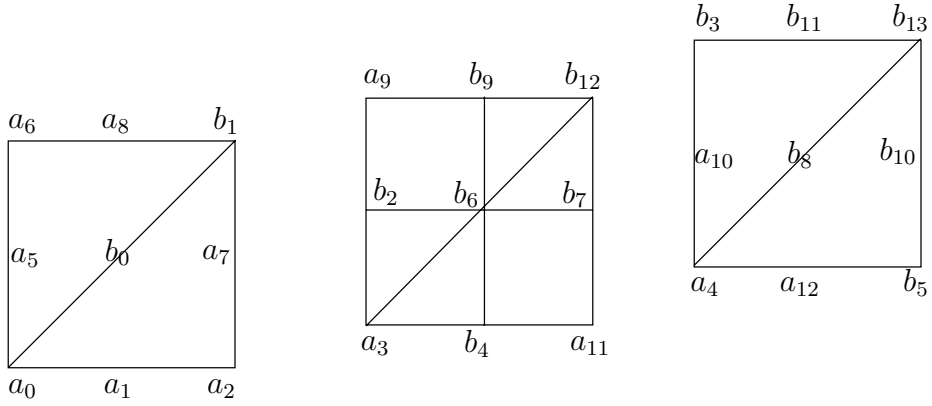


FIG. 6. *Illustration of the case of quadratic C^1 splines on Δ_Q. The points associated with the coefficients a_i, $i = 0, \ldots, 12$, are an MDS for this space.*

$1, \ldots, n$, (see section 2). To do this, we use the results from the previous section. In particular, we use Theorem 4.3 dealing with the case $n = 1$, which shows that $\mathcal{M}_Q$ is an MDS for $\mathcal{S}_q^1(\Delta_Q)$, where $Q = Q_{(1,1,1)}$. In the following, we define $\mathcal{M}$. To do this, we need three auxiliary sets $\mathcal{A}_{(i,j,k)}$, $\mathcal{B}_{(i,j,k)}$, $\mathcal{C}_{(i,j,k)}$, which are based on the construction of $\mathcal{M}_Q$. For $i, j, k = 1, \ldots, n$, we let

$$\mathcal{M}_{(i,j,k)} \;=\; \{\xi \in \mathcal{D}_{q,Q_{(i,j,k)}} : \xi - (i-1, j-1, k-1) \in \mathcal{M}_Q\} \tag{5.1}$$

be the "shifted versions" of the set $\mathcal{M}_Q$. Then, we let

$$\begin{aligned}
\mathcal{A}_{(i,j,k)} &= \{\xi_{\alpha,\beta,\gamma} \in \mathcal{M}_{(i,j,k)} :\ \alpha \geq 2\},\\
\mathcal{B}_{(i,j,k)} &= \{\xi_{\alpha,\beta,\gamma} \in \mathcal{M}_{(i,j,k)} :\ \beta \geq 2\},\\
\mathcal{C}_{(i,j,k)} &= \{\xi_{\alpha,\beta,\gamma} \in \mathcal{M}_{(i,j,k)} :\ \gamma \geq 2\},
\end{aligned}$$

for all $i, j, k = 1, \ldots, n$, where here and in the following for simplicity, we set

$$\xi_{\alpha,\beta,\gamma} := \xi_{\alpha,\beta,\gamma}^{Q_{(i,j,k)}} \in \mathcal{D}_{q,Q_{(i,j,k)}}, \qquad \alpha, \beta, \gamma = 0, \ldots, q.$$

Roughly speaking, $\mathcal{A}_{(i,j,k)}$ consists of all domain points in $Q_{(i,j,k)}$, which are not contained or in distance one to the two triangles which form the left square face of $Q_{(i,j,k)}$. Similarly, $\mathcal{B}_{(i,j,k)}$ and $\mathcal{C}_{(i,j,k)}$ consist of all points in $\mathcal{D}_{q,Q_{(i,j,k)}}$, which are not contained or in distance one to the two triangles which form the front and bottom square face of $Q_{(i,j,k)}$, respectively. Now, we define

$$\begin{aligned}
\mathcal{M} = \mathcal{M}_{(1,1,1)} &\cup \bigcup_{i=2}^{n} \Big[\mathcal{A}_{(i,1,1)} \cup \mathcal{B}_{(1,i,1)} \cup \mathcal{C}_{(1,1,i)}\Big]\\
&\cup \bigcup_{i,j=2}^{n} \Big[\big(\mathcal{A}_{(i,j,1)} \cap \mathcal{B}_{(i,j,1)}\big) \cup \big(\mathcal{A}_{(i,1,j)} \cap \mathcal{C}_{(i,1,j)}\big) \cup \big(\mathcal{B}_{(1,i,j)} \cap \mathcal{C}_{(1,i,j)}\big)\Big]\\
&\cup \bigcup_{i,j,k=2}^{n} \Big[\mathcal{A}_{(i,j,k)} \cap \mathcal{B}_{(i,j,k)} \cap \mathcal{C}_{(i,j,k)}\Big].
\end{aligned}$$

The next theorem particuarly shows that Theorem 3.1 holds true.

THEOREM 5.1. *Let Δ be a Freudenthal partition. Then, $\mathcal{M}$ is an MDS for $\mathcal{S}_q^1(\Delta)$ and the cardinality of $\mathcal{M}$ coincides with the number in* (3.1).

Proof. Let $q \geq 2$ and arbitrary coefficients $b_\xi = b_\xi(s),\ \xi \in \mathcal{M}$, of a spline $s \in \mathcal{S}_q^1(\Delta)$, be given. We first have to show that the remaining coefficients of s, i.e., the coefficients in $\mathcal{D}_q(\Delta) \setminus \mathcal{M}$, are uniquely determined, while all C^1 smoothness conditions described in Lemma 2.2 are satisfied.

Our method of proof is to show inductively that the coefficients $b_\xi,\ \xi \in \mathcal{D}_{q,Q_{(i,j,k)}}$, are uniquely determined for $i,j,k = 1,\dots,n$, where we use Theorem 4.3 and pass through the cubes $Q_{(i,j,k)}$ in an appropriate order. This natural order is as follows. First, we consider $(i,1,1),\ i = 1,\dots,n$. Then, we consider $(1,i,1)$ and $(1,1,i),\ i = 2,\dots,n$. Here, we have to take the C^1 continuity across exactly one square face of the cubes into account. We proceed by considering $(i,j,1)$, $(i,1,j)$, and $(1,i,j),\ i,j = 2,\dots,n$. Now, we have to take the C^1 continuity across exactly two square faces of the cube (which have a common edge) into account. Finally, we consider $(i,j,k),\ i,j,k = 2,\dots,n$. This is the most difficult situation, where we have to take the C^1 continuity across exactly three square faces of the cubes (which have a common point) into account.

Since $\mathcal{M}_{(1,1,1)} = \mathcal{M}_Q \subseteq \mathcal{M}$, it follows from Theorem 4.3 that all the coefficients b_ξ, where $\xi \in \mathcal{D}_{q,Q_{(1,1,1)}}$, are uniquely determined. Now, we consider the cube $Q_{(2,1,1)}$. This cube has exactly one square face in common with $Q_{(1,1,1)}$. This square face contains two triangular faces. Obviously, the coefficients associated with domain points within these triangular faces are uniquely determined. Moreover, since the coefficients b_ξ, where $\xi \in \{\xi_{q-1,\alpha,\beta},\ \xi_{q,\alpha,\beta},\ \xi_{q,\alpha+1,\beta+1}\} \subseteq \mathcal{D}_{q,Q_{(1,1,1)}}$, are uniquely determined, it follows from the smoothness conditions in (2.8) that the coefficients b_ξ, where $\xi = \xi_{1,\alpha+1,\beta+1} \in \mathcal{D}_{q,Q_{(2,1,1)}},\ \alpha,\beta = 0,\dots,q-1$, are determined. Note that an elementary computation shows that the conditions (2.5) for $\beta = 1$ and $\alpha = 1,\dots,q-1$ (w.r.t. $Q_{(2,1,1)}$) are automatically satisfied, and therefore the above coefficients are also uniquely determined. In particular, the coefficients b_ξ, where $\xi = \xi_{1,\beta,1}$ and $\xi = \xi_{1,1,\beta},\ \beta = 1,\dots,q$, are uniquely determined. The smoothness conditions (2.6) and (2.7) for $\alpha = 0$ (w.r.t. $Q_{(2,1,1)}$) now imply that the coefficients b_ξ, where $\xi = \xi_{1,\beta,0}$ and $\xi = \xi_{1,0,\beta},\ \beta = 1,\dots,q$, respectively, are uniquely determined. Finally, we use some elementary computations to show that an application of the smoothness condition (2.5) for $\alpha = 0$ and $\beta = 1$, as well as the smoothness conditions (2.6) and (2.7) for $\alpha = 0$ and $\beta = 0$, respectively, give the same value for the coefficient $b_{\xi_{1,0,0}}$, which is therefore uniquely determined. The choice of $\mathcal{A}_{(2,1,1)} \subseteq \mathcal{M}$ and an argumentation along the lines of the proof of Theorem 4.3 now show that all the coefficients b_ξ, where $\xi \in \mathcal{D}_{q,Q_{(2,1,1)}}$ are uniquely determined, while the smoothness conditions (2.8) for $i = 1,\ j = 1,\ k = 1$, and (2.5)–(2.7) for $i = 2,\ j = 1,\ k = 1$, are satisfied. It now follows from induction, the choice of points in $\mathcal{A}_{(i,1,1)} \cup \mathcal{B}_{(1,i,1)} \cup \mathcal{C}_{(1,1,i)} \subseteq \mathcal{M}$, and the same arguments that b_ξ is uniquely determined if $\xi \in \mathcal{D}_{q,Q_{(i,1,1)}} \cup \mathcal{D}_{q,Q_{(1,i,1)}} \cup \mathcal{D}_{q,Q_{(1,1,i)}}$, where $i = 2,\dots,n$.

Next, we consider the cube $Q_{(2,2,1)}$. According to the above ordering, this cube has exactly two square faces in common with some of the cubes considered before, namely, the cubes $Q_{(1,2,1)}$ and $Q_{(2,1,1)}$. It follows from the above that the coefficients associated with domain points in the triangular faces within these squares are uniquely determined. Moreover, since the coefficients b_ξ, where $\xi \in \{\xi_{q-1,\alpha,\beta},\ \xi_{q,\alpha,\beta},\ \xi_{q,\alpha+1,\beta+1}\} \subseteq \mathcal{D}_{q,Q_{(1,2,1)}}$ and $\xi \in \{\xi_{\alpha,q-1,\beta},\ \xi_{\alpha,q,\beta},\ \xi_{\alpha+1,q,\beta+1}\} \subseteq \mathcal{D}_{q,Q_{(2,1,1)}}$, are uniquely determined, it follows from the smoothness conditions in (2.8) and (2.9) that the coefficients b_ξ, where $\xi \in \{\xi_{1,\alpha+1,\beta+1},\ \xi_{\alpha+1,1,\beta+1}\} \subseteq \mathcal{D}_{q,Q_{(2,2,1)}},\ \alpha,\beta = 0,\dots,q-1$, are uniquely determined. In particular, the smoothness conditions (2.5) and (2.6) for

$\beta = 1$ and $\alpha = 0, \ldots, q-1$ (w.r.t. $Q_{(2,2,1)}$) are automatically satisfied, which can be seen by performing some elementary computations. Moreover, we note that the coefficients of the form b_ξ, where $\xi = \xi_{1,1,\beta},\ \beta = 0, \ldots, q$, are uniquely determined, since the spline s and its (three) first partial derivatives are already uniquely determined along the common edge of $Q_{(1,2,1)}$ and $Q_{(2,1,1)}$. An argument as above shows that the coefficients b_ξ, where $\xi = \xi_{1,\beta,0}$ and $\xi = \xi_{\beta,1,0},\ \beta = 1, \ldots, q$, respectively, are uniquely determined. The choice of $\mathcal{A}_{(2,2,1)} \cap \mathcal{B}_{(2,2,1)} \subseteq \mathcal{M}$ and an argumentation along the lines of the proof of Theorem 4.3 now show that all the coefficients b_ξ, where $\xi \in \mathcal{D}_{q,Q_{(2,2,1)}}$, are uniquely determined, while the smoothness conditions (2.8) for $i = 1,\ j = 2,\ k = 1$, (2.9) for $i = 2,\ j = 1,\ k = 1$, and (2.5)–(2.7) for $i = 2,\ j = 2,\ k = 1$ are satisfied. It now follows from induction, the choice of points in $\big(\mathcal{A}_{(i,j,1)} \cap \mathcal{B}_{(i,j,1)}\big) \cup \big(\mathcal{A}_{(i,1,j)} \cap \mathcal{C}_{(i,1,j)}\big) \cup \big(\mathcal{B}_{(1,i,j)} \cap \mathcal{C}_{(1,i,j)}\big) \subseteq \mathcal{M}$, and the same arguments that b_ξ is uniquely determined if $\xi \in \mathcal{D}_{q,Q_{(i,j,1)}} \cup \mathcal{D}_{q,Q_{(i,1,j)}} \cup \mathcal{D}_{q,Q_{(1,i,j)}}$, where $i, j = 2, \ldots, n$.

Finally, we consider the cube $Q_{(2,2,2)}$. According to the above ordering, this cube has exactly three square faces in common with cubes already considered, namely, the common faces with the cubes $Q_{(1,2,2)},\ Q_{(2,1,2)}$, and $Q_{(2,2,1)}$. It follows from the above that the coefficients associated with domain points in the triangular faces within these squares are uniquely determined. Moreover, since the coefficients b_ξ, where $\xi \in \{\xi_{q-1,\alpha,\beta},\ \xi_{q,\alpha,\beta},\ \xi_{q,\alpha+1,\beta+1}\} \subseteq \mathcal{D}_{q,Q_{(1,2,2)}}$, $\xi \in \{\xi_{\alpha,q-1,\beta},\ \xi_{\alpha,q,\beta},\ \xi_{\alpha+1,q,\beta+1}\} \subseteq \mathcal{D}_{q,Q_{(2,1,2)}}$, and $\xi \in \{\xi_{\alpha,\beta,q-1},\ \xi_{\alpha,\beta,q},\ \xi_{\alpha+1,\beta+1,q}\} \subseteq \mathcal{D}_{q,Q_{(2,2,1)}}$, are uniquely determined, it follows from the smoothness conditions in (2.8)–(2.10) that the coefficients b_ξ, where $\xi \in \{\xi_{1,\alpha+1,\beta+1},\ \xi_{\alpha+1,1,\beta+1},\ \xi_{\alpha+1,1,\beta+1}\} \subseteq \mathcal{D}_{q,Q_{(2,2,2)}}$, $\alpha, \beta = 0, \ldots, q-1$, are uniquely determined. In particular, the smoothness conditions (2.5)–(2.7) for $\beta = 1$ and $\alpha = 0, \ldots, q-1$ (w.r.t. $Q_{(2,2,2)}$) are automatically satisfied, which can be seen by performing some elementary computations. Moreover, we note that the coefficients of the form b_ξ, where $\xi \in \{\xi_{1,1,\beta},\ \xi_{1,\beta,1},\ \xi_{\beta,1,1}\},\ \beta = 0, \ldots, q$, are uniquely determined, since the spline s and its (three) first partial derivatives are already uniquely determined along the common edge of each pair of the cubes $Q_{(1,2,2)},\ Q_{(2,1,2)}$ and $Q_{(2,2,1)}$. The choice of $\mathcal{A}_{(2,2,2)} \cap \mathcal{B}_{(2,2,2)} \cap \mathcal{C}_{(2,2,2)} \subseteq \mathcal{M}$ and an argumentation along the lines of the proof of Theorem 4.3 now show that all the coefficients b_ξ, where $\xi \in \mathcal{D}_{q,Q_{(2,2,2)}}$, are uniquely determined, while the smoothness conditions (2.8) for $i = 1,\ j = 2,\ k = 2$, (2.9) for $i = 2,\ j = 1,\ k = 2$, (2.10) for $i = 2,\ j = 2,\ k = 1$, and (2.5)–(2.7) for $i = 2,\ j = 2,\ k = 2$ are satisfied. It now follows from induction, the choice of points in $\mathcal{A}_{(i,j,k)} \cap \mathcal{B}_{(i,j,k)} \cap \mathcal{B}_{(i,j,k)} \subseteq \mathcal{M}$ and the same arguments that b_ξ is uniquely determined if $\xi \in \mathcal{D}_{q,Q_{(i,j,k)}}$, where $i, j, k = 2, \ldots, n$.

This shows that all the coefficients of s are uniquely determined, while all the C^1 smoothness conditions described in Lemma 2.2 are satisfied. To complete the proof, we have to count the number of points in $\mathcal{M}$ and have to show that the cardinality of this set coincides with the number in (3.1).

Lemma 4.1 and Theorem 4.3 show that the set $\mathcal{M}_{(1,1,1)} = \mathcal{M}_Q$ contains $q^3+2\,q+1$ domain points, and it is obvious that this is also the number of points in every set $\mathcal{M}_{(i,j,k)}$ defined in (5.1), where $i, j, k = 1, \ldots, n$. Since the cardinality of $\mathcal{M}_{(2,1,1)} \setminus \mathcal{A}_{(2,1,1)}$ is $2\,q^2+2$, it follows that $\mathcal{A}_{(2,1,1)}$ contains $q^3-2\,q^2+2\,q-1$ domain points. The same number of points are contained in each of the sets $\mathcal{A}_{(i,1,1)},\ \mathcal{B}_{(1,i,1)}$, and $\mathcal{C}_{(1,1,i)},\ i = 2, \ldots, n$. The cardinality of $\mathcal{M}_{(2,2,1)} \setminus (\mathcal{A}_{(2,2,1)} \cap \mathcal{B}_{(2,2,1)})$ is $4\,q^2-3\,q+3$, and therefore $q^3-4\,q^2+5\,q-2$ domain points are contained in $\mathcal{A}_{(2,1,1)} \cap \mathcal{B}_{(2,2,1)}$. The same number of points are contained in each of the sets $\mathcal{A}_{(i,j,1)} \cap \mathcal{B}_{(i,j,1)},\ \mathcal{A}_{(i,1,j)} \cap \mathcal{C}_{(i,1,j)}$, and $\mathcal{B}_{(1,i,j)} \cap \mathcal{C}_{(1,i,j)},\ i, j = 2, \ldots, n$. Since the cardinality of $\mathcal{M}_{(2,2,2)} \setminus (\mathcal{A}_{(2,2,2)} \cap$

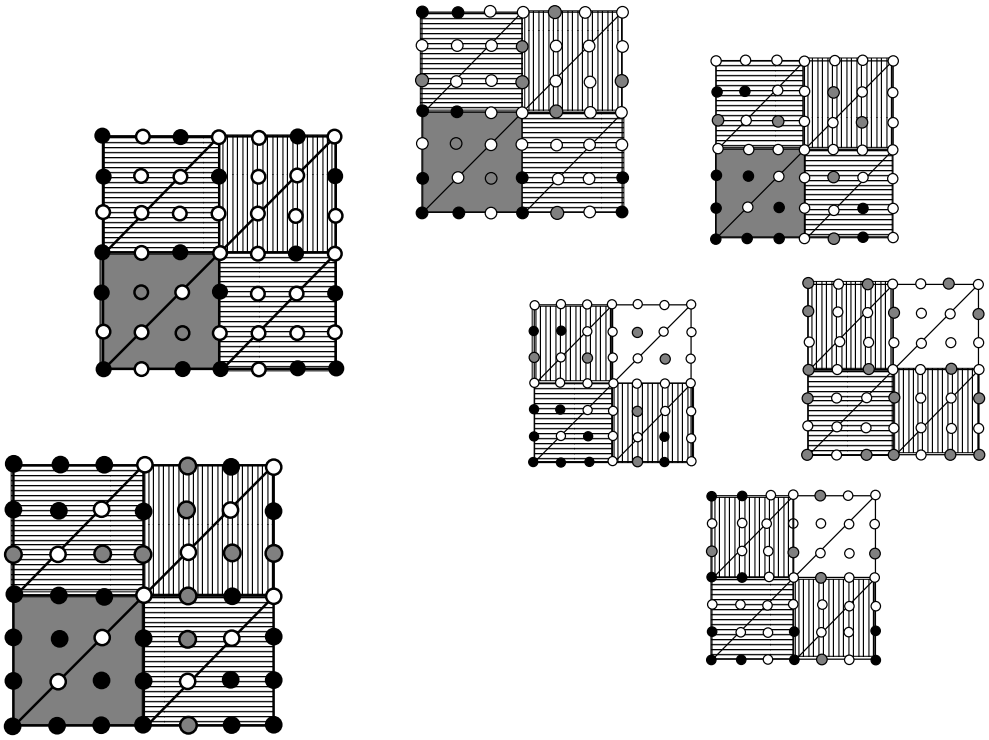


FIG. 7. *The MDS $\mathcal{M}$ for $n = 2$ in the case $q = 3$.*

$\mathcal{B}_{(2,2,2)} \cap \mathcal{C}_{(2,2,2)}$) is $6\ q^2 - 9\ q + 7$, it follows that $\mathcal{A}_{(2,2,2)} \cap \mathcal{B}_{(2,2,2)} \cap \mathcal{C}_{(2,2,2)}$ contains $q^3 - 6\ q^2 + 11\ q - 6$ domain points. The same number of points are contained in each of the sets $\mathcal{A}_{(i,j,k)} \cap \mathcal{B}_{(i,j,k)} \cap \mathcal{C}_{(i,j,k)},\ i, j, k = 2, \ldots, n$. We conclude that the total number of points in $\mathcal{M}$ is equal to

$$(q^3 + 2\ q + 1) + 3\ (n-1)\ (q^3 - 2\ q^2 + 2\ q - 1)$$
$$+3\ (n-1)^2\ (q^3 - 4\ q^2 + 5\ q - 2) + (n-1)^3\ (q^3 - 6\ q^2 + 11\ q - 6).$$

Some elementary computations now show that this is the number in (3.1). The proof is complete. □

Example 5.2. Figures 7 and 8 show examples for the MDS $\mathcal{M}$ for $n = 2$ in the cases $q = 3$ and $q = 4$, respectively. The figures show the different layers of domain points in a front to back manner, i.e., inside the planes $y = \ell/q,\ \ell = 0, \ldots, 2q$, organized in a spiral, where the different sizes of the layers indicate the distances to the front. The points from the set $\mathcal{M}$ are marked by black dots, the domain points associated with coefficients which are removed from $\mathcal{M}_{(i,j,k)}$ are marked by grey dots, while the remaining domain points determined by the various smoothness conditions are marked by white dots. The dimension of the spline spaces $\mathcal{S}_3^1(\Delta)$ and $\mathcal{S}_4^1(\Delta)$ in these examples are 88 and 250, respectively. In the proof of Theorem 5.1, the different cubes are considered in the order illustrated by the different symbols: grey, horizontally stripped, vertically stripped, and white.

6. Remarks. We close the paper with some remarks on possible extensions of our approach and discuss the differences to other spline spaces.

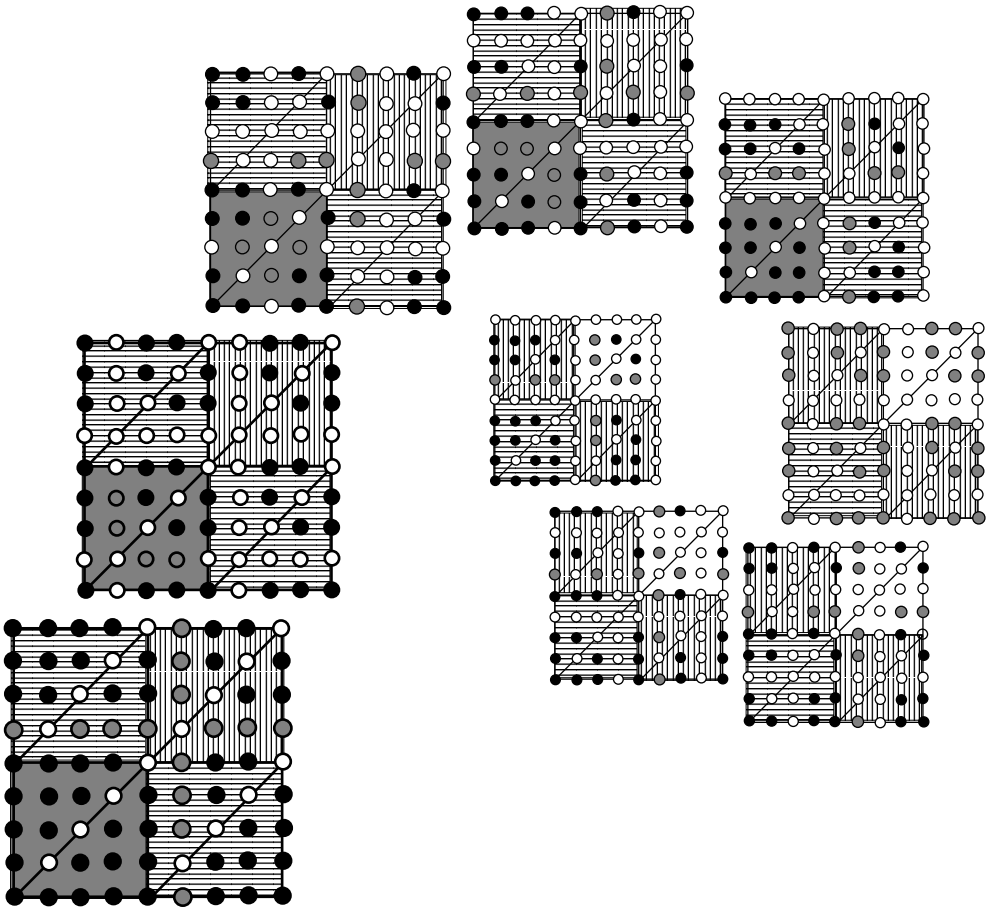


FIG. 8. *The MDS $\mathcal{M}$ for $n = 2$ in the case $q = 4$.*

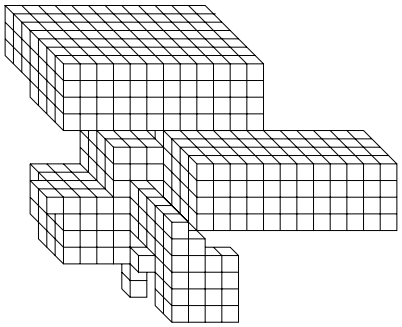


FIG. 9. *A more general domain Ω, where the inductive arguments of the proof of Theorem* 5.1 *can be applied.*

Remark 6.1. Our main results, Theorem 3.1 and Theorem 5.1, can be extended to more general partitions on different domains. An example for such a uniform cube partition is shown in Figure 9. Other examples of such domains are obtained from rectangular cube partitions, where there are n_j cubes with edge lengths 1 in the jth space direction, $j = 1, 2, 3$. In this case, the total number of cubes is equal to $n_1 n_2 n_3$

and the dimension of the corresponding C^1 spline space is given by

$$6\binom{q-1}{3} n_1n_2n_3 + 4\binom{q-1}{2}(n_1n_2 + n_1n_3 + n_2n_3) + 3\binom{q-1}{1}(n_1 + n_2 + n_3) + 4.$$

Moreover, we observe that the spacing in each direction does not need to be uniform, i.e., our arguments still hold when we start from a partition $\diamondsuit$, where $Q_{(i,j,k)} = [x_{i-1}, x_i] \times [y_{j-1}, y_j] \times [z_{k-1}, z_k]$, with $x_{i-1} < x_i,\ i = 1, \ldots, n_1,\ y_{j-1} < y_j,\ j = 1, \ldots, n_2, z_{k-1} < z_k,\ k = 1, \ldots, n_3$. However, in this case, the smoothness conditions of the spaces can no longer be described by one single formula as in Lemma 2.2. More precisely, one has to distinguish two cases: smoothness conditions across triangular faces for tetrahedra in different boxes involve different weights (barycentric coordinates evaluated at $\tilde{v}_3$), while smoothness conditions across triangular faces inside a box remain the same. In both cases, the smoothness conditions are still all of bivariate type, and our arguments hold for these more general partitions. However, in order to avoid complicated notations, we consider the uniform case, only. Moreover, we observe that C^r splines, $r \geq 2$, on Freudenthal partitions lead to more complex spline spaces, since the smoothness conditions are much more involved.

Remark 6.2. Since the trivariate splines we consider here are defined on a tetrahedral partition of a uniform grid, it is possible to compare them with different spline methods, which require the same grid, namely, approaches based on *tensor-product splines* and *box-splines*. First, we observe that the dimension of the trivariate splines is much larger than that of the tensor-product splines of the same total degree. For instance, tensor-product C^1 splines of lowest possible total degree are quadratic tensor-product splines (sometimes called triquadratic splines) and form a n^3+12n^2+6n+8-dimensional subspace of $\mathcal{S}_6^1(\Delta)$. Hence, using the full space $\mathcal{S}_6^1(\Delta)$ instead provides some additional flexibility, since its complete number of degrees of freedom is according to our result $60n^3 + 120n^2 + 45n + 4$. More generally, if tensor-product spline spaces exist, i.e., q is divisible by three, then we can understand these spaces as simple subspaces of $\mathcal{S}_q^1(\Delta)$ possessing a huge number of (unnecessary) supersmoothness conditions. This has also consequences for the approximation order, which for these subspaces is known to be not higher than $q/3 + 1 << q + 1$. A different approach is based on box-splines (see de Boor, Höllig and Riemenschneider [10]). These are piecewise polynomial functions of a certain degree and smoothness on uniform partitions with a local support which are usually defined as in Chui [12]. Most of the papers on these splines deal with the bivariate setting (see, for instance, de Boor and Höllig [9]), and it is known that box-splines may form a subspace or even a linearly dependent set of spline functions (see, for instance, Nürnberger [24]). For (slightly) more nonuniform partitions as described in the previous remark, the box-spline constructions would have to be—at least—adjusted. Moreover, we are not aware of any work on trivariate box-splines which yields explicit results on the dimension of the *full trivariate* spline space as we develop them here. For instance, Shi and Wang [37] focused on the existence of a trivariate box-spline and particularly showed this for $\mathcal{S}_q^1(\Delta)$ iff $q \geq 4$. On the other hand, for defining approximation operators as in Hecklin, Nürnberger, Schumaker and Zeilfelder [17] and Nürnberger, Rössl, Seidel and Zeilfelder [26, 32] (see also Schlosser, Hesser, Zeilfelder, Rössl, Männer, Nürnberger, and Seidel [33]) with advantageous properties for the applications, it is often necessary to know the precise structure of the trivariate spline spaces. We are not aware of a method or a practical test based on the usage of trivariate box-splines which has a comparable algorithmic behavior as it is reported in these trivariate spline methods. Also note that the implementation of the above-mentioned operators does not require an intermediate step of constructing explicitly a basis or a set of locally sup-

ported functions (see also the discussion in [39])—taking advantage of the piecewise Bernstein–Bézier form the methods can be directly applied to the given data and the methods from computer aided geometric design can be used for the effective further treatment of the reconstructed objects. Finally concerning the development of the above-mentioned operators, our approaches using low-degree trivariate splines provide the flexibility to subdivide (some of) the tetrahedra, and can therefore often be applied to deformed grids and more general tetrahedral partitions, too.

Remark 6.3. Bivariate and trivariate C^1 splines of lowest possible degree are extremely complex spaces. For instance, it is known that the dimension of quadratic C^1 bivariate splines on the three-directional mesh essentially coincides with the degrees of freedom of univariate splines, and we particularly showed here that this property is inherited by $\mathcal{S}_2^1(\Delta)$. On the other hand, it is still possible to work with quadratic C^1 splines. We note that the special structure of the Freudenthal partitions Δ *allows* to satisfy the necessary *geometric constraints* described in Worsey and Piper [41] (see also Sorokina and Worsey [38]) to apply the *Worsey–Piper split* to each tetrahedron of Δ.

Acknowledgments. The authors would like to thank the referees for their valuable comments.

REFERENCES

[1] G. Albertelli and R. A. Crawfis, *Efficient subdivision of finite-element datasets with consistent tetrahedra*, in Proceedings of IEEE Visualization, 1997, pp. 213–219.
[2] P. Alfeld, *A trivariate C^1 Clough-Tocher interpolation scheme*, Comput. Aided Geom. Design, 1 (1984), pp. 169–181.
[3] P. Alfeld, B. Piper, and L. L. Schumaker, *An explicit basis for C^1 quartic bivariate splines*, SIAM J. Numer. Anal., 24 (1987), pp. 891–911.
[4] P. Alfeld and L. L. Schumaker, *Smooth macro-elements based on Clough-Tocher triangle splits*, Numer. Math., 90 (2000), pp. 597–616.
[5] P. Alfeld and L. L. Schumaker, *Smooth macro-elements based on Powell-Sabin triangle splits*, Adv. Comput. Math., 16 (2002), pp. 29–46.
[6] P. Alfeld, L. L. Schumaker, and M. Sirvent, *The dimension and existence of local bases for multivariate spline spaces*, J. Approx. Theory, 70 (1992), pp. 243–264.
[7] P. Alfeld, L. L. Schumaker, and W. Whiteley, *The generic dimension of the space of C^1 splines of degree $d \geq 8$ on tetrahedral decompositions*, SIAM J. Numer. Anal., 30 (1993), pp. 889–920.
[8] C. de Boor, *B-form basics*, in Geometric Modeling: Algorithms and New Trends, G. Farin, ed., SIAM, Philadelphia, 1987, pp. 131–148.
[9] C. de Boor and K. Höllig, *Bivariate box-splines and smooth pp functions on a three-direction mesh*, J. Comput. Appl. Math., 9 (1983), pp. 13–28.
[10] C. de Boor, K. Höllig, and S. Riemenschneider, *Box Splines*, Springer, Berlin, 1993.
[11] H. Carr, T. Möller, and J. Snoeyink, *Simplicial subdivisions and sampling artifacts*, in Proceedings of IEEE Visualization, 2001, pp. 99–106.
[12] C. K. Chui, *Multivariate Splines*, CBMS-NSF Regional Conf. Ser. in Appl. Math., SIAM, Philadelphia, 1988, p. 189
[13] O. Davydov and F. Zeilfelder, *Scattered data fitting by direct extension of local polynomials to bivariate splines*, Adv. Comput. Math., (Special issue: Multivariate Splines), 21 (2004), pp. 223–271.
[14] G. Farin, *Triangular Bernstein–Bézier patches*, Comput. Aided Geom. Design, 3 (1986), pp. 83–127.
[15] H. Freudenthal, *Simplizialzerlegung von beschränkter Flachheit*, Ann. of Math., 43 (1942), pp. 580–582.
[16] T. Hangelbroek, G. Nürnberger, C. Rössl, H.-P. Seidel, and F. Zeilfelder, *Dimension of C^1-splines on type-6 tetrahedral partitions*, J. Approx. Theory, 131 (2004), pp. 157–184.
[17] G. Hecklin, G. Nürnberger, L. L. Schumaker, and F. Zeilfelder, *Efficient interpolation methods for cubic C^1 splines on partially subdivided Freudenthal tetrahedral partitions*, in preparation.

[18] M.-J. Lai and A. Le Méhauté, *A new kind of trivariate C^1 spline*, Adv. Comput. Math., (Special issue: Multivariate Splines), 21 (2004), pp. 223–271.
[19] M.-J. Lai and L. L. Schumaker, *Macro-elements and stable bases for splines on Clough-Tocher triangulations*, Numer. Math., 88 (2001), pp. 105–119.
[20] M.-J. Lai and L. L. Schumaker, *Quadrilateral macro-elements*, SIAM J. Math. Anal., 33 (2002), pp. 1107–1116.
[21] M.-J. Lai and L. L. Schumaker, *Splines on Triangulations*, to appear.
[22] G. Nielson and R. Franke, *Computing the separating surface from segmented data*, in Proceedings of IEEE Visualization, 1997, pp. 229–233.
[23] P. Ning and J. Bloomthal, *An evaluation of implicit surface tilers*, IEEE Comput. Graphics and Appl., 13 (1993), pp. 33–41.
[24] G. Nürnberger, *Approximation by Spline Functions*, Springer, Berlin, 1989.
[25] G. Nürnberger, V. Rayevskaya, L. L. Schumaker, and F. Zeilfelder, *Local Lagrange interpolation with bivariate splines of arbitrary smoothness*, Constr. Approx., 23 (2006), pg. 33–59.
[26] G. Nürnberger, C. Rössl, H.-P. Seidel, and F. Zeilfelder, *Quasi-Interpolation by quadratic piecewise polynomials in three variables*, Comput. Aided Geom. Design, 22 (2005), pp. 221–249.
[27] G. Nürnberger, L. L. Schumaker, and F. Zeilfelder, *Lagrange interpolation by C^1 cubic splines on triangulated quadrangulations*, Adv. Comput. Math. (Special issue: Multivariate Splines), 21 (2004), pp. 357–380.
[28] G. Nürnberger, L. L. Schumaker, and F. Zeilfelder, *Two Lagrange interpolation methods based on C^1 splines on tetrahedral partitions*, in Approximation Theory XI: Gatlinburg 2004, C. K. Chui, M. Neamtu and L. L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2005, pp. 101–118.
[29] G. Nürnberger and F. Zeilfelder, *Developments in bivariate spline interpolation*, J. Comput. Appl. Math., 121 (2000), pp. 125–152.
[30] G. Nürnberger and F. Zeilfelder, *Lagrange interpolation by bivariate C^1 splines with optimal approximation order*, Adv. Comput. Math. (Special issue: Multivariate Splines), 21 (2004), pp. 381–419.
[31] C. Rössl, F. Zeilfelder, G. Nürnberger, and H.-P. Seidel, *Reconstruction of volume data with quadratic super splines*, IEEE Trans. Visualization and Computer Graphics, 10 (2004), pp. 397–409.
[32] C. Rössl, F. Zeilfelder, G. Nürnberger, and H.-P. Seidel, *Spline approximation of general volumetric data*, in Proceedings of the ACM Symposium on Solid Modeling and Applications, 2004, pp. 1–8.
[33] G. Schlosser, J. Hesser, F. Zeilfelder, C. Rössl, R. Männer, G. Nürnberger, and H.-P. Seidel, *Fast visualization by shear-warp on quadratic super spline models using wavelet data decomposition*, C. T. Silva, E. Gröller, and H. Rushmeyer, eds., in Proceedings of IEEE Visualization, 2005, pp. 45–55.
[34] L. L. Schumaker, *Spline Functions: Basic Theory*, Wiley-Interscience, New York, 1980.
[35] L. L. Schumaker and T. Sorokina, *Quintic spline interpolation on type-4 tetrahedral partitions*, Adv. Comput. Math. (Special issue: Multivariate Splines), 21, 3–4 (2004), pp. 223–271.
[36] L.L. Schumaker and T. Sorokina, *A trivariate box macro element*, Constr. Approx., to appear.
[37] X. Shi and R. Wang, *Spline spaces and its B-splines on an $n+1$ direction mesh in $\mathbb{R}^n$*, J. Comput. Appl. Math., 144 (2002), pp. 241–250.
[38] T. Sorokina and A. Worsey, *A Multivariate Powell-Sabin Interpolant*, University of Georgia, Athens, GA, preprint.
[39] T. Sorokina and F. Zeilfelder, *Optimal quasi-interpolation by quadratic C^1 splines on four-directional meshes*, in Approximation Theory XI: Gatlinburg 2004, C. K. Chui, M. Neamtu, and L. L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2005, pp. 1–16.
[40] A. Worsey and G. Farin, *An n-dimensional Clough-Tocher interpolant*, Constr. Appr., 3 (1987), pp. 99–110.
[41] A. Worsey and B. Piper, *A trivariate Powell-Sabin interpolant*, Comput. Aided Geom. Design, 5 (1988), pp. 177–186.
[42] F. Zeilfelder, *Scattered data fitting with bivariate splines*, in Principles of Multiresolution in Geometric Modelling, M. Floater, A. Iske, and E. Quak, eds., Springer-Verlag, Heidelberg, 2002, pp. 243–286.
[43] A. Ženišek, *Polynomial approximation on tetrahedrons in the finite element method*, J. Approx. Theory, 7 (1973), pp. 334–351.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 368–384

VANISHING VISCOSITY LIMIT TO RAREFACTION WAVES FOR THE NAVIER–STOKES EQUATIONS OF ONE-DIMENSIONAL COMPRESSIBLE HEAT-CONDUCTING FLUIDS*

SONG JIANG†, GUOXI NI†, AND WENJUN SUN†

Dedicated to Professor Rolf Leis on the occasion of his 75th birthday

Abstract. We prove the solution of the Navier–Stokes equations for one-dimensional compressible heat-conducting fluids with centered rarefaction data of small strength exists globally in time, and moreover, as the viscosity and heat-conductivity coefficients tend to zero, the global solution converges to the centered rarefaction wave solution of the corresponding Euler equations uniformly away from the initial discontinuity.

Key words. compressible Navier–Stokes equations, vanishing viscosity limit, rarefaction waves, Euler equations

AMS subject classifications. 76N17, 35M10, 35Q30, 35L45, 35L65, 35B40

DOI. 10.1137/050626478

1. Introduction and the main result.

We study the asymptotic behavior, as the viscosity and heat-conductivity go to zero, of solutions to the Cauchy problem for the Navier–Stokes equations for a one-dimensional (1-d) compressible heat-conducting fluid (in Lagrangian coordinates):

$$\begin{cases} v_t - u_x = 0, \\ u_t + p_x = \epsilon\Big(\dfrac{u_x}{v}\Big)_x, \\ \Big(e + \dfrac{u^2}{2}\Big)_t + (up)_x = \Big(\kappa\dfrac{\theta_x}{v} + \epsilon\dfrac{uu_x}{v}\Big)_x \end{cases} \tag{1.1}$$

with (discontinuous) initial data

$$(u, v, e)(x, 0) = (u_0, v_0, e_0)(x), \qquad x \in \mathbb{R}, \tag{1.2}$$

where $v, u, \theta, p = p(e, v)$ and e denote the specific volume, the velocity, the temperature, the pressure, and the internal energy, respectively, and ϵ, κ are the viscosity and heat-conductivity coefficients, respectively. At infinity, the initial data u_0, v_0, e_0 are assumed to satisfy

$$\lim_{x\to\pm\infty}(u_0, v_0, e_0)(x) = (u_\pm, v_\pm, e_\pm), \tag{1.3}$$

where $u_\pm$,$v_\pm$ and $e_\pm$ are given constant states.

The system (1.1), describing the motion of the fluid, are the conservation laws of mass, momentum, and energy.

*Received by the editors March 10, 2005; accepted for publication (in revised form) December 14, 2005; published electronically May 12, 2006. This work was supported by the Special Funds for Major State Basic Research Projects, the NSFC (grant 10225105) and the CAEP (grant 2003-R-02) of China.

http://www.siam.org/journals/sima/38-2/62647.html

†LCP, Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, People's Republic of China (jiang@iapcm.ac.cn, gxni@iapcm.ac.cn, sun_wenjun@iapcm.ac.cn).

The asymptotic behavior of viscous flows, as the viscosity vanishes, is one of the important topics in the theory of compressible flows. It is expected that a general weak entropy solution to the Euler equations should be a (strong) limit of solutions to the corresponding Navier–Stokes equations with the same initial data as the viscosity and heat-conductivity tend to zero.

For the 1-d compressible isentropic Navier–Stokes equations

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v)_x = \epsilon\Big(\dfrac{u_x}{v}\Big)_x, \end{cases} \tag{1.4}$$

and the corresponding inviscid p-system

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v)_x = 0, \end{cases} \tag{1.5}$$

the vanishing viscosity limit for the Cauchy problem has been studied by several researchers. In [7] DiPerna uses the method of compensated compactness and establishes a.e. convergence of admissible solutions (u^ϵ, v^ϵ) of (1.4) to an admissible solution of (1.5), provided that (u^ϵ, v^ϵ) is uniformly L^∞ bounded and v^ϵ is uniform bounded away from zero. However, this uniform boundedness is difficult to verify in general, and the abstract analysis in [7] gets little information on the qualitative nature of the viscous solutions. In [14] Hoff and Liu investigate the inviscid limit problem for (1.4) in the case that the underlying inviscid flow is a single weak shockwave, and they show that solutions of the compressible Navier–Stokes equations with shock data exist and converge to the inviscid shocks, as viscosity vanishes, uniformly away from the shocks. Based on [9, 14], Xin in [30] shows that the solution to the Cauchy problem for the system (1.4) with weak centered rarefaction wave data exists for all time and converges to the weak centered rarefaction wave solution of the corresponding Euler equations, as the viscosity tends to zero, uniformly away from the initial discontinuity. Moreover, for a given centered rarefaction wave to the Euler equations with finite strength, he constructs a viscous solution to the compressible Navier–Stokes system with initial data depending on the viscosity, such that the viscous solution approaches the centered rarefaction wave as the viscosity goes to zero at the rate $|\ln\epsilon|\epsilon^{1/4}$ uniformly for all time away from $t = 0$. In the vanishing viscosity limit, the Prandtl boundary layers (characteristic boundaries) are studied for the multidimensional linearized compressible Navier–Stokes equations by using asymptotic analysis in [31, 32, 29], while the boundary layer stability in the case of noncharacteristic boundaries and one spatial dimension is discussed in [26, 23]. We mention that there is extensive literature on the vanishing artificial viscosity limit for hyperbolic systems of conservation laws; see, for example, [7, 8, 6, 9, 18, 17, 33, 10, 25, 3, 11, 12, 1], also cf. the monographs [2, 5, 24] and the references therein. We also mention that the convergence of the 1-d Broadwell model and the relaxation limit of a rate-type viscoelastic system to the isentropic Euler equations with centered rarefaction wave initial data are studied in [28, 15], respectively.

Our aim in this paper is to study the relation between the solution $(u^\epsilon, v^\epsilon, e^\epsilon)(x,t)$ of the Navier–Stokes equations for a compressible heat-conducting fluid (1.1) and the solution $(u, v, e)(x,t)$ of the corresponding inviscid Euler equations:

$$\begin{cases} v_t - u_x = 0, \\ u_t + p_x = 0, \\ \Big(e + \dfrac{u^2}{2}\Big)_t + (up)_x = 0 \end{cases} \tag{1.6}$$

with the initial data

$$(u,v,e)(x,0)=(\tilde{u}_0,\tilde{v}_0,\tilde{e}_0)(x),\qquad x\in\mathbb{R}, \tag{1.7}$$

satisfying

$$\lim_{x\to\pm\infty}(\tilde{u}_0,\tilde{v}_0,\tilde{e}_0)(x)=(u_\pm,v_\pm,e_\pm) \tag{1.8}$$

with the same constant states $(u_\pm,v_\pm,e_\pm)$ as in (1.3).

It is convenient to work with the equations for the entropy s and the absolute temperature θ. The second law of thermodynamics asserts that

$$\theta ds=de+pdv.$$

We assume, as is customary in thermodynamics, that given any two of the thermodynamics variables ρ,e,θ,s, and p, we can obtain the remaining three variables. If we choose (v,θ) as independent variables and write $(p,e,s)=(p,e,s)(v,\theta)$, we deduce that

$$s_v(v,\theta)=p_\theta(v,\theta),\quad s_\theta(v,\theta)=\frac{e_\theta(v,\theta)}{\theta},\quad e_v(v,\theta)=\theta p_\theta(v,\theta)-p(v,\theta).$$

Then, a straightforward calculation gives

$$s_t=\kappa\Big(\frac{\theta_x}{v\theta}\Big)_x+\kappa\frac{\theta_x^2}{v\theta^2}+\epsilon\frac{u_x^2}{v\theta}, \tag{1.9}$$

$$\theta_t+\frac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)}u_x=\frac{\kappa}{e_\theta(v,\theta)}\Big(\frac{\theta_x}{v}\Big)_x+\frac{\epsilon}{e_\theta(v,\theta)}\frac{u_x^2}{v}. \tag{1.10}$$

We may also choose (v,s) as independent variables and write

$$p=p(v,s),\quad \theta=\theta(v,s).$$

Thus, instead of (1.1), we shall study the system $(1.1)_1$, $(1.1)_2$, and (1.9), or $(1.1)_1$, $(1.1)_2$, and (1.10). Namely, we shall consider

$$\begin{cases} v_t-u_x=0,\\ u_t+p(v,s)_x=\epsilon\Big(\dfrac{u_x}{v}\Big)_x,\\ s_t=\kappa\Big(\dfrac{\theta_x}{v\theta}\Big)_x+\kappa\dfrac{\theta_x^2}{v\theta^2}+\epsilon\dfrac{u_x^2}{v\theta}, \end{cases} \tag{1.11}$$

with initial data

$$(u,v,s)(x,0)=(u_0,v_0,s_0)(x)=\begin{cases}(u_-,v_-,s_-), & x<0,\\ (u_+,v_+,s_+), & x>0,\end{cases} \tag{1.12}$$

where $u_\pm,v_\pm$ and $s_\pm$ are the constant states. The corresponding inviscid Euler equations read as

$$\begin{cases} v_t-u_x=0,\\ u_t+p(v,s)_x=0,\\ s_t=0. \end{cases} \tag{1.13}$$

We assume in this paper that the pressure p is a smooth function of its arguments satisfying

$$p_v(v,s) < 0 < p_{vv}(v,s) \quad \text{for} \quad v > 0. \tag{1.14}$$

Notice that the condition (1.14) assures that the system (1.13) has characteristic speeds

$$\lambda_1 = -\sqrt{-p_v}, \quad \lambda_2 = 0, \quad \lambda_3 = \sqrt{-p_v},$$

and there are two families of rarefaction waves for the Euler equations (1.13). For illustration, we describe only the 1-rarefaction waves, and thus assume $s^+ = s_- \equiv \overline{s}$. The case for the 3-rarefaction waves can be dealt with similarly.

Suppose the end states $(u_\pm, v_\pm, \overline{s})$ can be connected by 1-rarefaction waves. The centered 1-rarefaction wave connecting $(u_-, v_-, \overline{s})$ to $(u_+, v_+, \overline{s})$ is the self-similar solution $(u,v,s)(x,t) = (u^r, v^r, s^r)(x/t)$ of (1.13) defined by (see, e.g., [27, 4])

$$\begin{cases} s^r(\xi) = \overline{s}, \\ u^r(\xi) = u_- + \displaystyle\int_{v_-}^{v^r(\xi)} \lambda_1(z,\overline{s})dz, \\ \lambda_1(v^r,\overline{s})(x,t) \text{ increasing in } x, \ \ \lambda_1(v^r,\overline{s})(x,t) = -\sqrt{-p_v(v^r(x/t),\overline{s})}, \end{cases} \tag{1.15}$$

which is uniquely determined by the system (1.13) and the rarefaction wave initial data

$$(u,v,s)|_{t=0} \equiv (u_0^r, v_0^r, s_0^r)(x) = \begin{cases} (u_-, v_-, \overline{s}), & x < 0, \\ (u_+, v_+, \overline{s}), & x > 0. \end{cases} \tag{1.16}$$

For the internal energy $e(v,\theta)$ and the viscosity and heat-conductivity coefficients ϵ, κ, we assume that for some constant $C > 0$,

$$\begin{cases} e_\theta(v,\theta) > 0 \quad \text{for } v,\ \theta > 0, \\ \kappa = O(\epsilon) \text{ as } \epsilon \to 0, \quad \kappa(\epsilon)/\epsilon \geq C > 0 \quad \forall \epsilon > 0. \end{cases} \tag{1.17}$$

From the kinetic theory, the viscosity and heat-conductivity should be in the same order. In this sense, the assumption $\kappa = O(\epsilon)$ in (1.17) is reasonable.

For the sake of convenience, throughout this paper we denote

$$\alpha = |u_+ - u_-| + |v_+ - v_-|.$$

In this paper, we prove that the solution of system (1.11) with the centered rarefaction wave initial data (1.16) of small strength α exists for all time and converges to the centered rarefaction wave of the Euler equation (1.13) as $\epsilon \to 0$ uniformly away from the initial discontinuity. More precisely, the main result of this paper is stated in the following theorem.

THEOREM 1.1. *Let the constant states $(u_\pm, v_\pm, \overline{s})$ be connected by a centered 1-rarefaction wave $(u^r(\frac{x}{t}), v^r(\frac{x}{t}), s^r(\frac{x}{t}))$ defined by* (1.15). *Assume that* (1.14) *and* (1.17) *hold. Then, for α small enough, the compressible Navier–Stokes equations* (1.11) *with the rarefaction wave initial data* (1.16) *have a global piecewise smooth*

solution $(u^\epsilon(x,t), v^\epsilon(x,t), s^\epsilon(x,t))$*, such that the following hold:*

(i) $u^\epsilon, \theta^\epsilon$ *are continuous for* $t > 0$*,* v^ϵ *and* $u^\epsilon_x, v^\epsilon_x, \theta^\epsilon_x$ *are uniformly Hölder continuous in the set* $x < 0,\ t \geq \tau$ *and* $x > 0,\ t \geq \tau$ *for any* $\tau > 0$*;* $u^\epsilon_t, u^\epsilon_{xx}, v^\epsilon_{xt}, \theta^\epsilon_t, \theta^\epsilon_{xx}$ *are Hölder continuous on compact set* (x,t)*,* $x \neq 0, t > 0$*. Moreover, the jumps in* v^ϵ *at* $x = 0$ *satisfy*

$$|[v^\epsilon(0,t)]| \leq C_1 \exp\left(-C_2 t/\epsilon\right),$$

and so the other jumps, where C_1, C_2 *are positive constants independent of* t *and* ϵ*, and* $[\,\cdot\,]$ *denotes jumps in what follows.*

(ii) *The solution* $(u^\epsilon, v^\epsilon, s^\epsilon)$ *converges to the centered rarefaction wave* (u^r, v^r, s^r) *as* $\epsilon \to 0$ *uniformly away from* $t = 0$*; i.e., for any positive* h*, we have*

$$\lim_{\epsilon \to 0} \sup_{x \in \mathbb{R}, t > h} \left| (u^\epsilon(x,t), v^\epsilon(x,t), s^\epsilon(x,t)) - \left(u^r\left(\frac{x}{t}\right), v^r\left(\frac{x}{t}\right), s^r\left(\frac{x}{t}\right)\right) \right| = 0.$$

(iii) *For any fixed viscosity* $\epsilon > 0$*, the solution* $(u^\epsilon, v^\epsilon, s^\epsilon)$ *approaches the centered rarefaction wave* (u^r, v^r, s^r) *uniformly as time goes to infinity; i.e.,*

$$\lim_{t \to \infty} \sup_{x \in \mathbb{R}^1} \left| (u^\epsilon(x,t), v^\epsilon(x,t), s^\epsilon(x,t)) - \left(u^r\left(\frac{x}{t}\right), v^r\left(\frac{x}{t}\right), s^r\left(\frac{x}{t}\right)\right) \right| = 0.$$

Remark 1.1. (i) The exponential decay with respect to t of the jumps in v^ϵ also remains valid for $[u^\epsilon_x]$ and $[\theta^\epsilon_x]$.

(ii) The smallness of α is needed in (2.11) in section 2 to make $\|\varphi_{0y}\|_\pm$ small (cf. Remark 2.1).

To prove Theorem 1.1 and to overcome the difficulties induced by nonisentropy of the flow, we shall adapt and modify the arguments in [30, 13, 22]. Namely, we first use a natural scaling argument to reduce the proof to the nonlinear time-asymptotic stability analysis of rarefaction waves for the compressible Navier–Stokes equation (1.11) under nonsmooth initial perturbations. Then, observing that the approximation of the smooth rarefaction waves to the rarefaction wave of Euler equation depends on both the strength and the initial perturbation, we exploit the smoothing property induced by the parabolic parts in (1.12) and the smallness of α, and employ delicate energy estimates and carefully control jumps to obtain the theorem.

We point out here that in view of Theorem 1.1, an initial jump discontinuity at $x = 0$ can be allowed in (1.2). The evolution of this jump discontinuity is an important aspect in our analysis. It has been shown in [13] that the discontinuity evolution follows a curve $\dot{x} = -[u]/[v]$ in the x-t plane, and the jump discontinuity in v, u_x and θ_x decays exponentially in time, while the discontinuity in u and θ is smoothed out at positive time; see [13] for details. We shall exploit this fact in the proof of Theorem 1.1.

In section 2 we reformulate the problem and give the proof of Theorem 1.1, while section 3 is dedicated to the derivation of a priori estimates used in section 2.

Throughout this paper, we use the following notation:

$$\mathbb{R}^- := (-\infty, 0),\ \ \mathbb{R}^+ := (0, \infty), \quad \|\cdot\| \equiv \|\cdot\|_{L^2(\mathbb{R})},\ \ \|\cdot\|_{L^p} \equiv \|\cdot\|_{L^p(\mathbb{R})},$$
$$\|\cdot\|^2_\pm \equiv \|\cdot\|^2_{L^2(\mathbb{R}^-)} + \|\cdot\|^2_{L^2(\mathbb{R}^+)}, \quad \int_\pm \cdot dy \equiv \int_{-\infty}^0 \cdot dy + \int_0^\infty \cdot dy.$$

2. Reformulation and the proof of Theorem 1.1. In this section, we will reduce the proof of Theorem 1.1 to the nonlinear time-asymptotic stability analysis of rarefaction waves for the system (1.11) under nonsmooth perturbations.

First, we derive some necessary estimates on the rarefaction waves of the Euler equations (1.13) based on the inviscid Burgers equation, in particular, we construct an explicit smooth 1-rarefaction wave which well approximates a given centered 1-rarefaction wave. We start with the Riemann problem for the Burgers equation:

$$\begin{cases} w_t + \left(\frac{w^2}{2}\right)_x = 0, \\ w(x,0) = w_0^r(x), \end{cases} \tag{2.1}$$

where $w_0^r(x)$ is given by

$$w_0^r(x) = \begin{cases} w_-, & x < 0, \\ w_+, & x > 0. \end{cases}$$

If $w_- < w_+$, then the problem (2.1) has the centered rarefaction wave solution $w^r(x,t) = w^r(x/t)$ given by

$$w^r(x,t) = \begin{cases} w_-, & x/t \le w_-, \\ x/t, & w_- \le x/t \le w_+, \\ w_+, & x/t \ge w_+. \end{cases}$$

To construct a smooth rarefaction wave solution of the Burgers equation which approximates the centered rarefaction wave, we set for $\delta > 0$,

$$w_\delta(x) = w(\delta x) = \frac{w_+ + w_-}{2} + \frac{(w_+ - w_-)}{2}\tanh(\delta x)$$

and for each $\delta > 0$, we solve the following initial value problem:

$$\begin{cases} w_t + \left(\frac{w^2}{2}\right)_x = 0, \\ w(x,0) = w_\delta(x). \end{cases} \tag{2.2}$$

Next, we state certain properties that will be used later (see [30, 22] for a proof).

LEMMA 2.1. *For each $\delta > 0$, the problem* (2.2) *has a unique global smooth solution $w_\delta^r(x,t)$, such that the following hold:*

(i) $w_- < w_\delta^r(x,t) < w_+$, $\partial_x w_\delta^r(x,t) > 0$ *for* $x \in \mathbb{R}$, $t \ge 0$, $\delta > 0$.

(ii) *For any $1 \le p \le \infty$, there is a constant $C(p)$ depending only on p, such that*

$$\|\partial_x w_\delta^r(\cdot,t)\|_{L^p} \le C(p)\min\{(w_+ - w_-)\delta^{1-1/p}, (w_+ - w_-)^{1/p} t^{-1+1/p}\},$$
$$\|\partial_x^2 w_\delta^r(\cdot,t)\|_{L^p} \le C(p)\min\left\{(w_+ - w_-)\delta^{2-1/p}, \delta^{1-1/p}\frac{1}{t}\right\},$$
$$\|\partial_x^3 w_\delta^r(\cdot,t)\|_{L^p} \le C(p)\min\left\{(w_+ - w_-)\delta^{3-1/p}, \delta^{2-1/p}\frac{1}{t}\right\}.$$

(iii)

$$\lim_{t\to+\infty}\sup_{x\in\mathbb{R}}|w_\delta^r(x,t) - w^r(x,t)| = 0.$$

Now, set $w_\pm = \lambda_1(v_\pm, \overline{s})$, and we define $V(x,t), U(x,t), S(x,t), \Theta(x,t)$, the smooth approximation of $(v^r, u^r, s^r, \theta^r)$, by

$$\lambda_1(V(x,t),\overline{s}) = w_\delta^r(x,t), \quad U(x,t) = u_\pm + \int_{v_\pm}^{V(x,t)} \sqrt{-p_v(z,\overline{s})}dz,$$
$$S(x,t) = \overline{s}, \quad \Theta(x,t) = \theta(V(x,t),\overline{s}).$$

Then, it is not difficult to see that $V(x,t), U(x,t), S(x,t), \Theta(x,t)$ satisfy

$$\begin{cases} V_t - U_x = 0, \\ U_t + p(V,\Theta)_x = 0, \\ S_t(V,\Theta) = 0, \\ \Theta_t + \dfrac{\Theta p_\theta(V,\Theta)}{e_\theta} U_x = 0, \end{cases} \tag{2.3}$$

and due to Lemma 2.1, the following lemma holds for V, U, S, Θ.

LEMMA 2.2. *The functions $V(x,t), U(x,t), S(x,t)$, and $\Theta(x,t)$ constructed above satisfy the following:*

(i) *$V_t = U_x > 0$ for all $x \in \mathbb{R}$, $t \geq 0$.*

(ii) *For any $1 \leq p \leq \infty$, there is a positive constant $C(p)$ depending only on p, such that*

$$\|(V_x, U_x, \Theta_x)(\cdot,t)\|_{L^p} \leq C(p)\min\{\alpha\delta^{1-1/p}, \alpha^{1/p}t^{-1+1/p}\},$$
$$\|(V_{xx}, U_{xx}, \Theta_{xx})(\cdot,t)\|_{L^p} \leq C(p)\min\{\alpha\delta^{2-1/p},\ \delta^{1-1/p}t^{-1}\},$$
$$\|(V_{xxx}, U_{xxx}, \Theta_{xxx})(\cdot,t)\|_{L^p} \leq C(p)\min\{\alpha\delta^{3-1/p},\ \delta^{2-1/p}t^{-1}\}.$$

(iii)

$$\lim_{t\to\infty}\sup_{x\in\mathbb{R}} |(V,U,S,\Theta)(t,x) - (v^r, u^r, s^r, \theta^r)(t,x)| = 0.$$

(iv)

$$|(V_t, U_t, \Theta_t)(x,t)| \leq C|(V_x, U_x, \Theta_x)(x,t)|.$$

Consequently, from Lemmas 2.1 and 2.2, it follows that $(U,V,\Theta)(x,t)$ converges to $(u^r, v^r, \theta^r)(x,t)$ as $t \to \infty$.

Now, we reformulate the problem by a natural scaling. Due to the scale invariance of the Riemann problem (1.13), (1.16), we rescale the Cauchy problem $(1.1)_1, (1.1)_2$ and (1.10) by

$$y = x/\epsilon, \quad \tau = t/\epsilon, \qquad \epsilon > 0$$

to obtain

$$\begin{cases} v_\tau - u_y = 0, \\ u_\tau + p_y(v,\theta) = \Big(\dfrac{u_y}{v}\Big)_y, \\ \theta_\tau + \dfrac{\theta p_\theta(v,\theta)}{e_\theta(v,\theta)} u_y = \dfrac{\mu}{e_\theta(v,\theta)}\Big(\dfrac{\theta_y}{v}\Big)_y + \dfrac{1}{e_\theta(v,\theta)}\dfrac{u_y^2}{v} \end{cases} \tag{2.4}$$

with initial data

$$(u,v,\theta)(y,0)=(u_0,v_0,\theta_0)(y),\quad y\in\mathbb{R}, \tag{2.5}$$

where $\mu=\kappa/\epsilon$ and by virtue of the assumptions (1.17),

$$\underline{\mu}\le\mu\le\overline{\mu}\quad\text{uniformly in }\epsilon>0\ \text{ for some positive constants }\underline{\mu},\ \overline{\mu}.$$

And in the case of the rarefaction wave initial data (1.16), the initial data (2.5) are

$$(u_0,v_0,\theta_0)(y)=\begin{cases}(u_-,v_-,\theta_-), & x<0,\\ (u_+,v_+,\theta_+), & x>0.\end{cases} \tag{2.6}$$

If there exists a unique global solution $(u,v,\theta)(y,\tau)$ to the problem (2.4), (2.5) with the same regularity as stated in Theorem 1.1, then the solution $(u^\epsilon,v^\epsilon,\theta^\epsilon)(x,t)$ to the problem $(1.1)_1,(1.1)_2$, (1.10), and (1.12) is given by

$$(u^\epsilon,v^\epsilon,\theta^\epsilon)(x,t)=(u,v,\theta)(x/\epsilon,t/\epsilon). \tag{2.7}$$

Hence, it follows that Theorem 1.1 can be proved if one can show

$$\lim_{\tau\to\infty}\sup_{y\in\mathbb{R}}\left|(u,v,\theta)(y,\tau)-(u^r,v^r,\theta^r)\left(\frac{y}{\tau}\right)\right|=0, \tag{2.8}$$

where $(u^r,v^r,\theta^r)(y/\tau)=(u^r,v^r,\theta^r)(x/t)$ is the centered 1-rarefaction wave solution defined by (1.15). Thus, the proof of Theorem 1.1 is reduced to showing that the centered rarefaction wave is a time-asymptotic state for the solution of (2.4) with discontinuous initial data (2.6); this will be a consequence of the following (more general) stability theorem.

THEOREM 2.3. *Let $(u^r,v^r,\theta^r)(y/\tau)$ be the centered 1-rarefaction wave as in Theorem* 1.1. *Consider the Cauchy problem for* (2.4), (2.5), *where (u_0,v_0,θ_0) and its derivatives are sufficiently smooth away from $y=0$, but up to $y=0$ with a simple jump discontinuity at $y=0$. Assume that*

$$(u_0-u_\pm,v_0-v_\pm,\theta_0-\theta_\pm)\in L^2(\mathbb{R}^\pm),\quad v_{0y}\in L^2(\mathbb{R}^-)\cap L^2(\mathbb{R}^+).$$

Then, there is a positive constant η_0, such that if

$$\|(u_0-u_\pm,v_0-v_\pm,\theta_0-\theta_\pm)\|_{L^2(\mathbb{R}^\pm)}+\|v_{0y}\|_\pm+\alpha\le\eta_0,$$

then the Cauchy problem (2.4), (2.5) *has a unique global solution $(u,v,\theta)(y,\tau)$ in the same function class as in Theorem* 1.1. *Moreover,*

$$\lim_{\tau\to\infty}\sup_{y\in\mathbb{R}}\left|(u,v,\theta)(y,\tau)-(u^r,v^r,\theta^r)\left(\frac{y}{\tau}\right)\right|=0.$$

Theorem 2.3 looks like nonlinear stability of centered rarefaction waves for the compressible Navier–Stokes equations; see, e.g., [16, 20, 21, 19, 22]. The main difference is that for the nonlinear stability of centered rarefaction waves, initial perturbation is smooth, while here one has to deal with initial perturbation with discontinuities, the time evolution of which has to be controlled properly. But, some ideas from the study of nonlinear stability can be borrowed here.

The proof of Theorem 2.3 is broken up into several steps. We start with the observation that by making use of the smooth rarefaction wave $(U, V, \Theta)(y, \tau)$ constructed above (e.g., one may take $\delta = 1$), one can decompose the solution $(u, v, \theta)(y, \tau)$ of (2.4), (2.5) into

$$(\varphi, \psi, \phi)(y, \tau) = (v - V, u - U, \theta - \Theta)(y, \tau), \quad \xi(y, \tau) = s(y, \tau) - \overline{s}.$$

Substituting the above decomposition into (2.4), (2.5), we obtain the system for the functions φ, ψ, ϕ, ξ:

$$\begin{cases} \varphi_\tau - \psi_y = 0, \\ \psi_\tau + \Big(p(v, \theta) - p(V, \Theta)\Big)_y = \Big(\dfrac{u_y}{v}\Big)_y, \\ \phi_\tau + \dfrac{\theta p_\theta(v, \theta)}{e_\theta(v, \theta)}\psi_y + \left(\dfrac{\theta p_\theta(v, \theta)}{e_\theta(v, \theta)} - \dfrac{\Theta p_\theta(V, \Theta)}{e_\theta(V, \Theta)}\right) U_y = \dfrac{1}{e_\theta(v, \theta)}\left(\mu\left(\dfrac{\theta_y}{v}\right)_y + \dfrac{u_y^2}{v}\right), \\ \xi_\tau = \mu\left(\dfrac{\theta_y}{v\theta}\right)_y + \mu\dfrac{\theta_y^2}{v\theta^2} + \dfrac{u_y^2}{v\theta} \end{cases} \tag{2.9}$$

with initial data

$$(\varphi, \psi, \phi, \xi)(y, 0) = (\varphi_0, \psi_0, \phi_0, \xi_0) \equiv (v_0 - V_0, u_0 - U_0, \theta_0 - \Theta_0, s_0 - \overline{s}), \tag{2.10}$$

where $(\varphi_0, \psi_0, \phi_0, \xi_0)$ and its derivatives are sufficiently smooth away from $y = 0$ but up to $y = 0$, and $(\varphi_0, \psi_0, \phi_0, \xi_0) \in L^2(\mathbb{R})$, $\varphi_{0y} \in L^2(\mathbb{R}^-) \cap L^2(\mathbb{R}^+)$.

We shall show that the Cauchy problem (2.9), (2.10) possesses a unique global solution $(\varphi, \psi, \phi, \xi)(y, \tau)$ in the same function class as for $(u^\epsilon, v^\epsilon, \theta^\epsilon)$ in Theorem 1.1. Moreover, (φ, ψ, ϕ) goes to zero uniformly as $\tau \to \infty$. This convergence then yields Theorem 2.3 due to Lemmas 2.1 and 2.2.

PROPOSITION 2.4. *There exists a positive constant η_0, such that if*

$$N^2(0) := \|(\varphi_0, \psi_0, \phi_0)\|^2 + \|\varphi_{0y}\|_\pm^2 + \alpha \le \eta_0, \tag{2.11}$$

then the Cauchy problem (2.9), (2.10) *has a unique global solution $(\varphi, \psi, \phi)(y, \tau)$ in the same function class as for $(u^\epsilon, v^\epsilon, \theta^\epsilon)(x, t)$ in Theorem* 1.1*, satisfying the following:*

(i)

$$\sup_{\tau \ge 0}(\|(\varphi, \psi, \phi)(\tau)\|^2 + \|\varphi_y(\tau)\|_\pm^2) + \int_0^\infty \|(\varphi_y, \psi_y, \phi_y)(\tau)\|_\pm^2 d\tau \le C\{N^2(0) + \delta^{1/4}\}.$$

(ii) *For any $\tau_0 > 0$, there is a constant $C(\tau_0) > 0$, such that*

$$\sup_{\tau \ge \tau_0}(\|(\psi_y, \phi_y)(\tau)\|_\pm^2 + \|(\psi_{yy}, \phi_{yy})(\tau)\|_\pm^2) + \int_{\tau_0}^\infty \|(\psi_{yy}, \psi_{y\tau}, \phi_{yy}, \phi_{y\tau})(\tau)\|_\pm^2 d\tau \le C(\tau_0)\{N^2(0) + \delta^{1/4}\}.$$

(iii) *The jump discontinuity of $\varphi(y, \tau)$ at $y = 0$ is bounded from above by*

$$\big|[\varphi](\tau)\big| \le C_1 \exp\{-C_2\tau\} \qquad \forall\, \tau \in (0, \infty).$$

Here C, C_1, C_2 are positive constants independent of τ.

Remark 2.1. It is not difficult to see that for the rarefaction wave initial data (1.16), the smallness of $\|(\varphi_0, \psi_0, \phi_0)\|$ in the condition (2.11) is satisfied provided that δ is appropriately large but without smallness of α, while the smallness of $\|\varphi_{0y}\|_\pm$ holds provided that for fixed δ, α is small enough.

Proof. To show Proposition 2.4, we combine the local existence and regularity result in [13] with an a priori energy estimate based on the nature of the underlying rarefaction wave. First, we state the following local existence, the proof of which can be found in [13].

LEMMA 2.5 (see Hoff [13]). *Suppose that $N(0)$ is suitably small so that there exist two positive constants $\underline{v}$ and $\overline{v}$ with $\underline{v} \le v_0^1(y) \le \overline{v}$ for all $y \in \mathbb{R}$. Then, there is a constant $T > 0$, such that the Cauchy problem* (2.9), (2.10) *has a solution (φ, ψ, ϕ) on $\mathbb{R} \times [0, T]$ in the same function class as for $(u^\epsilon, v^\epsilon, \theta^\epsilon)$ in Theorem* 1.1. *Moreover, φ, ψ, ϕ satisfies the following:*

(i) *There exists a positive constant C, such that*

$$\sup_{\tau \ge 0}(\|(\varphi, \psi, \phi)(\tau)\|^2 + \|\varphi_y(\tau)\|_\pm^2) + \int_0^T \|(\varphi_y, \psi_y, \phi_y)(\tau)\|_\pm^2 d\tau \le C\{N^2(0) + \delta^{1/4}\}.$$

(ii) *There is a positive constant C, such that*

$$\sup_{0 \le \tau \le T}(\|(\psi_y, \phi_y)(\tau)\|_\pm^2 + \|(\psi_{yy}, \phi_{yy})(\tau)\|_\pm^2) + \int_{\tau_0}^T \|(\psi_{yy}, \psi_{y\tau}, \phi_{yy}, \phi_{y\tau})(\tau)\|_\pm^2 d\tau \le C\{N^2(0) + \delta^{1/4}\}.$$

(iii) *There are constants $C_1, C_2 > 0$ independent of T, such that*

$$|[p(v, e)]| = \left|\left[\frac{u_y}{v}\right]\right| \le C_1 \exp\{-C_2 \tau\}.$$

By virtue of Lemma 2.5 and the continuation in time of the local solution, we see that to complete the proof of Proposition 2.4, it suffices to prove the following a priori estimate, the proof of which will be postponed to the next section.

PROPOSITION 2.6 (a priori estimate). *Let the assumptions in Lemma* 2.5 *be satisfied. Assume that the Cauchy problem* (2.9), (2.10) *has a solution $(\varphi, \psi, \phi)(y, \tau)$ on $\mathbb{R} \times [0, \tau]$ for some $\tau > 0$ in the same function class as in Lemma* 2.5. *Denote*

$$N^2(\tau_0, \tau) := \sup_{\tau_0 \le s \le \tau} \{\|(\varphi, \psi, \phi)(\tau)\|^2 + \|\varphi_y(\tau)\|_\pm^2\}, \quad 0 \le \tau_0 \le \tau.$$

Then, there are positive constants η_1 and C independent of τ_1, such that for each fixed τ_0, if

$$N^2(\tau_0, \tau_1) \le \eta_1,$$

then the following estimates hold:

$$N^2(\tau_0, \tau_1) + \int_{\tau_0}^{\tau_1} \|(\varphi_y, \psi_y, \phi_y)(\tau)\|_\pm^2 d\tau \le C\{\|(\varphi, \psi, \phi)(\tau_0)\|^2 + \|\varphi_y(\tau_0)\|_\pm^2 + \delta^{1/4}\},$$

$$\sup_{0 \le \tau \le T} \|(\psi_y, \phi_y)(\tau)\|_\pm^2 + \int_{\tau_0}^{\tau_1} \|(\psi_{yy}, \phi_{yy})\|_\pm^2 d\tau \le C\{\|(\varphi, \psi, \phi)(\tau_0)\|_\pm^2 + \|(\varphi_y, \psi_y, \phi_y)(\tau_0)\|_\pm^2 + \delta^{1/4}\}.$$

Proof of Theorem 2.3. By the systems (2.3) and (2.9), Lemma 2.2, and the Cauchy–Schwarz and Sobolev inequalities, we easily find that

$$\int_0^\infty \|(\varphi_t,\psi_t,\phi_t)(\tau)\|_{L^\infty}^2 d\tau<\infty,$$

which together with Proposition 2.4 yields $\lim_{\tau\to\infty}\sup_y|(\varphi(y,\tau),\psi(y,\tau),\phi(y,\tau))|\to 0$. Hence, in view of Lemma 2.2, we have proved Theorem 2.3. □

3. Uniform a priori estimates. In this section we derive the key a priori estimates given in Proposition 2.6. First, we introduce the normalized entropy $\eta(v,u,s,V,U,S)$ around (V,U,S):

$$\begin{aligned}\eta(v,u,s,V,U,S):=&\Big(e(v,\theta)+\frac{u^2}{2}\Big)-\Big(e(V,\Theta)+\frac{U^2}{2}\Big)\\&-\{-p(V,\Theta)(v-V)+U(u-U)+\Theta(s-S)\},\end{aligned}$$

where we have used the fact that $e_v(v,s)=-p(v,\theta),\ e_s(v,s)=0$.

An easy computation implies that η satisfies the following equation:

$$\begin{aligned}&\eta_\tau(v,u,s,V,U,S)+\{(p(v,\theta)-p(V,\Theta)\psi)\}_y+\Big(\Theta\frac{\psi_y^2}{v\theta}+\mu\Theta\frac{\phi_y^2}{v\theta^2}\Big)\\&\quad+\{p(v,s)-p(V,\overline{s})-p_v(V,\overline{s})\psi-p_s(V,\overline{s})\xi\}U_y\\&\quad=\Big(\frac{\psi\psi_y}{v}+\mu\frac{\phi\phi_y}{v\theta}\Big)_y+\Big(-\frac{U_y\psi\phi_y}{v^2}+2\frac{U_y\phi\psi_y}{v\theta}-\mu\frac{\Theta\phi\varphi_y}{v^2\theta}+\mu\frac{\Theta_y\phi\phi_y}{v\theta^2}\Big)\\&\quad+\Big(\frac{U_xx\psi}{v}+\mu\frac{\Theta_{yy}\phi}{v\theta}\Big)+\Big(-\frac{V_yU_y\psi}{v^2}+\frac{U_y^2\phi}{v\theta}-\mu\frac{V_y\Theta_y\phi}{v^2\theta}\Big).\end{aligned}\tag{3.1}$$

Employing (3.1), one has the following lemma.

LEMMA 3.1. *Suppose that the assumptions of Proposition* 2.6 *hold. Then,*

$$\begin{aligned}&\|(\varphi,\psi,\phi)(\tau)\|^2+\int_{\tau_0}^\tau(\|\sqrt{V_\tau}(\varphi,\phi)(\hat\tau)\|^2+\|(\psi_y,\phi_y)(\hat\tau)\|_\pm^2)d\hat\tau\\&\quad\le C\Big\{\|(\varphi,\psi,\phi)(\tau_0)\|^2+\delta^{1/4}+N(\tau_0,\tau)^{2/3}\int_{\tau_0}^\tau\|\varphi_y(\hat\tau)\|_\pm^2d\hat\tau\Big\}.\end{aligned}\tag{3.2}$$

Proof. Integrating (3.1) with respect to τ and y, we get

$$\|(\varphi,\psi,\phi)(\tau)\|^2+\int_{\tau_0}^\tau(\|\sqrt{V_\tau}(\varphi,\phi)\|^2+\|(\psi_y,\phi_y)\|_\pm^2)(\hat\tau)d\hat\tau\le C\Big\{\|(\varphi,\psi,\phi)(\tau_0)\|^2+\sum_{j=1}^{j=4}R_j\Big\},\tag{3.3}$$

where

$$\begin{aligned}R_1&=\int_{\tau_0}^\tau\int_\pm\Big\{-(p(v,\theta)-p(V,\Theta)\psi)_y+\Big(\frac{\psi\psi_y}{v}+\mu\frac{\phi\phi_y}{v\theta}\Big)_y\Big\}(y,\hat\tau)dyd\hat\tau,\\R_2&=\int_{\tau_0}^\tau\int_\pm(|\psi U_{yy}|+|\phi\Theta_{yy}|)(y,\hat\tau)dyd\hat\tau,\\R_3&=\int_{\tau_0}^\tau\int_\pm(|U_y\psi\varphi_y|+|U_y\psi_y\phi|+|\Theta_y\varphi_y\phi|+|\phi\phi_y\Theta_y|)(y,\hat\tau)dyd\hat\tau,\\R_4&=\int_{\tau_0}^\tau\int_\pm(|V_yU_y\psi|+|U_y^2\phi|+|\phi\Theta_yV_y|)(y,\hat\tau)dyd\hat\tau.\end{aligned}$$

Here we have used the assumption (1.17), the smallness of $N(\tau_0,\tau)$ such that $\underline{v}\le v\le\overline{v}$ and $\underline{\theta}\le\theta\le\overline{\theta}$ for some positive constants $\underline{v},\overline{v},\underline{\theta},\overline{\theta}$, the convexity of $p(v,s)$ with respect to v and s, and the equivalence of $|(\varphi,\psi,\xi)|^2$ to $|(\varphi,\psi,\phi)|^2$.

Recalling the definition of $N(\tau,\tau_0)$ and applying Lemma 2.2, for given α, R_j can be estimated as follows:

$$
\begin{aligned}
R_1 &= \int_{\tau_0}^{\tau}\Big\{\Big[-p(v,\theta)+p(V,\Theta)+\frac{\psi_y}{v}\Big]\psi+\Big[\frac{\phi\phi_y}{v\theta}\Big]\Big\}(\hat\tau)d\hat\tau=0,\\
R_2 &\le C\int_{\tau_0}^{\tau}\|(\psi,\phi)(\hat\tau)\|^{1/2}\|(\psi_y,\phi_y)(\hat\tau)\|^{1/2}\|U_{yy}(\hat\tau)\|_{L^1}d\hat\tau\\
&\le C\int_{\tau_0}^{\tau}\Big\{N(\tau_0,\tau)\|(\phi_y,\psi_y)(\hat\tau)\|^2+\|U_{yy}(\hat\tau)\|_{L^1}^{4/3}\Big\}d\hat\tau\\
&\le C\Big\{N(\tau_0,\tau)\int_{\tau_0}^{\tau}\|(\phi_y,\psi_y)(\hat\tau)\|^2d\hat\tau+\delta^{1/4}\Big\},
\end{aligned}
$$

$$
\begin{aligned}
R_4 &\le C\int_{\tau_0}^{\tau}\|(\psi,\phi)(\hat\tau)\|^{1/2}\|(\psi_y,\phi_y)(\hat\tau)\|^{1/2}\|U_y(\hat\tau)\|^2d\hat\tau\\
&\le C\int_{\tau_0}^{\tau}\Big\{N(\tau_0,\tau)\|(\phi_y,\psi_y)(\hat\tau)\|^2+\|U_y(\hat\tau)\|^{8/3}\Big\}d\hat\tau\\
&\le C\Big\{N(\tau_0,\tau)\int_{\tau_0}^{\tau}\|(\phi_y,\psi_y)(\hat\tau)\|^2d\hat\tau+\delta^{1/4}\Big\},
\end{aligned}
$$

and

$$
\begin{aligned}
R_3 &\le C\int_{\tau_0}^{\tau}\|(\psi,\phi,\varphi)(\hat\tau)\|^{1/2}\|(\psi_y,\phi_y,\varphi_y)(\hat\tau)\|^{3/2}\|U_y(\hat\tau)\|d\hat\tau\\
&\le C\int_{\tau_0}^{\tau}\Big\{N(\tau_0,\tau)^{2/3}\|(\phi_y,\psi_y,\varphi_y)(\hat\tau)\|^2+\|U_y(\hat\tau)\|^4\Big\}d\hat\tau\\
&\le C\Big\{N(\tau_0,\tau)^{2/3}\int_{\tau_0}^{\tau}\|(\phi_y,\psi_y)(\hat\tau)\|^2d\hat\tau+\delta^{1/4}\Big\},
\end{aligned}
$$

where we have used the Sobolev inequality and the following inequality:

$$
\begin{aligned}
\int_{\tau_0}^{\tau}\|\partial_x^iU(\hat\tau)\|_{L^p}^{a+b}d\hat\tau &\le \sup_{r\in[\tau_0,\tau]}\|\partial_x^iU(\hat\tau)\|_{L^p}^{a}\int_{\tau_0}^{\tau}\|\partial_x^iU(\hat\tau)\|_{L^p}^{b}d\hat\tau\\
&\le C\delta^{(i-1/p)a}\int_{\tau_0}^{\tau}\|\partial_x^iU(\hat\tau)\|_{L^p}^{b}d\hat\tau.
\end{aligned}
$$

Substituting the above estimates for R_j $(j=1,\dots,4)$ into (3.3), we obtain (3.2). This completes the proof. □

We now proceed to derive bounds for the term $\int_{\tau_0}^{\tau}\|\varphi_x(\hat\tau)\|_{\pm}^2d\hat\tau$, and we have the following lemma.

LEMMA 3.2. *Suppose that the assumptions of Proposition* 2.6 *hold. Then,*

$$
\begin{aligned}
\|\varphi_y(\tau)\|_{\pm}^2+\int_{\tau_0}^{\tau}\|\varphi_y(\hat\tau)\|_{\pm}^2d\hat\tau &\le C\Big\{\|(\varphi_y,\psi)(\tau_0)\|^2+\|\psi(\tau)\|_{\pm}^2\\
&\quad+\int_{\tau_0}^{\tau}\Big(\|(\psi_y,\phi_y)(\hat\tau)\|_{\pm}^2+\|\sqrt{V_\tau(\hat\tau)}(\varphi,\phi)(\hat\tau)\|_{\pm}^2\Big)d\hat\tau+\delta\Big\}.
\end{aligned}
\tag{3.4}
$$

Proof. By (2.9), we easily find that

$$\begin{aligned}\Big(\frac{1}{2}\Big(\frac{\varphi_y}{v}\Big)^2-\frac{\varphi_y}{v}\psi\Big)_\tau-p_v(v,\theta)\frac{\varphi_y^2}{v}-\Big(\frac{\psi_y^2}{v}+\frac{p_\theta\varphi_y\phi_y}{v}\Big)+\Big(\frac{\psi\psi_y}{v}\Big)_y\\ =\Big\{V_y(p_v(v,\theta)-p_v(V,\Theta))\frac{\varphi_y}{v}+\Theta_y(p_\theta(v,\theta)-p_\theta(V,\Theta))\frac{\varphi_y}{v}\\ +\frac{U_y\psi\varphi_y}{v^2-V_y\psi\psi_y/v^2}\Big\}+\frac{V_y\psi_y\varphi_y}{v^3}-\frac{U_{yy}\varphi_y}{v^2}+\frac{V_yU_y\varphi_y}{v^3}.\end{aligned}\tag{3.5}$$

Integrating (3.5) with respect to y,τ over $(\tau_0,\tau)\times\mathbb{R}$, we infer

$$\begin{aligned}\|\varphi_y(\tau)\|_\pm^2+\int_{\tau_0}^\tau\|\varphi_y(\hat\tau)\|_\pm^2d\hat\tau\le C\Big\{\|(\varphi_y,\psi)(\tau_0)\|^2+\|\psi(\tau)\|^2\\ +\int_{\tau_0}^\tau\Big(\|(\psi_y,\phi_y)\|_\pm^2+\|\sqrt{V_t}(\varphi,\phi)\|_\pm^2\Big)(\hat\tau)d\hat\tau+\sum_{j=5}^{j=7}R_j\Big\},\end{aligned}\tag{3.6}$$

with

$$\begin{aligned}&R_5=-\int_{\tau_0}^\tau\int_\pm\Big(\frac{\psi\psi_y}{v}\Big)_yd\hat\tau,\quad R_6=\int_{\tau_0}^\tau\int_\pm(|\psi U_y\varphi_y|+|V_y\psi\psi_y|)(y,\hat\tau)dyd\hat\tau,\\ &R_7=\int_{\tau_0}^\tau\int_\pm(|\psi_yV_y\varphi_y|+|U_{yy}\varphi_y|+|U_yV_y\varphi_y|)(y,\hat\tau)dyd\hat\tau,\end{aligned}$$

where R_5,R_6,R_7 can be bounded as follows, using Sobolev's imbedding theorem and Lemma 2.5(iii):

$$\begin{aligned}R_5&=\int_{\tau_0}^\tau\Big[\frac{\psi\psi_y}{v}\Big]d\hat\tau=\int_{\tau_0}^\tau\psi\Big[\frac{\psi_y}{v}\Big]d\hat\tau\le\int_{\tau_0}^\tau\|\psi\|^{1/2}\|\psi_y\|_\pm^{1/2}\Big[\frac{u_y}{v}\Big]d\hat\tau\\ &\le\frac{1}{4}\int_{\tau_0}^\tau\|\psi_y\|_\pm^2d\hat\tau+C\sup_{\tau_0\le s\le\tau}\|\psi(s)\|^{2/3}\Big|[\varphi(\tau_0)]\Big|^{4/3}\int_{\tau_0}^\tau\exp\Big(-\frac{4}{3}C_3(\hat\tau-\tau_0)\Big)d\hat\tau\\ &\le\frac{1}{4}\int_{\tau_0}^\tau\|\psi_y\|_\pm^2d\hat\tau+\frac{1}{3}\sup_{\tau_0\le s\le\tau}\|\psi(s)\|^2+C\{\|(\varphi(\tau_0)\|^2+\|\varphi_y(\tau_0)\|^2)\},\end{aligned}$$

$$\begin{aligned}R_6&\le C\int_{\tau_0}^\tau\|\psi(\hat\tau)\|^{1/2}\|\psi_y(\hat\tau)\|^{1/2}\|U_y(\hat\tau)\|\,\|(\varphi_y,\psi_y)(\hat\tau)\|d\hat\tau\\ &\le C\Big\{N(\tau_0,\tau)^{2/3}\int_{\tau_0}^\tau\|(\varphi_y,\psi_y)(\hat\tau)\|^2d\hat\tau+\int_{\tau_0}^\tau\|U_y(\hat\tau)\|^4d\hat\tau\Big\}\\ &\le C\Big\{N(\tau_0,\tau)^{2/3}\int_{\tau_0}^\tau\|(\varphi_y,\psi_y)(\hat\tau)\|^2d\hat\tau+\delta^{1/4}\Big\},\end{aligned}$$

$$\begin{aligned}R_7&\le C\alpha\int_{\tau_0}^\tau\|\varphi_y\|_\pm\|\psi_y\|_\pm d\hat\tau+\int_{\tau_0}^\tau(\|U_{yy}\|\,\|\varphi_y\|_\pm+\|V_y\|_{L^4}\|U_y\|_{L^4}\|\varphi_y\|_\pm)d\hat\tau\\ &\le\frac{1}{2}\int_{\tau_0}^\tau\|\varphi_y(\hat\tau)\|_\pm^2d\hat\tau+C\int_{\tau_0}^\tau(\|\psi_y\|_\pm^2+\|U_y\|_{L^4}^4+\|V_y\|_{L^4}^4+\|U_{yy}\|^2)d\hat\tau\\ &\le\frac{1}{2}\int_{\tau_0}^\tau\|\varphi_y(\hat\tau)\|^2d\hat\tau+C\Big\{\int_{\tau_0}^\tau\|\psi_y(\hat\tau)\|^2d\hat\tau+\delta^{1/4}\Big\}.\end{aligned}$$

Inserting the estimates for R_j $(j = 5, 6, 7)$ into (3.6), we arrive at

$$\begin{aligned}\|\varphi_y(\tau)\|_\pm^2 + \int_{\tau_0}^{\tau} \|\varphi_y(\hat\tau)\|_\pm^2 d\hat\tau \le C\Big\{ &\|(\varphi_y,\psi)(\tau_0)\|_\pm^2 + \|\psi(\tau)\|^2 \\ &+ \int_{\tau_0}^{\tau} (\|(\psi_y,\phi_y)(\hat\tau)\|_\pm^2 + \|\sqrt{V_\tau}(\varphi,\phi)(\hat\tau)\|^2) d\hat\tau + \delta^{1/4}\Big\}. \end{aligned} \tag{3.7}$$

Finally, combining Lemma 3.1 with Lemma 3.2, we conclude that

$$\begin{aligned}\|(\varphi,\psi,\phi,\varphi_y)(\tau)\|^2 + \int_{\tau_0}^{\tau} \Big(\|\sqrt{V_\tau(\hat\tau)}(\varphi,\phi)(\hat\tau)\|^2 + \|(\varphi_y,\psi_y,\phi_y)(\hat\tau)\|_\pm^2\Big) d\hat\tau \\ \le C(\|(\varphi,\psi,\phi)(\tau_0)\|^2 + \|\varphi_y(\tau_0)\|_\pm^2 + \delta^{1/4}). \end{aligned} \tag{3.8}$$

Comparing with the standard energy estimate for the compressible Navier–Stokes equations, we refer (3.8) to the basic energy estimate. □

Next, we proceed to estimate higher order derivatives of ψ, ϕ in the space $L^\infty(\tau_0, \tau; L^2(\mathbb{R}^\pm))$.

LEMMA 3.3. *Suppose that the assumptions of Proposition* 2.6 *hold. Then,*

$$\|\psi_y(\tau)\|_\pm^2 + \int_{\tau_0}^{\tau} \|\psi_{yy}(\tau)\|_\pm^2 d\tau \le C(\|\psi_y(\tau_0)\|_\pm + \|(\varphi,\psi,\phi)(\tau_0)\|^2 + \delta^{1/4}). \tag{3.9}$$

Proof. Multiplying the second equation of (2.9) by $-\psi_{yy}$, one obtains

$$\begin{aligned}\Big(\frac{\psi_y^2}{2}\Big)_\tau + \frac{\psi_{yy}^2}{v} - (\psi_\tau\psi_y)_y &= (p_v(v,\theta)\varphi_y + p_\theta(v,\theta)\phi_y)\psi_{yy} + \frac{\varphi_y\psi_y\psi_{yy}}{v^2} \\ &+ V_y\{p_v(v,\theta) - p_v(V,\Theta)\}\psi_{yy} + \Theta_y\{p_\theta(v,\theta) - p_\theta(V,\Theta)\}\psi_{xx} \\ &+ \frac{V_y\psi_y\psi_{yy}}{v^2} + \frac{U_y\varphi_y\psi_{yy}}{v^2} - \frac{U_{yy}\psi_{yy}}{v} + \frac{V_yU_y\psi_{yy}}{v^2}, \end{aligned} \tag{3.10}$$

which, by integrating with respect to y and τ, leads to

$$\begin{aligned}\|\psi_y(\tau)\|_\pm^2 + \int_{\tau_0}^{\tau} \|\psi_{yy}(\hat\tau)\|_\pm^2 d\hat\tau &\le \|\psi_y(\tau_0)\|_\pm^2 + \int_{\tau_0}^{\tau}\int_\pm (\psi_\tau\psi_y)_y dy d\hat\tau \\ &+ C\Big\{ \int_{\tau_0}^{\tau} \Big(\|(\varphi_y,\psi_y,\phi_y)(\hat\tau)\|_\pm^2 + \|\sqrt{V_t}(\varphi,\phi)(\hat\tau)\|^2\Big) d\hat\tau \\ &+ \int_{\tau_0}^{\tau}\int_\pm \Big(|\varphi_y\psi_y\psi_{yy}| + |V_y\psi_y\psi_{yy}| + |U_y\varphi_y\psi_{yy}| \\ &+ |\psi_{yy}|(|U_{yy}| + |U_y^2|)\Big) dy d\hat\tau\Big\}. \end{aligned} \tag{3.11}$$

The terms on the right-hand side of (3.11) can be bounded as follows:

$$\begin{aligned}\int_{\tau_0}^{\tau}\int_\pm (\psi_\tau\psi_y)_y dy d\hat\tau &= [\psi\psi_y]|_{\tau_0}^{\tau} - \int_{\tau_0}^{\tau} \psi[\psi_y]_\tau d\hat\tau = \psi[\psi_y]|_{\tau_0}^{\tau} - \int_{\tau_0}^{\tau} \psi[\psi_y]_\tau d\hat\tau \\ &\le \frac{1}{8}\int_{\tau_0}^{\tau} \|\psi_{yy}(\hat\tau)\|_\pm^2 d\hat\tau + C\int_{\tau_0}^{\tau} (\|\varphi\|^2 + \|\psi\|^2 + \|\psi_y\|_\pm^2)(\hat\tau) d\hat\tau, \end{aligned}$$

where we have used the fact that the jump $[u_y]$ decays exponentially in τ (cf. the estimate of R_5 in the proof of Lemma 3.2); and

$$\begin{aligned}\int_{\tau_0}^{\tau}\int_\pm |(\varphi_y\psi_y\psi_{yy})(y,\hat\tau)| dx d\hat\tau &\le C\int_{\tau_0}^{\tau} \|\varphi_y(\hat\tau)\|_\pm \|\psi_y(\hat\tau)\|_\pm^{1/2} \|\psi_{yy}(\cdot,\hat\tau)\|_\pm^{3/2} d\hat\tau \\ &\le \frac{1}{8}\int_{\tau_0}^{\tau} \|\psi_{yy}(\hat\tau)\|_\pm^2 d\hat\tau + CN(\tau_0,\tau)\int_{\tau_0}^{\tau} \|\psi_y(\hat\tau)\|_\pm^2 d\hat\tau; \end{aligned}$$

$$\int_{\tau_0}^{\tau}\int_{\pm}|\psi_{yy}(y,\hat\tau)|(|V_y\psi_y|+|U_y\varphi_y|)(y,\hat\tau)dyd\hat\tau$$
$$\le\frac{1}{8}\int_{\tau_0}^{\tau}\|\psi_{yy}(\hat\tau)\|_{\pm}^2d\hat\tau+C\delta\int_{\tau_0}^{\tau}\|(\varphi_y,\psi_y)(\hat\tau)\|_{\pm}d\hat\tau;$$

and

$$\int_{\tau_0}^{\tau}\int_{\pm}|(\psi_{yy}(y,\tau))|(|U_{yy}|+|U_y|^2)(y,\tau)dyd\tau$$
$$\le\frac{1}{8}\int_{\tau_0}^{\tau}\|\psi_{yy}(\hat\tau)\|_{\pm}^2d\hat\tau+C\int_{\tau_0}^{\tau}(\|U_{yy}\|^2+\|U_y\|_{L^4}^4)(\hat\tau)d\hat\tau$$
$$\le\frac{1}{8}\int_{\tau_0}^{\tau}\|\psi_{yy}(\hat\tau)\|_{\pm}^2d\hat\tau+C\delta^{1/4}.$$

Substituting the above estimates into (3.10), we obtain (3.9). □

Similarly, we can bound the derivatives of ϕ as follows.

LEMMA 3.4. *Assume that the assumptions of Proposition* 2.6 *hold. Then,*

$$\|\phi_y(\tau)\|_{\pm}^2+\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat\tau)\|_{\pm}^2d\hat\tau\le C(\|(\varphi,\psi,\phi)(\tau_0)\|^2+||\phi_y(\tau_0)\|_{\pm}^2+\delta^{1/4}).\tag{3.12}$$

Proof. Multiplying the third equation of (2.9) by $-\phi_{yy}$, then integrating with respect to y and τ, and utilizing (3.7) and (3.8), we deduce that

$$\begin{aligned}&\|\phi_y(\tau)\|^2+\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat\tau)\|^2d\hat\tau\le\|\phi_y(\tau_0)\|^2+\int_{\tau_0}^{\tau}\|\psi_y(\hat\tau)\|^2d\hat\tau\\&+\int_{\tau_0}^{\tau}\int_{\pm}(\phi_\tau\phi_y)_yd\hat\tau\quad+C\Big\{\int_{\tau_0}^{\tau}\int_{\pm}\Big(|\phi_{yy}|(|\varphi_y\phi_y|+|\psi_y^2|)\\&+|U_y\phi_{yy}|(|\phi|+|\varphi|)+|\phi_{yy}|(|V_y\phi_y|+|\Theta_y\varphi_y|\\&\quad+|U_y\psi_x|)+|\phi_{yy}|(|U_{yy}|+|U_y^2|)\Big)(y,\hat\tau)dyd\hat\tau\Big\},\end{aligned}\tag{3.13}$$

where the right-hand side can be estimated as follows:

$$\int_{\tau_0}^{\tau}\int_{\pm}(\phi_\tau\phi_y)_ydyd\hat\tau=[\phi\phi_y]\Big|_{\tau_0}^{\tau}-\int_{\tau_0}^{\tau}\phi[\phi_y]_\tau=\phi[\phi_y]\Big|_{\tau_0}^{\tau}-\int_{\tau_0}^{\tau}\phi[\phi_y]_\tau$$
$$\le\frac{1}{16}\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat\tau)\|_{\pm}^2d\hat\tau+C\int_{\tau_0}^{\tau}(\|\psi\|^2+\|\phi\|^2+||\phi_y\|_{\pm}^2)(\hat\tau)d\hat\tau;$$

$$\int_{\tau_0}^{\tau}\int_{\pm}|(\varphi_y\phi_y\phi_{yy})(y,\hat\tau)|dyd\hat\tau\le C\int_{\tau_0}^{\tau}\|\varphi_y(\hat\tau)\|_{\pm}\|\phi_y(\hat\tau)\|_{\pm}^{1/2}\|\phi_{yy}(\hat\tau)\|_{\pm}^{3/2}d\hat\tau$$
$$\le\frac{1}{16}\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat\tau)\|_{\pm}^2d\hat\tau+CN(\tau_0,\tau)\int_{\tau_0}^{\tau}\|\phi_y(\hat\tau)\|_{\pm}^2d\hat\tau;$$

$$\int_{\tau_0}^{\tau}\int_{\pm}|(\psi_y^2\phi_{yy})(y,\hat\tau)|dyd\hat\tau\le C\int_{\tau_0}^{\tau}\|\psi_y(\hat\tau)\|_{\pm}^{3/2}\|\psi_{yy}(\hat\tau)\|_{\pm}^{1/2}\|\phi_{yy}(\hat\tau)\|_{\pm}d\hat\tau$$
$$\le\frac{1}{16}\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat\tau)\|_{\pm}^2d\hat\tau+CN(\tau_0,\tau)\int_{\tau_0}^{\tau}(\|\psi_{yy}(\hat\tau)\|_{\pm}^2+\|\psi_y(\hat\tau)\|_{\pm}^2)d\hat\tau;$$

$$\int_{\tau_0}^{\tau}\int_{\pm}|U_y\phi_{yy}|(|\varphi|+|\phi|)(y,\hat{\tau})dyd\hat{\tau}$$
$$\leq \frac{1}{16}\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat{\tau})\|_{\pm}^2 d\hat{\tau}+C\delta\int_{\tau_0}^{\tau}\|\sqrt{V_\tau}(\varphi,\phi)(\hat{\tau})\|d\hat{\tau};$$

$$\int_{\tau_0}^{\tau}\int_{\pm}|\phi_{yy}|(|V_y\phi_y|+|\Theta_y\varphi_y|+|U_y\psi_y|)(y,\hat{\tau})dyd\hat{\tau}$$
$$\leq \frac{1}{16}\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat{\tau})\|_{\pm}^2 d\hat{\tau}+C\delta\int_{\tau_0}^{\tau}\|(\varphi_y,\psi_y,\phi_y)(\hat{\tau})\|_{\pm}d\hat{\tau}$$

and

$$\int_{\tau_0}^{\tau}\int_{\pm}|\phi_{yy}|(|U_{yy}|+|U_y^2|)(y,\hat{\tau})dyd\hat{\tau}\leq\frac{1}{16}\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat{\tau})\|_{\pm}^2 d\hat{\tau}$$
$$+C\int_{\tau_0}^{\tau}(\|U_{yy}\|^2+\|U_y\|_{L^4}^4)(\hat{\tau})d\hat{\tau}\leq\frac{1}{16}\int_{\tau_0}^{\tau}\|\phi_{yy}(\hat{\tau})\|_{\pm}^2 d\hat{\tau}+C\delta^{1/4}.$$

Substitution of the above estimates into (3.13) gives Lemma 3.4 immediately. □

Now, combining Lemmas 3.1–3.4, we obtain Proposition 2.6.

Acknowledgment. The authors would like to thank the referees for their suggestions which improved the presentation of the present paper.

REFERENCES

[1] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math., 161 (2005), pp. 223–342.

[2] A. Bressan, *BV-Solutions to Hyperbolic Systems By Vanishing Viscosity*, S.I.S.S.A., Trieste, Italy, 2000.

[3] A. Bressan and T. Yang, *On the convergence rate of vanishing viscosity approximations*, Comm. Pure Appl. Math., 57 (2004), pp. 1075–1109.

[4] R. Courant and K. O. Friedrichs, *Supersonic Flows and Shock Waves*, Wiley-Interscience, New York, 1948.

[5] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, 2000.

[6] X. Q. Ding, G.-Q. Chen, and P. Luo, *Convergence of the Lax-Friedrich scheme for the system of equations of isentropic gas dynamics,* I, Acta Math. Scientia, 7 (1987), pp. 467–480 (in Chinese).

[7] R. J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27–70.

[8] R. J. DiPerna, *Convergence of the viscosity method for isentropic gas dynamics*, Comm. Math. Phys., 91 (1983), pp. 1–30.

[9] J. Goodman and Z. Xin, *Viscous limits for piecewise smooth solutions to systems of conservation laws*, Arch. Rational Mech. Anal., 121 (1992), pp. 235–265.

[10] E. Grenier and F. Rousset, *Stability of one-dimensional boundary layers by using Green's functions*, Comm. Pure Appl. Math., 54 (2001), pp. 1343–1385.

[11] O. Gùes, G. Métivier, M. Williams, and K. Zumbrun, *Multidimensional viscous shocks,* II*: The small viscosity limit*, Comm. Pure Appl. Math., 57 (2004), pp. 141–218.

[12] O. Gùes, G. Métivier, M. Williams, and K. Zumbrun, *A new approach to the stability of multidimensional viscous shocks*, preprint, 2004.

[13] D. Hoff, *Global well-posedness of Cauchy problem for the Navier-Stokes equations of non-isentropic flow with discontinuous initial data*, J. Differential Equations, 95 (1992), pp. 33–74.

[14] D. Hoff and T.-P. Liu, *The inviscid limit for Navier-Stokes equations of compressible, isentropic flow with shock data*, Indiana Univ. Math. J., 38 (1989), pp. 861–915.

[15] L. Hsiao and R. Pan, *Zero relaxation limit to centered rarefaction waves for a rate-type viscoelastic system*, J. Differential Equations, 157 (1999), pp. 20–40.
[16] S. Kawashima, A. Matsumura, and K. Nishihara, *Asymptotic behavior of solutions for the equations of a viscous heat-conductive gas*, Proc. Japan Acad. Ser. A Math. Sci., 62 (1986), pp. 249–252.
[17] P.-L. Lions, B. Perthame, and P. E. Souganidis, *Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates*, Comm. Pure Appl. Math., 49 (1996), 599–638.
[18] P.-L. Lions, B. Perthame, and E. Tadmor, *Kinetic formulation of the isentropic gas dynamics and p-system*, Comm. Math. Phys., 163 (1994), pp. 415–431.
[19] T.-P. Liu and Z. Xin, *Nonlinear stability of rarefaction waves for compressible Navier-Stokes equations*, Comm. Math. Phys., 118 (1988), pp. 451–465.
[20] A. Matsumura and K. Nishihara, *Asymptotics toward the rarefaction waves of the solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 3 (1986), pp. 1–13.
[21] A. Matsumura and K. Nishihara, *Global stability of the rarefaction waves of a one-dimensional model system for compressible viscous gas*, Comm. Math. Phys., 144 (1992), pp. 325–335.
[22] K. Nishihara, T. Yang, and H. J. Zhao, *Nonlinear stability of strong rarefaction waves for compressible Navier–Stokes equations*, SIAM J. Math. Anal., 35 (2004), pp. 1561–1597.
[23] F. Rousset, *Stability of small amplitude boundary layers for mixed hyperbolic-parabolic systems*, Trans. Amer. Math. Soc., 355 (2003), pp. 2991–3008.
[24] D. Serre, *Systems of Conservation Laws*, I, II, Cambridge University Press, Cambridge, UK, 2000.
[25] D. Serre, *Sur la stabilité des couches limites de viscosité*, Ann. Inst. Fourier, 51 (2001), pp. 109–129.
[26] D. Serre and K. Zumbrun, *Boundary layer stability in real vanishing viscosity limit*, Comm. Math. Phys., 221 (2001), pp. 267–292.
[27] J. Smoller, *Shock Waves and Reactive Diffusion Equations*, Springer-Verlag, New York, Berlin, 1983.
[28] W.-C. Wang and Z. Xin, *Fluid-dynamic limit for the centered rarefaction wave of the Broadwell equation*, J. Differential Equations, 150 (1998), pp. 438–461.
[29] Y. Wang and Z. Xin, *Zero-viscosity limit of the linearized compressible Navier-Stokes equations with highly oscillatory forces in the half-plane*, IMS preprint 2004-08, Chinese University of Hong Kong, Hong Kong.
[30] Z. Xin, *Zero dissipation limit to rarefaction waves for the one-dimensional Navier-Stokes equations of compressible isentropic gases*, Comm. Pure Appl. Math., 46 (1993), pp. 621–665.
[31] Z. Xin, *Viscous boundary layers and their stability*, I, J. Partial Differential Equations, 11 (1998), pp. 97–124.
[32] Z. Xin and T. Yanagisawa, *Zero-viscosity limit of the linearized Navier-Stokes equations for a compressible viscous fluid in the half-plane*, Comm. Pure Appl. Math., 52 (1999), pp. 479–541.
[33] S.-H. Yu, *Zero-dissipation limit of solutions with shocks for systems of hyperbolic conservation laws*, Arch. Ration. Mech. Anal., 146 (1999), pp. 275–370.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 385–413

CONCENTRATION OF GINZBURG–LANDAU ENERGIES WITH SUPERCRITICAL GROWTH*

N. DESENZANI† AND I. FRAGALÀ‡

Abstract. We study the asymptotic behavior of energies of Ginzburg–Landau type, for maps from $\mathbb{R}^{n+k}$ into $\mathbb{R}^k$, and when the growth exponent p is strictly larger than k. We prove a compactness and Γ-convergence result, with respect to a suitable topology on the Jacobians, seen as n-dimensional currents. The limit energy is defined on the class of n-integral boundaries M, and its density involves a family of optimal profile constants depending locally on the multiplicity of M.

Key words. Γ-convergence, Ginzburg–Landau functionals, currents, Jacobians, degree

AMS subject classifications. 49J45, 49Q20, 49Q15

DOI. 10.1137/S0036141004443981

1. Introduction.

The investigation of phase transition and concentration phenomena from a variational point of view involves functionals of the kind

$$E_\varepsilon(u) := \int_\Omega |\nabla u|^p + \frac{1}{\varepsilon^p} W(u), \qquad u \in W^{1,p}(\Omega;\mathbb{R}^k), \tag{1.1}$$

where $n \geq 0$ and $k \geq 1$ are given integers, Ω is a bounded regular domain in $\mathbb{R}^{n+k}$, $p > 1$ is a real exponent, and W is a nonnegative continuous potential on $\mathbb{R}^k$, vanishing only on the unit sphere S^{k-1}. A rigorous mathematical analysis of the asymptotic behavior of E_ε as the positive parameter ε tends to zero began in the 1970s with the celebrated paper by Modica and Mortola [30]. They proved that, taking $p = 2$, $n \geq 0$, and $k = 1$ (so that W is null at -1 and $+1$), the minima of E_ε on the class of functions with prescribed mean converge to a constant times the area of a minimal surface of dimension n. The multiplicative constant is determined by an "optimal profile problem," which selects the 1-dimensional transition of lowest cost between the two wells of the potential. The basic physical motivation was the Cahn–Hilliard model for phase transition of immiscible fluids [8], and the technique adopted to attack the problem was the Γ-convergence introduced by De Giorgi and Franzoni in [10]. For an outline of the proof of the Modica–Mortola theorem close in spirit to the aims of the present work, and for the numerous later extensions by different authors, we refer the reader to the survey paper [1] and enclosed references. Therein the reader may also find a quick introduction to Γ-convergence, whose general theory and applications are developed in [6, 9].

The analogous problem in the vector-valued case involves in its simplest form the Ginzburg–Landau energies obtained by taking $n = 0$ and $k = p = 2$ in (1.1) (so that W is null on S^1). In this case, the physical background is related to phase transition models for superconductors [17] or superfluids [18]. A detailed study of the asymptotic behavior of minimizers of Ginzburg–Landau energies on the class of

*Received by the editors May 19, 2004; accepted for publication (in revised form) June 26, 2005; published electronically May 12, 2006.
http://www.siam.org/journals/sima/38-2/44398.html
†Dipartimento di Matematica "F. Enrigues," Università degli Studi di Milano, Via Saldini 50, 20133 Milano, Italy (desenzani@mat.unimi.it).
‡Dipartimento di Matematica "F. Brioschi," Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (fragala@mate.polimi.it).

functions with prescribed boundary datum $g \in C^\infty(\partial\Omega; S^1)$ has been carried out by Bethuel, Brezis, and Hélein; see the monograph [5]. Roughly, when the degree of g is nonzero, energy concentration occurs near a finite number of isolated points—the so-called Ginzburg–Landau vortices. Moreover, all singularities have degree 1, and they tend to repel each other. These results have been obtained essentially through PDE methods for the Euler–Lagrange equation of E_ε, combined with some complex analysis devices.

A more variational approach allows one to deal with energies of the type (1.1) in any space dimensions $n \geq 0$ and $k \geq 2$, and taking $p = k$. The asymptotic behavior as $\varepsilon \to 0$ of such functionals E_ε has been studied in a recent paper by Alberti, Baldo, and Orlandi [3], and independently (for $k = 2$) by Jerrard and Soner [26]. When the boundary datum has a topological singularity, the energy of minimizers turns out to be of order $|\log \varepsilon|$ and concentrates near a surface of dimension n in Ω. A good topology which allows for the detection of energy concentration is a suitable *flat norm* $\mathbb{F}_\Omega$ for the Jacobians Ju_ε of u_ε; indeed, using a suitable Hodge-type operator between k-covectors and n-vectors, the k-forms Ju_ε may be identified with generalized oriented n-surfaces, or more precisely, with n-dimensional *integral boundaries* $\star Ju_\varepsilon$ in the sense of currents. So, within the general setting of currents theory, the behavior of u_ε can be described in a rigorous way by means of a compactness and Γ-convergence result [3, Theorem 1.1]: with a suitable Dirichlet condition, the minima of E_ε converge to a dimensional constant times the area measure of a minimizing n-current. We also refer the reader to [4, 7, 15, 19, 20, 22, 24, 27, 28, 29, 31] and the references therein for related concentration, compactness, and lower bound results.

In this paper, we study the variational convergence of the functionals E_ε in (1.1) still for arbitrary $n \geq 0$ and $k \geq 2$, but now for p *strictly larger* than k. We stress that, while in the scalar case $k = 1$ the choice of p is essentially not relevant for the problem, in the vector case $k \geq 2$ the choice $p = k$ is in some sense "critical." Indeed, for every $p > 1$, the Γ-limit of E_ε in L^1 is equal to $\int_\Omega |\nabla u|^p$ for $u \in W^{1,p}(\Omega, S^{k-1})$, and $+\infty$ otherwise. Now, the trace operator maps $W^{1,p}(\Omega, S^{k-1})$ onto $W^{1-\frac{1}{p},p}(\partial\Omega, S^{k-1})$ if and only if $p < k$ (see, e.g., [21]). Thus, for $p < k$, imposing any Dirichlet boundary datum does not produce energy concentration (i.e., the energy of minimizers remains bounded). On the other hand, in the supercritical case $p > k$, for suitable boundary data the energy of minimizers turns out to be of order ε^{k-p}. In particular, the appropriate sequence of rescaled functionals to be considered in order to determine a meaningful Γ-limit is obtained by multiplying the functionals E_ε in (1.1) by ε^{p-k}. In this paper, the Γ-limit of $\varepsilon^{p-k}E_\varepsilon$ is determined in the absence of boundary conditions; however, by suitable modifications, our main result (cf. Theorem 1.1 below) can be extended in order to include Dirichlet boundary data; see [11].

Though it seems to have no particular physical interpretation, from a purely mathematical point of view the problem in the case $p > k$ reveals some interesting features, different from the case $p = k$. The main new difficulties concern the influence of the potential W on energy concentration, and the consequent effects on the form of the Γ-limit. This phenomenon can be observed by looking at the behavior of a so-called *recovery sequence*, namely, a sequence $\{u_\varepsilon\}$ which satisfies condition (ii) in Theorem 1.1 below when $p > k$ (respectively, the analogous condition (ii) in Theorem 1.1 of [3] when $p = k$). Indeed, when $p = k$ we see that a zone carrying a vanishing energy is located between the limit integral boundary M and the set where $E_\varepsilon(u_\varepsilon)$ are of order $|\log \varepsilon|$; therefore, the shape of the potential does not modify the Γ-limit. On the contrary, when $p > k$, there is a whole neighborhood of M where $E_\varepsilon(u_\varepsilon)$ are of

order ε^{k-p}. As a consequence, the amount of energy concentration is governed by an auxiliary variational problem, which involves the potential and depends locally on the integer multiplicity of M; so we have to deal with a whole family of "optimal profile constants" τ_d for d varying in $\mathbb{Z}$. Their role is in some respect analogous to the one played by the constant τ which appears in the Γ-limit of E_ε in the scalar case: in terms of the unscaled energy $\int_{\mathbb{R}^n} |\nabla u|^p + W(u)$, when $k = 1$, τ represents the minimal cost for a transition between the two wells of the potential, whereas when $k \geq 2$, τ_d represents the minimal cost for a singularity of degree d. However, though the situation reveals a close resemblance to the scalar case, it is considerably more complicated, not only because τ_d varies with d, but also because, for fixed d, it is quite delicate to determine the class of admissible functions in the variational problem which defines τ_d. Actually, in order to obtain the Γ-lim sup inequality, all the competitors should have the same "trace at infinity," whereas, in order to prove the Γ-lim inf inequality, they should satisfy the strictly weaker condition of having just the same "degree at infinity." Fortunately this dichotomy, which is one of the most intriguing aspects of the problem, is only apparent: in fact the two involved variational problems turn out to have the same infimum, so that no gap occurs between the Γ-limsup and the Γ-liminf.

Our main result reads as follows. For the definition of the constants τ_d, $d \in \mathbb{Z}$, we refer the reader to section 4; moreover, we denote by α_k the Lebesgue measure of the unit ball in $\mathbb{R}^k$.

THEOREM 1.1. *Let $n \geq 0$, $k \geq 2$, and $p > k$. For $\varepsilon > 0$, let $F_\varepsilon := \varepsilon^{p-k} E_\varepsilon$, E_ε being the functionals defined in* (1.1). *Assume that the positive continuous potential W vanishes only on S^{k-1} and satisfies the growth conditions*

$$\liminf_{|y|\to 1} \frac{W(y)}{(1-|y|)^{p/(p-k+1)}} > 0 \quad \text{and} \quad \liminf_{|y|\to+\infty} \frac{W(y)}{|y|^p} > 0. \tag{1.2}$$

Then the following statements hold:

(i) *Compactness and Γ-lim inf inequality. Whenever $\sup_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$, up to subsequences we have $\mathbb{F}_\Omega(\star J u_\varepsilon - \alpha_k M) \to 0$, where M is an n-dimensional integral boundary in Ω with multiplicity σ, and*

$$\liminf_\varepsilon F_\varepsilon(u_\varepsilon) \geq \int_M \tau_{\sigma(x)}\, d\mathcal{H}^n(x). \tag{1.3}$$

(ii) Γ-lim sup *inequality. For every n-dimensional integral boundary M in Ω with multiplicity σ, there exists a sequence $\{u_\varepsilon\} \subset W^{1,p}(\Omega;\mathbb{R}^k)$ such that $\mathbb{F}_\Omega(\star J u_\varepsilon - \alpha_k M) \to 0$ and*

$$\limsup_\varepsilon F_\varepsilon(u_\varepsilon) \leq \int_M \tau_{\sigma(x)}\, d\mathcal{H}^n(x). \tag{1.4}$$

We point out that a deeper understanding of the infimum problems which define the constants τ_d might allow us to clarify whether the dependence of τ_d on d is linear, and hence to establish whether the Γ-limit given by Theorem 1.1 is a multiple of the area. However, we feel that this is a challenging task. Indeed, while the admissible fields in the variational problems defining τ_d satisfy just a *global* degree condition, to understand the way the infima τ_d depend on d one should obtain some information on the *local* degree of minimizing sequences. (Actually, one should know how concentrated or spread out the subsets of $\mathbb{R}^k$ where the global degree is attained are.) An adequate technique to derive this local information from a global degree seems to be

somehow missing. Still for $n = 0$, one may also try to locate the singularities: it could be of some interest to determine whether their behavior differs in some aspect from that of Ginzburg–Landau vortices for $p = k$.

Let us briefly describe how the contents are organized. Section 2 is conceived so as to make the paper self-contained: we summarize the basic notation and preliminaries, omitting all proofs and details, for which we indicate suitable references. In section 3, we give a proof of the compactness statement in Theorem 1.1 (i), which turns out to be simpler than the corresponding result in [3]; essentially, we use the deformation theorem and the uniform boundedness in mass for a suitably defined notion of projected modified Jacobians. Section 4 is entirely devoted to the study of the optimal profile constants τ_d. In particular, we give "weak" and "strong" definitions of τ_d, a priori different from each other, and we show that they coincide. In section 5, we prove the Γ-liminf inequality, first for $n = 0$, by exploiting the weak definition of τ_d, and then for $n > 0$, by means of a slicing technique. Finally, in section 6, we prove the Γ-limsup inequality, using as a key tool an existence result for maps with prescribed singularities given in [2], combined with the strong definition of τ_d.

2. Preliminaries.

2.1. Notation. Throughout the paper, sets and functions are assumed to be Borel measurable, and, when no ambiguity may arise, we omit the indication of the measure in integrals. We adopt the letters K and C for constants that are, respectively, universal (i.e., depend only on the parameters p, k, and n and possibly on the potential W) and not universal. In general, we do not use indices, so that K and C may take different values even within the same line. Moreover, we adopt the following list of standard symbols.

B^d_r	the open ball in $\mathbb{R}^d$ with center the origin and radius r
S^{d-1}	the unit sphere in $\mathbb{R}^d$
$\mathcal{L}^d$	the Lebesgue measure on $\mathbb{R}^d$
$\mathcal{H}^d$	the d-dimensional Hausdorff measure
$\delta(P)$	the Dirac mass at the point P
$\mu \llcorner A$	the restriction of the measure μ to the set A
α_d	the Lebesgue measure of B^d_1
$A \subset\subset B$	the set A is compactly contained into the set B
χ_A	the characteristic function of the set A
$\mathrm{dist}(x, A)$	the distance of the point x from the set A
$\sharp A$	the cardinality of the finite set A
$\cdot$	the Euclidean scalar product
$\lvert\ \rvert$	the Euclidean norm
$f \ll g$	$\lim(\lvert f\rvert/\lvert g\rvert)$ vanishes
$\mathrm{osc}(f, E)$	the oscillation of $f \in C^0(A; B)$ in the set $E \subseteq A$
$\mathrm{Lip}_{(\mathrm{loc})}(A; B)$	the space of (locally) Lipschitz functions $f : A \to B$
$\mathrm{Lip} f$	the Lipschitz constant of $f \in \mathrm{Lip}(A; B)$
$\big(L^p_{(\mathrm{loc})}(A; B), \|\ \|_{p,A}\big)$	the space of (locally) Lebesgue functions $f : A \to B$
$\big(W^{1,p}_{(\mathrm{loc})}(A; B), \|\ \|_{1,p,A}\big)$	the space of (locally) Sobolev functions $f : A \to B$

From now on, we always assume that $p > k \geq 2$ (while $n \geq 0$) and that Ω is an open bounded domain in $\mathbb{R}^{n+k}$ with a Lipschitz boundary. Given $A \subset \Omega$ and $u \in W^{1,p}(A; \mathbb{R}^k)$, we set

$$F_\varepsilon(u, A) := \int_A e_\varepsilon(u), \qquad e_\varepsilon(u) := \varepsilon^{p-k}|\nabla u|^p + \frac{1}{\varepsilon^k} W(u).$$

Finally, for convenience we fix a positive function ψ on $[0,+\infty)$ which vanishes only at 1 and satisfies $\psi(|y|) \le W(y)$ for every $y \in \mathbb{R}^k$.

2.2. Currents. For every integer $h \in \{0,\dots,n+k\}$, we denote by $\Lambda_h(\mathbb{R}^{n+k})$ and $\Lambda^h(\mathbb{R}^{n+k})$, respectively, the spaces of *h-vectors* and *h-covectors* in $\mathbb{R}^{n+k}$, endowed with the standard notion of duality product $\cdot$, wedge product $\wedge$, and Euclidean norm $|\ |$. An h-vector or covector is called *simple* if it can be represented as the wedge product of h-vectors. Recalling that an *h-form* on Ω is a map from Ω into $\Lambda^h(\mathbb{R}^{n+k})$, we set $\mathcal{D}^h(\Omega)$ as the class of all smooth h-forms with compact support in Ω. By the pairing between vectors and covectors, the dual of $\mathcal{D}^h(\Omega)$ may be identified with the space $\mathcal{D}_h(\Omega)$ of *h-currents* over Ω, namely, distributions with values into $\Lambda_h(\mathbb{R}^{n+k})$. In particular, when a current T is a (locally) finite measure with values into $\Lambda_h(\mathbb{R}^{n+k})$, we say that it has (*locally*) *finite mass*. In this case, we denote by $|T|$ the variation of the measure T, by $\|T\| := |T|(\Omega)$ the total variation of T, and by $\|T\|_A = |T|(A)$ the mass of T in the open set $A \subset \Omega$; furthermore, we call sptT the support of the measure T.

Let us draw our attention to some special classes of currents with locally finite mass. When $M \subseteq \Omega$ is an oriented h-dimensional manifold at least of class C^1, we denote by $[\![M]\!]$ the h-current which applied to an h-form ω gives its integral over M in the classical sense of differential geometry. More generally, when $M \subseteq \Omega$ is an h-rectifiable set (namely, it can be covered by countably many h-surfaces of class C^1 up to an $\mathcal{H}^h$-negligible set) endowed with an orientation ξ (namely, a mapping which associates with $\mathcal{H}^h$-a.e. $x \in M$ a simple unitary h-vector spanning the approximate tangent space to M at x) and a multiplicity σ (namely, a real function locally integrable with respect to $\mathcal{H}^h \llcorner M$), we set (M,ξ,σ) to be the current defined by

$$(M,\xi,\sigma)(\omega) := \int_M \sigma(\omega\cdot\xi)\, d\mathcal{H}^h \qquad \forall \omega \in \mathcal{D}^h(\Omega). \tag{2.1}$$

With some abuse of notation, when dealing with a current (M,ξ,σ), we often denote it just by M. If a current can be represented as in (2.1) for some integer-valued multiplicity σ, we call it an *h-rectifiable current*. A current (M,ξ,σ) is said to be a *real polyhedral current* if M is a sum of finitely many h-simplexes, with ξ and σ constant on each of them; in case σ is also integer-valued, M is called an *integral polyhedral current*.

For $h \ge 1$, the *boundary* of an h-current T is the $(h-1)$-current defined by the identity $\partial T(\omega) := T(d\omega)$, $d\omega$ being the differential of the $(h-1)$-form ω. In particular, by the Stokes theorem, $\partial[\![M]\!] = [\![\partial M]\!]$; notice also that, if T is a boundary, namely, it is the boundary of some current, then $\partial T = 0$. A current T is called a *normal current* if both T and ∂T have locally finite mass and an *integral current* if both T and ∂T are rectifiable.

We also need to introduce the push forward and the flat norm of currents. Given a linear map $L : \mathbb{R}^{n+k} \to \mathbb{R}^m$, the *pull back* of an h-covector $w \in \Lambda^h(\mathbb{R}^m)$ by L is the h-covector $L^\sharp w \in \Lambda^h(\mathbb{R}^{n+k})$ defined by $L^\sharp w \cdot (v_1 \wedge \cdots \wedge v_h) := w \cdot (Lv_1 \wedge \cdots \wedge Lv_h)$ for every $v_1,\dots,v_h \in \mathbb{R}^{n+k}$. Accordingly, if Ω' is an open set of $\mathbb{R}^m$, given a smooth map $f : \Omega \to \Omega'$, the pull back of an h-form $\omega \in \mathcal{D}^h(\Omega')$ is the h-form $f^\sharp\omega \in \mathcal{D}^h(\Omega)$ defined by $f^\sharp\omega(x) := (Df(x))^\sharp \omega(f(x))$. The *push forward* of an h-current $T \in \mathcal{D}_h(\Omega)$ is the h-current $f_\sharp T \in \mathcal{D}_h(\Omega')$ defined through the duality $f_\sharp T(\omega) = T(f^\sharp\omega)$ for every $\omega \in \mathcal{D}^h(\Omega')$. Whenever T has compact support, the push forward and the boundary operator commute; that is, $\partial(f_\sharp T) = f_\sharp(\partial T)$. If in addition T has locally finite mass,

the variation measures of T and $f_\sharp T$ are related by the inequality

$$|f_\sharp T|(A) \le (\mathrm{Lip} f)^h\, |T|(f^{-1}(A)) \qquad \forall A \text{ open } \subset \Omega. \tag{2.2}$$

Given $T \in \mathcal{D}_h(\Omega)$, we shall denote by $\mathbb{F}_\Omega(T)$ the following *flat norm* (see [3]):

$$\mathbb{F}_\Omega(T) := \begin{cases} \inf\{\|S\| : S \in \mathcal{D}_{h+1}(\Omega),\ T = \partial S\} & \text{if } T \text{ is a boundary,} \\ +\infty & \text{otherwise.} \end{cases}$$

We point out that $\mathbb{F}_\Omega$ is strictly related to the usual flat norm $\mathcal{F}_\Omega$ for integral currents, as defined in [13, section 4.1.24]. In particular, if $S_i \to S$ in the flat norm $\mathcal{F}_\Omega$, we have $\mathbb{F}_\Omega(\partial S_i - \partial S) \to 0$, which in turn implies $\partial S_i \to \partial S$ in the dual of forms of class C^1_0; see [3] for more details.

For a broader description of the elements of currents theory sketched above, we refer the reader to [13, 16, 30].

2.3. Jacobians. Given a smooth map $u : \Omega \to \mathbb{R}^k$, its Jacobian is the k-form defined on Ω by

$$Ju := u^\sharp(dy) = du^1 \wedge \cdots \wedge du^k, \tag{2.3}$$

where $dy = dy_1 \wedge \cdots \wedge dy_k$ is the standard volume form on $\mathbb{R}^k$, and the 1-form du^i is the differential of the ith component of u. As long as u is of class $W^{1,k}(\Omega;\mathbb{R}^k)$, (2.3) makes sense and defines a continuous operator $u \mapsto Ju$ from $W^{1,k}(\Omega;\mathbb{R}^k)$ into $L^1(\Omega;\Lambda^k(\mathbb{R}^{n+k}))$. Moreover, Ju may be represented as $\sum_\alpha \det \nabla_\alpha u(x) dx_\alpha$, where the sum is extended over all k-multi-indices $\alpha = (i_1, \dots, i_k)$ with $1 \le i_1 < \cdots < i_k \le n+k$, $\nabla_\alpha u$ is the $(k \times k)$-matrix with columns $\nabla_{i_1} u, \dots, \nabla_{i_k} u$, and dx_α is the k-covector $dx_{i_1} \wedge \cdots \wedge dx_{i_k}$.

For a more geometric interpretation of the Jacobian, it is convenient to consider, instead of the k-form Ju, the n-current $\star Ju$, obtained via the following identification $\star$ of k-vectors and n-covectors. Given a multi-index α of order k, $\star dx_\alpha$ is the n-vector $\mathrm{sgn}(\hat\alpha, \alpha) e_{\hat\alpha}$, where $\hat\alpha = (j_1, \dots, j_n)$ is the set of indices not contained in α, $e_{\hat\alpha}$ is the n-vector $e_{j_1} \wedge \cdots \wedge e_{j_n}$, and $\mathrm{sgn}(\hat\alpha, \alpha)$ is the sign of the permutation $(\hat\alpha, \alpha)$. Then $\star$ extends in a natural way to an operator bringing k-forms into n-currents. Such an operator transforms differentials into boundaries, as one can check that, for every $\omega \in \mathcal{D}^k(\Omega)$, there holds $\star(d\omega) = (-1)^n \partial(\star\omega)$. Now observe that

$$Ju = \frac{1}{k} d(u^\sharp \omega_0), \tag{2.4}$$

where, denoting by $\widehat{dy_i}$ the wedge product of all dy_j with $j \ne i$, ω_0 is the k-form

$$\omega_0(y) := \sum_i (-1)^{i-1} y_i \widehat{dy_i}. \tag{2.5}$$

In view of (2.4), Ju is always a differential, which implies that $\star Ju$ is always a boundary. Further, by (2.4) and the definition of flat norm $\mathbb{F}_\Omega$ given in section 2.2, we deduce that

$$\mathbb{F}_\Omega(\star Ju) \le \|u\|_{\infty,\Omega} \|\nabla u\|^{k-1}_{k-1,\Omega}. \tag{2.6}$$

Finally, let us stress that (2.4) makes sense (differently from (2.3)) even for functions $u : \Omega \to \mathbb{R}^k$ of class $L^\infty_{\mathrm{loc}} \cap W^{1,k-1}_{\mathrm{loc}}$, provided the right-hand side is interpreted as a distribution. Mappings u for which such a distribution turns out to be a measure have been called functions with *bounded k-variation* and were investigated in [25], to which we refer for a more detailed account of the Jacobians theory.

2.4. Degree. Let u be a continuous map between two oriented h-manifolds M and N in $\mathbb{R}^{n+k}$, with $h = 1, \dots, n+k$, M compact, and $\partial N = \emptyset$. Given $y \in N \setminus u(\partial M)$, we denote by $\deg(u, M, N, y)$ the Brouwer degree of u at y. For its general definition and main properties, we refer the reader to [14, 22]. Let us just recall that, if u is of class C^1 and y is a regular value for u, then $\deg(u, M, N, y)$ is the algebraic sum of the number of inverse images $x \in u^{-1}(y)$, each one counted with a plus or minus sign according to whether the linear operator $\nabla u(x)$ is orientation-preserving or orientation-reversing. When $\partial M = \emptyset$ and N is connected, $\deg(u, M, N, y)$ turns out to be independent of y; in this case, it will be denoted by $\deg(u, M, N)$, or by $\deg(u, M)$ if $h = k-1$ and $N = S^{k-1}$.

We shall heavily exploit the behavior of the degree under continuous homotopy: by Hopf's theorem [22, Chapter 5, Theorem 1.10], two functions u and v from a $(k-1)$-dimensional compact manifold M without boundary with values into S^{k-1} are in the same homotopy class if and only if $\deg(u, M) = \deg(v, M)$. Also, we shall use the following remark, whose proof is elementary and may be found, e.g., in [2, section 2.10]: if a function u from an open subset M of $\mathbb{R}^k$ into $\mathbb{R}^k$ maps ∂M into S^{k-1}, then $\deg(u, M, \mathbb{R}^k, y)$ equals $\deg(u, \partial M)$ for $y \in B_1^k$ and 0 for $y \in \mathbb{R}^k \setminus \overline{B_1^k}$.

2.5. Area and coarea formulae. For completeness, we recall the following simple versions of the area and coarea formulae which will be often used in the paper. Let $E \subset \mathbb{R}^k$, and $f : \mathbb{R}^k \to \mathbb{R}$. For every function $u \in W^{1,p}(E; \mathbb{R}^k)$ $(p > k)$ and every Lipschitz function $g : \mathbb{R}^k \to \mathbb{R}$, we have (see, respectively, [16, section 3.1.5] and [12, section 3.4])

$$\int_E f(u)|Ju(x)| = \int_{\mathbb{R}^k} f(y)\, \sharp\{u^{-1}(y) \cap E\}\, d\mathcal{L}^k, \tag{2.7}$$

$$\int_E f(x)|\nabla g(x)| = \int_{\mathbb{R}} dt \int_{g^{-1}(t)\cap E} f\, d\mathcal{H}^{k-1}. \tag{2.8}$$

3. Compactness. The starting point in proving the compactness statement in Theorem 1.1 (i) is the following key inequality:

$$F_\varepsilon(u, A) \geq K \int_A W(u)^{1-k/p}|Ju|. \tag{3.1}$$

It is obtained simply by applying first Young's inequality $a + b \geq K a^\lambda b^{1-\lambda}$ (with $a = \varepsilon^{p-k}|\nabla u|^p$, $b = \varepsilon^{-k} W(u)$, and $\lambda = k/p$) and then the algebraic inequality $|\nabla u|^k \geq |Ju|$. Thanks to (3.1), we can gain compactness without using the lower bound proved by Jerrard in [24], which instead is fundamental in the case $p = k$. Actually, given a sequence of functions $\{u_\varepsilon\}$ with equibounded energies, by applying (3.1) we can deduce a uniform bound in mass not for the very Jacobians of u_ε but for suitably defined "modified Jacobians." Let us introduce them, together with some additional tools and notation.

3.1. Modified Jacobians, grids, and projection maps. Let $\rho : [0, +\infty) \to [0, +\infty)$ be a smooth kernel such that $\rho(t) = 0$ for $t \geq 1/3$ and $\int_{B^k_{1/3}} \rho(|y|) = \alpha_k$. For every smooth function $u : \Omega \to \mathbb{R}^k$, we call *the modified Jacobian of* u the current $J'u := u^\sharp(\rho(|y|)\, dy) = \rho(|u|)Ju$. Thus, the basic distinguishing feature of $J'u$ with respect to Ju is that the support of $J'u$ does not intersect the set of points $x \in \Omega$ where $|u(x)| \geq 1/2$.

Given $l > 0$ and $x_0 \in \mathbb{R}^{n+k}$, we call *the grid of step* l *and center* x_0 the union $\mathcal{G} = \mathcal{G}^l(x_0)$ of all closed cubes of the form $x_0 + lz + [0, l]^{n+k}$ for z varying in $\mathbb{Z}^n$.

(We often omit the center or even the step, and we simply write $\mathcal{G}^l$ or $\mathcal{G}$.) For $h = 1, \dots, n+k$, we call h-cells of $\mathcal{G}$ the h-dimensional faces of the cubes of $\mathcal{G}$, and h-skeleton of $\mathcal{G}$ their union, which we denote by R_h. In particular, by saying that an open set A is delimited by $\mathcal{G}$, we mean that $\partial A \subset R_{n+k-1}$. We indicate by $\mathcal{G}'$ the dual grid of $\mathcal{G} = \mathcal{G}^l(x_0)$, namely, $\mathcal{G}' := \mathcal{G}^l\big(x_0 + (l/2, \dots, l/2)\big)$, and by R'_h the h-skeleton of $\mathcal{G}'$. For every h-cell Q of $\mathcal{G}$, we denote by Q' the unique $(n+k-h)$-cell of $\mathcal{G}'$ which intersects $\mathcal{G}$. When dealing with a sequence of grids $\mathcal{G}_\varepsilon$, $R_{\varepsilon,h}$ will denote the corresponding h-skeletons.

Let a grid of step l and a bounded open set V in $\mathbb{R}^{n+k}$ be given. We denote by V_d the closed d-neighborhood of V, with $d := 2\sqrt{n+k}\, l$ being twice the diameter of the cubes in $\mathcal{G}$. Then a suitable version [3, Lemma 3.8] of the Federer–Fleming deformation theorem [32, section 29] ensures the existence of a locally Lipschitz *projection map* $\Phi : \mathbb{R}^{n+k} \setminus R_{k-1} \to R'_n$ such that the following retraction property holds: for any n-current T in $\mathbb{R}^{n+k}$ with finite mass, such that

$$\mathrm{spt}T \subset \mathbb{R}^k \setminus R_{k-1} \quad \text{and} \quad (\mathrm{spt}\partial T) \cap V_d = \emptyset, \tag{3.2}$$

the push forward $\Phi_\sharp T$ is a well-defined current in $\mathbb{R}^{n+k}$ satisfying

$$\mathrm{spt}\Phi_\sharp T \subset R'_n \quad \text{and} \quad \mathrm{spt}\partial(\Phi_\sharp T) \cap V = \emptyset, \tag{3.3}$$

$$\|\Phi_\sharp T\| \le K\|T\|, \tag{3.4}$$

$$\mathbb{F}_V(T - \Phi_\sharp T) \le K l^{n+1} \int_{V_d} \frac{d|T|(x)}{[\mathrm{dist}(x, R_{k-1})]^n}. \tag{3.5}$$

If in addition $T = [\![M]\!]$ is the current associated with a smooth n-surface M, then for every k-cell Q of $\mathcal{G}$ such that M is transversal to Q and Q' intersects V, the multiplicity of $\Phi_\sharp T$ on Q' is a constant integer. It agrees, up to a sign depending on the orientations of Q and Q', with the intersection number of M and Q (see [22, section 5.2]).

Observe now that, if u is a function smooth in a neighborhood W of V_d, with $|u| \ge 1/2$ on $R_{k-1} \cap W$, then the current T obtained by extending $\star J' u$ to zero out of W fulfills both assumptions in (3.2). Therefore, its push forward $\Phi_\sharp T$ satisfies conditions (3.3), (3.4), and (3.5); in particular, it agrees in V with a real polyhedral boundary. Moreover, applying first the above-mentioned statement on the multiplicity of $\Phi_\sharp [\![M]\!]$ to the currents associated with the level sets of u and then the coarea formula, one may compute the multiplicity of $\Phi_\sharp(\star J' u)$ on an n-cell Q' of $\mathcal{G}'$ which intersects V. It equals up to a sign $\alpha_k \deg(u/|u|, \partial Q)$ (where α_k come out from the integral constraint satisfied by the kernel ρ; see [3, Lemma 3.9] for a complete proof).

PROPOSITION 3.1. *Let a sequence of smooth maps $\{u_\varepsilon\}$ with $\sup_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$, and open sets U, V, and W with $U \subset\subset V \subset\subset W \subset\subset \Omega$ be given. Then the following statements hold:*

(i) *Boundedness of modified Jacobians. We have $\sup_\varepsilon \| \star J' u_\varepsilon\|_\Omega < +\infty$.*

(ii) *Compactness of projected modified Jacobians. We can find a sequence of grids $\mathcal{G}_\varepsilon$ of infinitesimal step $l(\varepsilon) \gg \varepsilon$, such that $|u_\varepsilon|$ converge uniformly to 1 on $R_{\varepsilon,k-1} \cap W$; moreover, if Φ_ε are the retraction maps associated with $\mathcal{G}_\varepsilon$ as above, we have $\sup_\varepsilon \|(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon)\|_V < +\infty$, and we can extract a subsequence such that $\mathbb{F}_U[(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon) - \alpha_k M] \to 0$, where M is a n-dimensional integral boundary with finite mass in U.*

(iii) *Flat asymptotical equivalence. The sequences of currents $\{\star J u_\varepsilon\}$, $\{\star J' u_\varepsilon\}$, and $\{(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon)\}$ is asymptotically equivalent in the flat norm over U.*

From the previous proposition, it is easy to deduce the following.

Proof of the compactness statement in Theorem 1.1 (i). First, we take an open set $\Omega' \supset\supset \Omega$, and we extend u_ε to $\tilde{u}_\varepsilon$ on Ω' so that the energies remain equibounded, i.e., $\sup_\varepsilon F_\varepsilon(\tilde{u}_\varepsilon, \Omega') < +\infty$. This can be made by reflection around $\partial\Omega$; cf. Remark 8.2 of [3]. Next we observe that, by density, it is not restrictive to assume that u_ε are smooth. Then, it is enough to apply Proposition 3.1, with Ω replaced by Ω' and U replaced by Ω: by (ii), up to subsequences there exists an n-dimensional integral boundary M with finite mass in Ω, such that $\mathbb{F}_\Omega[(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon) - \alpha_k M] \to 0$; by (iii), we have also $\mathbb{F}_\Omega(\star J u_\varepsilon - \alpha_k M) \to 0$. □

The remainder of this section is devoted to the proof of Proposition 3.1. We need two preliminary lemmas, which will be repeatedly used in later sections as well. They are quite close to some technical results presented in [3, section 8].

LEMMA 3.2. *Given a nonnegative function $v \in L^1(\Omega)$, a positive finite measure μ in Ω, and a positive parameter l, we can position a grid $\mathcal{G}$ of step l so that both the following inequalities are satisfied;*

$$l^{n+1} \int_{R_{k-1}\cap\Omega} v \le K \int_\Omega v, \tag{3.6}$$

$$l^n \int_\Omega \frac{d\mu(x)}{[\mathrm{dist}(x, R_{k-1})]^n} \le K\mu(\Omega). \tag{3.7}$$

Proof. If the grid $\mathcal{G}^l$ is centered at x_0, the left-hand sides of (3.6) and (3.7) are functions of x_0, say, respectively, $f_1(x_0)$ and $f_2(x_0)$. Let us compute the integral means of $f_i(x_0)$ over $Q_l := [0, l]^{n+k}$ for $i = 1, 2$.

We denote by $\mathcal{I}$ the family of all subsets I of $\{1, \dots, n+k\}$ with $\sharp I = k-1$. For $I \in \mathcal{I}$, $I = \{i_1, \dots, i_{k-1}\}$, we let $R_{k-1}(I)$ be the union of the $(k-1)$-cells of $\mathcal{G}^l(x_0)$ which are parallel to the space spanned by the subset $\{e_{i_1}, \dots, e_{i_{k-1}}\}$ of the canonical basis $\{e_1, \dots, e_{n+k}\}$ of $\mathbb{R}^{n+k}$; also, we let $Q_l(I)$ be the set of points $x_0 \in Q_l$ whose coordinates $x_{0i_1}, \dots, x_{0i_{k-1}}$ equal zero. We have

$$\begin{aligned} \fint_{Q_l} f_1(x_0) &= \sum_{I\in\mathcal{I}} \fint_{Q_l} l^{n+1} \int_{R_{k-1}(I)\cap\Omega} v = \sum_{I\in\mathcal{I}} \fint_{Q_l(I)} l^{n+1} \int_{R_{k-1}(I)\cap\Omega} v \\ &= \sum_{I\in\mathcal{I}} \int_{Q_l(I)} \int_{R_{k-1}(I)\cap\Omega} v = \binom{n+k}{k-1} \int_\Omega v =: K_1 \int_\Omega v. \end{aligned}$$

Here the second equality holds since, for $I = \{i_1, \dots, i_{k-1}\}$, the set $R_{k-1}(I) \cap \Omega$ does not depend on the coordinates $x_{0i_1}, \dots, x_{0i_{k-1}}$ of x_0, while the fourth equality follows from Fubini's theorem and the computation of $\sharp\mathcal{I}$.

Denoting by R the $(k-1)$-skeleton of the grid with step 1 centered at the origin, we have

$$\begin{aligned} \fint_{Q_l} f_2(x_0) &= \fint_{Q_l} l^n \int_\Omega \frac{d\mu(x)}{[l\,\mathrm{dist}(l^{-1}(x - x_0), R)]^n}\, d\mathcal{L}^{n+k}(x_0) \\ &= \int_\Omega \fint_{Q_l} \frac{d\mathcal{L}^{n+k}(x_0)}{[\mathrm{dist}(l^{-1}(x - x_0), R)]^n} d\mu(x) \\ &= \left[\fint_{Q_1} \frac{d\mathcal{L}^{n+k}(y)}{[\mathrm{dist}(y, R)]^n} \right] \mu(\Omega) =: K_2 \mu(\Omega), \end{aligned}$$

where in the last line K_2 is a universal constant because the integral between square brackets is finite (see, e.g., [3, Lemma 8.3]).

Now, fix $K > K_1 + K_2$, and set $E_1 := \{x_0 \in Q_l : f_1(x_0) > K \int_\Omega v\}$, $E_2 := \{x_0 \in Q_l : f_2(x_0) > K\mu(\Omega)\}$. Using the above computations, we infer that $\mathcal{L}^{n+k}(E_i) \le (K_i/K)l^{n+k}$ for $i = 1, 2$. So $\mathcal{L}^{n+k}(E_1 \cup E_2) < l^{n+k}$; hence the complement of $E_1 \cup E_2$ in Q_l is nonempty. □

LEMMA 3.3. *Let $u \in W^{1,p}(\Omega; \mathbb{R}^k)$, and let $Q \subset \Omega$ be a $(k-1)$-cell of a grid of step l. If ψ is defined as in the end of section* 2.1*, for every $\varepsilon > 0$ there holds*

$$\fint_Q \psi(|u|) \le K\varepsilon^k l^{-k-n}\left\{l^{n+1}\int_Q e_\varepsilon(u)\right\}. \tag{3.8}$$

Moreover, if we fix $\lambda > (k-1)/p$, and Ψ is a primitive of $\psi^{(1-\lambda)/\lambda p}$, we have

$$\operatorname{osc}(\Psi(|u|), Q)^{\lambda p} \le K\varepsilon^{k-\lambda p} l^{\lambda p-k-n}\left\{l^{n+1}\int_Q e_\varepsilon(u)\right\}. \tag{3.9}$$

Proof. Inequality (3.8) is straightforward by taking into account that $\psi(|u|) \le \varepsilon^k e_\varepsilon(u)$. To prove (3.9), since $\lambda p > k-1$, we may apply the Morrey–Sobolev embedding theorem to obtain

$$\operatorname{osc}(\Psi(|u|), Q)^{\lambda p} \le K l^{\lambda p[1-(k-1)/\lambda p]} \int_Q |\nabla(\Psi(|u|))|^{\lambda p}. \tag{3.10}$$

By Young's inequality and the definition of Ψ, we also have

$$\varepsilon^{k-\lambda p}\int_Q e_\varepsilon(u) \ge K\int_Q |\nabla u|^{\lambda p}\psi(|u|)^{1-\lambda} \ge K\int_Q |\nabla(\Psi(|u|))|^{\lambda p}. \tag{3.11}$$

Combining (3.10) and (3.11), we obtain (3.9). □

We are now in a position to give the following proof.

Proof of Proposition 3.1. (i) Let $K > 0$ be chosen so that $\psi(t)^{1-k/p} \ge K\rho(t)$ for all $t \in [0, +\infty)$. Then (3.1) yields $F_\varepsilon(u_\varepsilon) \ge K\| \star J'u_\varepsilon\|_\Omega$, and the statement follows from the assumption $\sup_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$.

(ii) Let us define $l(\varepsilon) := \varepsilon^\alpha$, with $0 < \alpha < 1/(n+1)$. By Lemma 3.2, the assumption $\sup_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$, and assertion (i) is already proved. For every fixed ε we may position a grid $\mathcal{G}_\varepsilon$ of step $l(\varepsilon)$ so that both conditions below hold:

$$\sup_\varepsilon \left\{l(\varepsilon)^{n+1}\int_{R_{\varepsilon,k-1}\cap\Omega} e_\varepsilon(u_\varepsilon)\right\} < +\infty, \tag{3.12}$$

$$\sup_\varepsilon \left\{l(\varepsilon)^n \int_\Omega \frac{d\,|\star J'u_\varepsilon|(x)}{[\operatorname{dist}(x, R_{\varepsilon,k-1})]^n}\right\} < +\infty. \tag{3.13}$$

We claim that $|u_\varepsilon| \to 1$ uniformly on $R_{\varepsilon,k-1} \cap W$. To prove such a claim, we choose λ and Ψ as in the statement of Lemma 3.3, and we denote by Q a generic $(k-1)$-cell of $R_{\varepsilon,k-1}$ entirely contained in Ω. Since ψ vanishes only at 1 and Ψ is continuous and strictly increasing, it is enough to show that

$$\fint_Q \psi(|u_\varepsilon|) \to 0 \quad \text{and} \quad \operatorname{osc}(\Psi(|u_\varepsilon|), Q) \to 0 \quad \text{uniformly in } Q. \tag{3.14}$$

By Lemma 3.3, (3.12), and the choice of $l(\varepsilon)$, we have

$$\fint_Q \psi(|u_\varepsilon|) \le C\varepsilon^k l(\varepsilon)^{-k-n} = C\varepsilon^{k-\alpha(k+n)},$$
$$\operatorname{osc}(\Psi(|u_\varepsilon|), Q)^{\lambda p} \le C\varepsilon^{k-\lambda p} l(\varepsilon)^{\lambda p-k-n} = C\varepsilon^{k-\lambda p+\alpha(\lambda p-k-n)}.$$

Therefore, the first condition in (3.14) is fulfilled, because $\alpha < k/(k+n)$. As for the second condition in (3.14), it is satisfied provided $\lambda < (k+\alpha(\alpha-1)^{-1}n)/p$. This latter inequality turns out to be compatible (thanks to the choice α) with the one already imposed on λ (namely, $\lambda > (k-1)/p$). Thus (3.14) is proved.

Now, since the functions u_ε are smooth on W with $|u_\varepsilon| \to 1$ on $R_{\varepsilon,k-1} \cap W$, for ε small enough the currents obtained by extending to zero $\star J' u_\varepsilon$ out of W satisfy both conditions in (3.2) (indeed, for ε small enough, $|u_\varepsilon| \geq 1/2$ on $R_{\varepsilon,k-1} \cap W$ and, if $d = 2\sqrt{n+k}\, l(\varepsilon)$, V_d is compactly contained in W). Therefore, using (3.3), (3.4), and assertion (i) (already proved), we deduce that $\{(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon)\}$ is a sequence of polyhedral boundaries in V, with uniformly bounded masses and multiplicities in $\alpha_k\mathbb{Z}$ (cf. section 3.1). As $U \subset\subset V$, by an adaptation of the Federer–Fleming closure theorem (see [3, Proposition 2.8]), we can extract a subsequence such that $\mathbb{F}_U[(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon) - \alpha_k M] \to 0$, where M is an n-dimensional integral boundary with finite mass in U.

(iii) The asymptotical equivalence between $\{\star J u_\varepsilon\}$ and $\{\star J' u_\varepsilon\}$ follows from the assumption $\sup_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$, together with the following estimate, holding for all functions u smooth on U:

$$\mathbb{F}_U(\star J u - \star J' u) \leq K\varepsilon F_\varepsilon(u, U). \tag{3.15}$$

To show (3.15), observe first that $k\rho(|y|)\, dy$ is the differential of the $(k-1)$-form $\omega' := \sigma(|y|)\omega_0(y)$, where $\sigma(t) := kt^{-k}\int_0^t \rho(s)s^{k-1}\, ds$ and ω_0 is the $(k-1)$-form defined in (2.5). So there holds

$$\star J u - \star J' u = (-1)^n \frac{1}{k}\partial[\star u^\sharp(\omega_0 - \omega')]. \tag{3.16}$$

Moreover, the difference between the forms ω_0 and ω' satisfies the estimate

$$|\omega_0(y) - \omega'(y)| \leq K|1 - \sigma(|y|)|\, |y| \leq K||y| - 1| \leq KW(y)^{(p-k+1)/p}, \tag{3.17}$$

where the second inequality follows from the identity $\sigma(t) = t^{-k}$ for $t \geq 1/3$, and the third inequality holds due to the growth assumptions (1.2) on W.

Now, using the definition of flat norm together with (3.16), estimate (3.17), and Hölder's inequality, we infer

$$\begin{aligned}
\mathbb{F}_U(\star J u - \star J' u) &\leq K\| \star u^\sharp(\omega_0 - \omega')\|_U \leq K\int_U |\omega_0(u) - \omega'(u)||\nabla u|^{k-1} \\
&\leq K\int_U W(u)^{(p-k+1)/p}|\nabla u|^{k-1} \\
&\leq K\left\{\int_U W(u)\right\}^{(p-k+1)/p}\left\{\int_U |\nabla u|^p\right\}^{(k-1)/p} \\
&\leq K\left\{\varepsilon^k F_\varepsilon(u,U)\right\}^{(p-k+1)/p}\left\{\varepsilon^{k-p}F_\varepsilon(u,U)\right\}^{(k-1)/p} \\
&= K\varepsilon F_\varepsilon(u,U).
\end{aligned}$$

Thus (3.15) is proved, and it remains to show the asymptotical equivalence between $\{\star J' u_\varepsilon\}$ and $\{(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon)\}$. Since the currents obtained by extending to zero $\star J' u_\varepsilon$ out of W satisfy (3.5), and the grids $\mathcal{G}_\varepsilon$ verify (3.13), we have

$$\mathbb{F}_U\big(\star J' u_\varepsilon - (\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon)\big) \leq Kl(\varepsilon)\left\{l(\varepsilon)^n \int_\Omega \frac{d\,|\star J' u_\varepsilon|(x)}{[\mathrm{dist}(x, R_{\varepsilon,k-1})]^n}\right\} \to 0. \qquad \square$$

4. Optimal profile constants. Throughout this section, we assume that $n = 0$ (so, in particular, all functions in $W^{1,p}(\Omega, \mathbb{R}^k)$ are continuous by the Sobolev embedding theorem). Our goal is to introduce the family of constants τ_d which intervene in Theorem 1.1, and to state their main properties. We proceed according to the following plan. First, we give the strong definition of τ_d, prescribing the trace at infinity (cf. Definition 4.1), and we prove that τ_d is strictly positive for all $d \neq 0$. Next, we show that all functions u with finite energy on $\mathbb{R}^k$ admit a suitably defined notion of degree at infinity, and we introduce the corresponding relaxed definition of optimal profile constant $\overline{\tau}_d$ (cf. Definition 4.8). Finally, we prove that τ_d and $\overline{\tau}_d$ coincide and, as a consequence, that τ_d is a subadditive function of d.

DEFINITION 4.1. *For every $d \in \mathbb{Z}$, let ϕ_d be a fixed Lipschitz map from the sphere S^{k-1} into itself of degree d. We say that a function $u \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k)$ has* trace ϕ_d at infinity, *and we write* $\mathrm{tr}(u, \infty) = \phi_d$ *if there exists $\overline{r} > 0$ such that $u(x) = \phi_d(x/|x|)$ for all x with $|x| \geq \overline{r}$. Then we set*

$$\tau_d := \inf \left\{ F_1(u, \mathbb{R}^k) \ : \ u \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k) \text{ such that } \mathrm{tr}(u, \infty) = \phi_d \right\}. \tag{4.1}$$

Remark 4.2. It is straightforward to check that the energy $F_1(u, \mathbb{R}^k)$ is finite for all functions $u \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k)$ having trace ϕ_d at infinity.

PROPOSITION 4.3. *There exists a positive constant c, independent of d, such that $\tau_d \geq c|d|$ for every $d \in \mathbb{Z}$.*

Proof. Let $u \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k)$ with $u(x) = \phi_d(x/|x|)$ for $|x| \geq \overline{r}$. Using (3.1) and the area formula (2.7), we get

$$\begin{aligned} F_1(u, \mathbb{R}^k) &\geq K \int_{\mathbb{R}^k} W(u)^{1-k/p} |Ju| \geq K \int_{B^k_{\overline{r}}} W(u)^{1-k/p} |Ju| \\ &= K \int_{\mathbb{R}^k} W(y)^{1-k/p} \sharp \{u^{-1}(y) \cap B^k_{\overline{r}}\} \\ &\geq K \int_{\mathbb{R}^k} W(y)^{1-k/p} |\deg(u, B^k_{\overline{r}}, \mathbb{R}^k, y)| = K|d| \int_{B^k_1} W(y)^{1-k/p}, \end{aligned}$$

where the last equality holds because $u(\partial B^k_{\overline{r}}) \subseteq S^{k-1}$ (cf. section 2.4). ∎

DEFINITION 4.4. *Let a function $u \in C^0(\mathbb{R}^k; \mathbb{R}^k)$ be given. We say that a value $y \in \mathbb{R}^k$ is* not attained by u at infinity *if there exists $\overline{r} = \overline{r}(y) > 0$ such that $y \notin u(\mathbb{R}^k \setminus B^k_{\overline{r}})$. In this case, we can set*

$$\deg(u, \infty, y) := \deg\left(\frac{u - y}{|u - y|}, \partial B^k_r \right) \quad \forall r \geq \overline{r}(y),$$

since the right-hand side of the above definition does not depend on $r \geq \overline{r}(y)$.

LEMMA 4.5. *Let $F_1(u, \mathbb{R}^k) < +\infty$. Then the following hold:*

(i) *For all $\alpha > 0$, B^k_α contains some value not attained by u at infinity.*

(ii) *There exists a sequence of radii $\{r_n\}$, with $\lim_n r_n = +\infty$, such that*

$$\lim_n \left\{ r_n \int_{\partial B^k_{r_n}} e_1(u) \right\} = 0 \quad \text{and} \quad \lim_n \left\{ \sup_{\partial B^k_{r_n}} ||u| - 1| \right\} = 0. \tag{4.2}$$

(iii) *For any two values $y_1, y_2 \in B^k_1$ not attained by u at infinity, there holds $\deg(u, \infty, y_1) = \deg(u, \infty, y_2)$.*

Before proving the above lemma, we add a related definition and remark which will be useful in what follows.

DEFINITION 4.6. *Let $F_1(u,\mathbb{R}^k) < +\infty$. Then, we call the* degree at infinity of *u—and denote it by* $\deg(u,\infty)$*—the quantity* $\deg(u,\infty,y)$ *for* any *value $y \in B_1^k$ not attained by u at infinity. Indeed, such a value exists in view of Lemma* 4.5 (i)*, and it does not depend on the choice of $y \in B_1^k$ in view of Lemma* 4.5 (iii).

Remark 4.7. (i) Let $F_1(u,\mathbb{R}^k) < +\infty$. For every $\delta > 0$, we may find a radius R (arbitrarily large) such that

$$R\int_{\partial B_R^k} e_1(u) < \delta\ , \quad \big||u|-1\big| < \delta \quad \text{on } \partial B_R^k\ , \quad \deg\left(\frac{u}{|u|},\partial B_R^k\right) = \deg(u,\infty). \tag{4.3}$$

Indeed, by Lemma 4.5 we can select a value $y \in B_{1-2\delta}^k$ and an arbitrarily large radius R such that $y \notin u(\mathbb{R}^k \setminus B_R)$ and the first two conditions in (4.3) are satisfied. Then also the third one is fulfilled, as

$$\deg\left(\frac{u}{|u|},\partial B_R^k\right) = \deg\left(\frac{u-y}{|u-y|},\partial B_R^k\right) = \deg(u,\infty,y) = \deg(u,\infty).$$

(ii) Let $F_1(u,\mathbb{R}^k) < +\infty$, and assume in addition that $|u| \geq 1/2$ on ∂Q_n for a sequence of cubes Q_n centered at the origin, whose sides diverge as $n \to +\infty$. Then, by arguing in a similar way as in the item (i) above, one obtains that $\deg(u,\infty) = \deg(u/|u|,\partial Q_n)$ for n large enough.

Proof of Lemma 4.5. (i) Let $\alpha > 0$. By the assumption $F_1(u,\mathbb{R}^k) < +\infty$ and (3.1), we may choose $r = r(\alpha)$ large enough so that

$$\int_{\mathbb{R}^k\setminus B_r^k} W(u)^{1-k/p}|Ju| < \int_{B_\alpha^k} W(y)^{1-k/p}. \tag{4.4}$$

Since by the area formula (2.7) the left-hand side of (4.4) is larger than or equal to $\int_{B_\alpha^k}[W(y)^{1-k/p}\sharp\{u^{-1}(y)\cap(\mathbb{R}^k\setminus B_r^k)\}]$, we deduce that $\sharp\{u^{-1}(y)\cap(\mathbb{R}^k\setminus B_r^k)\}$ cannot be strictly positive for all $y \in B_\alpha^k$.

(ii) By the assumption $F_1(u,\mathbb{R}^k) < +\infty$, a contradiction argument readily yields the existence of a sequence of radii $\{r_n\}$, with $\lim_n r_n = +\infty$, satisfying the first equality in (4.2). Let us show that this sequence also fulfills the second condition in (4.2). Choose λ and Ψ as in Lemma 3.4, and notice that such a lemma still holds (with the very same proof) when u belongs to $W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k;\mathbb{R}^k)$ and the $(k-1)$-cell Q of size l is replaced by ∂B_l^k. Then inequalities (3.8) and (3.9), applied with $n = 0$ and $\varepsilon = 1$, give

$$\fint_{\partial B_{r_n}^k}\psi(|u|) \leq K\, r_n^{-k}\left\{r_n\int_{\partial B_{r_n}^k} e_1(u)\right\},$$
$$\mathrm{osc}(\Psi(|u|),\partial B_{r_n}^k)^{\lambda p} \leq K\, r_n^{\lambda p-k}\left\{r_n\int_{\partial B_{r_n}^k} e_1(u)\right\}.$$

We conclude by taking $\lambda \leq k/p$ and arguing as in the proof of Proposition 3.1.

(iii) Let $y_1, y_2 \in B_1^k$ be two values not attained by u at infinity, and let r_1, r_2 be the corresponding radii such that $y_i \notin u(\mathbb{R}^k \setminus B_{r_i}^k)$ for $i = 1,2$. Since $|y_i| < 1$, by assertion (ii) (already proved), we may find a radius $\overline{r} > \max\{r_1, r_2\}$ such that

$|u(x)| > \max\{|y_1|, |y_2|\}$ for all $x \in \partial B^k_{\overline{r}}$. Therefore,

$$\begin{aligned}\deg(u, \infty, y_1) &= \deg\left(\frac{u(x)-y_1}{|u(x)-y_1|}, \partial B^k_{\overline{r}}\right) = \deg\left(\frac{u(x)-y_2}{|u(x)-y_2|}, \partial B^k_{\overline{r}}\right)\\ &= \deg(u, \infty, y_2). \qquad \square\end{aligned}$$

DEFINITION 4.8. *For every $d \in \mathbb{Z}$, we set*

$$\overline{\tau}_d := \inf\big\{F_1(u, \mathbb{R}^k) \ : \ u \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k) \ \text{ such that } \ \deg(u, \infty) = d\big\}. \tag{4.5}$$

PROPOSITION 4.9. *For every $d \in \mathbb{Z}$, the constants τ_d and $\overline{\tau}_d$ defined, respectively, by* (4.1) *and* (4.5) *coincide. In particular, τ_d equals τ_{-d}, and it is independent of the choice of ϕ_d.*

Proof. We introduce an auxiliary quantity τ'_d defined by

$$\begin{aligned}\tau'_d := \inf\big\{&F_1(u, \mathbb{R}^k) \ : \ u \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k) \text{ such that } \exists\, \overline{r} > 0 \text{ with}\\ &|u(x)| = 1 \text{ on } \mathbb{R}^k \setminus B^k_r \ \text{ and } \deg(u, \partial B^k_r) = d \ \ \forall r \geq \overline{r}\big\}.\end{aligned} \tag{4.6}$$

Denoting, respectively, by $\mathcal{A}_d$, $\mathcal{A}'_d$, and $\overline{\mathcal{A}}_d$ the classes of admissible functions in the right-hand sides of (4.1), (4.6), and (4.5), there holds $\mathcal{A}_d \subset \mathcal{A}'_d \subset \overline{\mathcal{A}}_d$. Therefore, $\tau_d \geq \tau'_d \geq \overline{\tau}_d$. We can then achieve the proof in two steps by showing first that $\tau_d \leq \tau'_d$ and then that $\tau'_d \leq \overline{\tau}_d$.

Step 1. To prove that $\tau_d \leq \tau'_d$, let us show that, for any given $u \in \mathcal{A}'_d$ with $F_1(u, \mathbb{R}^k) < +\infty$ and for any $\varepsilon > 0$, there exists $v \in \mathcal{A}_d$ such that

$$F_1(v, \mathbb{R}^k) \leq F_1(u, \mathbb{R}^k) + \varepsilon. \tag{4.7}$$

Let $\overline{r} > 0$ be such that $|u(x)| = 1$ on $\mathbb{R}^k \setminus B^k_r$ and $\deg(u, \partial B^k_r) = d$ for all $r \geq \overline{r}$. For any $r_1 \geq \overline{r}$, the continuous mapping $g_{r_1} : S^{k-1} \to S^{k-1}$ defined by $g_{r_1}(y) := u(r_1 y)$ has degree d. Therefore, we may find a Lipschitz homotopy $\varphi : S^{k-1} \times [0,1] \to S^{k-1}$ such that $\varphi(\cdot, 0) = g_{r_1}(\cdot)$ and $\varphi(\cdot, 1) = \phi_d(\cdot)$. For $r_2 > r_1$, let $\lambda : [r_1, r_2] \to [0,1]$ be an affine function with $\lambda(r_1) = 0$ and $\lambda(r_2) = 1$. We set

$$v(x) := \begin{cases} u(x) & \text{if } |x| \leq r_1,\\ \varphi\big(\frac{x}{|x|}, \lambda(|x|)\big) & \text{if } r_1 \leq |x| \leq r_2,\\ \phi_d\big(\frac{x}{|x|}\big) & \text{if } |x| \geq r_2.\end{cases}$$

Clearly, $v \in \mathcal{A}_d$, and we claim that (4.7) is fulfilled provided r_1 and r_2 are suitably chosen. Indeed, we have

$$F_1(v, \mathbb{R}^k) - F_1(u, \mathbb{R}^k) = \underbrace{F_1(v, \mathbb{R}^k \setminus B^k_{r_2})}_{\mathrm{I}} + \underbrace{F_1(v, B^k_{r_2} \setminus B^k_{r_1})}_{\mathrm{II}} - \underbrace{F_1(u, \mathbb{R}^k \setminus B^k_{r_1})}_{\mathrm{III}}.$$

Term I is infinitesimal when $r_2 \to +\infty$, as $F_1(v, \mathbb{R}^k) < +\infty$ (cf. Remark 4.2). Similarly, term III is infinitesimal when $r_1 \to +\infty$ by the assumption $F_1(u, \mathbb{R}^k) < +\infty$. As for term II, we have $\mathrm{II} = \int_{B^k_{r_2} \setminus B^k_{r_1}} |\nabla v|^p$, and

$$|\nabla v(x)| \leq \Big\{\sup_{S^{k-1} \times [0,1]} |\nabla \varphi|\Big\}\left\{\frac{(k-1)^{1/2}}{|x|} + \frac{1}{r_2 - r_1}\right\} \qquad \forall x \in B^k_{r_2} \setminus B^k_{r_1}.$$

Then, if we choose $r_2 = r_1(1 + r_1^{-\alpha})$, with $0 < \alpha < (p-k)/(p-1)$, term II turns out to be infinitesimal when $r_1 \to +\infty$, because

$$\int_{B^k_{r_2}\setminus B^k_{r_1}} \frac{1}{|x|^p} = K r_1^{k-p}\big[1 - (1 + r_1^{-\alpha})^{k-p}\big] \sim r_1^{k-p-\alpha},$$
$$\int_{B^k_{r_2}\setminus B^k_{r_1}} \frac{1}{(r_2 - r_1)^p} = K r_1^{k+(\alpha-1)p}\big[(1 + r_1^{-\alpha})^k - 1\big] \sim r_1^{k-p+\alpha(p-1)}.$$

Step 2. To prove that $\tau'_d \le \overline{\tau}_d$, let us show that, for any given $u \in \overline{\mathcal{A}}_d$ with $F_1(u, \mathbb{R}^k) < +\infty$, and for any $\varepsilon > 0$, there exists $v \in \mathcal{A}'_d$ such that (4.7) holds. For a fixed $\delta > 0$, we can choose $R > 0$ arbitrarily large such that conditions (4.3) in Remark 4.7 (i) hold. Let $\alpha : S^{k-1} \times [R, +\infty) \to \mathbb{R}$ be defined by

$$\alpha(y,t) := \begin{cases} 1 + (t-R)\mathrm{sgn}\big(1 - |u(Ry)|\big) & \text{if } 0 \le t - R \le \big||u(Ry)|^{-1} - 1\big|, \\ |u(Ry)|^{-1} & \text{if } t - R \ge \big||u(Ry)|^{-1} - 1\big|, \end{cases}$$

and set

$$v(x) := \begin{cases} u(x) & \text{if } |x| \le R, \\ \alpha\big(\frac{x}{|x|}, |x|\big)\, u\big(R\frac{x}{|x|}\big) & \text{if } |x| \ge R. \end{cases}$$

Notice that $v(x) = |u(Rx/|x|)|^{-1} u(Rx/|x|)$ for all x with $|x| \ge R' := R + \sup_{y \in S^{k-1}} \big||u(Ry)|^{-1} - 1\big|$, and hence $v \in \mathcal{A}'_d$. We claim that v fulfills (4.7) provided δ and R are chosen, respectively, sufficiently small and sufficiently large. Indeed, we have

$$F_1(v, \mathbb{R}^k) - F_1(u, \mathbb{R}^k) = \underbrace{\int_{\mathbb{R}^k \setminus B^k_R} |\nabla v|^p}_{\mathrm{I}} + \underbrace{\int_{B^k_{R'} \setminus B^k_R} W(v)}_{\mathrm{II}} - \underbrace{F_1(u, \mathbb{R}^k \setminus B^k_R)}_{\mathrm{III}}.$$

Term III is infinitesimal as $R \to +\infty$ since $F_1(u, \mathbb{R}^k) < +\infty$. Term II is infinitesimal as $\delta \to 0$. Indeed, by the second condition in (4.3) the difference $(R' - R)$ becomes arbitrarily small as $\delta \to 0$, so that the measure of the integration domain in II is infinitesimal with δ (and clearly the integrand remains bounded, because $|v(x)| \le (1 + R' - R)(1 + \delta)$ for all $x \in B^k_{R'} \setminus B^k_R$). Concerning term I, for $|x| \ge R$ we have

$$|\nabla v(x)| \le K\bigg\{|\alpha|\,|\nabla u|\frac{R}{|x|} + |u|\,|\alpha_y|\frac{1}{|x|} + |u|\,|\alpha_t|\bigg\}; \tag{4.8}$$

here and below, α and its derivatives are computed at $(x/|x|, |x|)$, while u and its gradient are computed at $Rx/|x|$. Now, for a suitable constant C_δ which depends on δ but remains bounded as $\delta \to 0$, there holds

$$|\alpha| \le C_\delta, \qquad |u| \le C_\delta, \qquad |\alpha_y| \le C_\delta R|\nabla u|. \tag{4.9}$$

From (4.8) and (4.9), we infer that

$$|\nabla v| \le C_\delta\bigg\{|\nabla u|\frac{R}{|x|} + |\alpha_t|\bigg\}.$$

Taking into account that $\alpha_t(y,t)$ vanishes for $t \ge R'$, and that, by the first condition in (4.3),

$$\int_{\mathbb{R}^k \setminus B^k_R} \bigg|\nabla u\Big(R\frac{x}{|x|}\Big)\bigg|^p \frac{R^p}{|x|^p} = KR\int_{\partial B^k_R} |\nabla u|^p \le K\delta,$$

we deduce that term I is infinitesimal as $\delta \to 0$.

Finally, the last assertion in the statement follows immediately from the definition of $\bar\tau_d$. ∎

PROPOSITION 4.10. *The function $\mathbb{Z} \ni d \mapsto \tau_d \in [0, +\infty)$ is subadditive.*

Proof. By Proposition 4.9, it is enough to prove that, for fixed integers d_1 and d_2, $\overline\tau_{d_1+d_2} \le \tau_{d_1} + \tau_{d_2}$. Moreover, by the same proposition we know that τ_{d_i} are independent of the choice of the trace functions ϕ_{d_i}. Hence, we may assume with no loss of generality that there exists a fixed vector $y_0 \in S^{k-1}$ such that $\phi_{d_1}(y) = y_0$ for y in the hemisphere $S^{k-1} \cap \{y_k \ge 0\}$ and $\phi_{d_2}(y) = y_0$ for y in the hemisphere $S^{k-1} \cap \{y_k \le 0\}$ (here and in the remainder of the proof, we denote by y_k the kth component of a vector $y \in \mathbb{R}^k$). Then, if $\mathcal{A}_d$ and $\overline{\mathcal{A}}_d$ denote, respectively, the classes of admissible functions in (4.1) and (4.5), it suffices to construct, for any given pair $(u_1, u_2) \in \mathcal{A}_{d_1} \times \mathcal{A}_{d_2}$ with $F_1(u_i, \mathbb{R}^k) < +\infty$, a function $v \in \overline{\mathcal{A}}_{d_1+d_2}$ such that $F_1(v, \mathbb{R}^k) = F_1(u_1, \mathbb{R}^k) + F_1(u_2, \mathbb{R}^k)$. Choose $r > \max\{r_1, r_2\}$, with r_i satisfying $u_i(x) = \phi_{d_i}(x/|x|)$ when $|x| \ge r_i$. Then set $P_r := (0, \dots, 0, r)$, and define v by

$$v(x) := \begin{cases} u_1(x + P_r) & \text{if } x_k \le 0, \\ u_2(x - P_r) & \text{if } x_k > 0. \end{cases}$$

Since $u_1(x) = y_0$ on $\{x_k \ge r\}$ and $u_2(x) = y_0$ on $\{x_k \le -r\}$, it is straightforward that $F_1(v, \mathbb{R}^k) = F_1(u_1, \mathbb{R}^k) + F_1(u_2, \mathbb{R}^k)$. It remains to check that $v \in \overline{\mathcal{A}}_{d_1+d_2}$. By construction, there holds

$$v(x) = \begin{cases} \phi_{d_1}\left(\frac{x+P_r}{|x+P_r|}\right) & \text{on } \{|x + P_r| \ge r_1\} \cap \{x_k \le 0\}, \\ \phi_{d_2}\left(\frac{x-P_r}{|x-P_r|}\right) & \text{on } \{|x - P_r| \ge r_2\} \cap \{x_k > 0\} \end{cases}$$

(in particular, $v(x) = y_0$ on $\{|x + P_r| \ge r_1\} \cap \{|x - P_r| \ge r_2\} \cap \{|x_k| \le r\}$). So, if we consider a cube Q centered at the origin, with sides larger than $4r$, we have $|v| = 1$ on ∂Q and $\deg(v, \partial Q) = d_1 + d_2$. In view of Remark 4.7 (ii), we conclude that $\deg(v, \infty) = d_1 + d_2$. ∎

5. Γ-lim inf inequality.

5.1. Case $n = 0$. The reason why we first deal with such a case relies basically on the following *scaling property*, enjoyed by the functionals F_ε for $n = 0$, as it is easily checked by a change of variables:

$$F_1\left(u(\varepsilon x), \frac{A}{\varepsilon}\right) = F_\varepsilon(u(x), A). \tag{5.1}$$

Thanks to (5.1), we can exploit the family of optimal profile constants introduced in section 4 in order to prove the following version of the Γ-lim inf inequality in Theorem 1.1 (i). Until otherwise specified, we keep the assumption $n = 0$ without any further mention.

THEOREM 5.1. *Let $\{u_\varepsilon\} \subset W^{1,p}(\Omega; \mathbb{R}^k)$ be a sequence such that $\mathbb{F}_\Omega(\star J u_\varepsilon - \alpha_k M) \to 0$ as $\varepsilon \to 0$, where $M = \sum_{i=1}^m \sigma^i \delta(x^i)$ is a 0-dimensional integral boundary (with $\sigma^i \in \mathbb{Z}$ and $x^i \in \Omega$). Then $\liminf_\varepsilon F_\varepsilon(u_\varepsilon) \ge \sum_{i=1}^m \tau_{\sigma^i}$.*

The proof of Theorem 5.1 given below may be outlined as follows. We consider the sequence of grids $\mathcal{G}_\varepsilon$ of step $l(\varepsilon) \gg \varepsilon$ constructed in Proposition 3.1, and we use (5.1) in order to scale the functionals F_ε over those cubes of the grids where the degree is located. Then, since the size of the scaled cubes diverges as $\varepsilon \to 0$, we need as a

fundamental ingredient the following estimate for the energy F_1 over enlarging cubes, whose proof is postponed.

PROPOSITION 5.2. *Let $Q_r := [-r,r]^k$, and let $u_r \in W^{1,p}(Q_r;\mathbb{R}^k)$ satisfy the following assumptions:*

$$\sup_r F_1(u_r,\, Q_r) < +\infty, \tag{5.2}$$

$$\sup_r \left\{ r \int_{\partial Q_r} e_1(u_r) \right\} < +\infty, \tag{5.3}$$

$$|u_r| \geq 1/2 \text{ on } \partial Q_r, \tag{5.4}$$

$$\deg\left(\frac{u_r}{|u_r|}, \partial Q_r\right) = d. \tag{5.5}$$

Then $\liminf_{r\to+\infty} F_1(u_r, Q_r) \geq \tau_d$.

Proof of Theorem 5.1. We may assume with no loss of generality that u_ε are smooth and that $\liminf_\varepsilon F_\varepsilon(u_\varepsilon) = \lim_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$. Given open sets U, V, and W, with $U \subset\subset V \subset\subset W \subset\subset \Omega$, we take a sequence of grids $\mathcal{G}_\varepsilon$ as in Proposition 3.1, and we denote by Q_ε^j for $j = 1,\dots,N(\varepsilon)$ those cubes of $\mathcal{G}_\varepsilon$ whose centers x_ε^j belong to U; for ε small enough, such cubes are entirely contained in Ω. If we equip Q_ε^j with the canonical orientation inherited by $\mathbb{R}^k$, the restrictions of the currents $(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon)$ to U may be written in the form

$$(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon) \llcorner U = \alpha_k \sum_{j=1}^{N(\varepsilon)} \sigma_\varepsilon^j \delta(x_\varepsilon^j), \tag{5.6}$$

where $\sigma_\varepsilon^j = \deg(u_\varepsilon/|u_\varepsilon|, \partial Q_\varepsilon^j)$. Since the currents $(\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon)$ are uniformly bounded in mass over U, and the coefficients σ_ε^j are integers, we deduce that the sum in (5.6) can be actually extended over $j \in \{1,\dots,N\}$ for some N independent of ε; moreover, up to subsequences there exist integers $\sigma_0^1,\dots,\sigma_0^N$ such that $\sigma_\varepsilon^j \equiv \sigma_0^j$ for ε small enough. Using (5.1), we have

$$\liminf_\varepsilon F_\varepsilon(u_\varepsilon) \geq \sum_{j=1}^N \liminf_\varepsilon F_1\left(u_\varepsilon(\varepsilon x + x_\varepsilon^j), \frac{Q_\varepsilon^j}{\varepsilon} - x_\varepsilon^j\right). \tag{5.7}$$

Set $r = r(\varepsilon) := l(\varepsilon)/\varepsilon$ (so that $\{r(\varepsilon)\}$ diverges as $\varepsilon \to 0$ because $l(\varepsilon) \gg \varepsilon$), and fix an index $j \in \{1,\dots,N\}$. We have $(Q_\varepsilon^j/\varepsilon) - x_\varepsilon^j = [-r,r]^k$, and we claim that the sequence $u_r(x) := u_\varepsilon(\varepsilon x + x_\varepsilon^j)$ satisfies all the hypotheses of Proposition 5.2. Indeed, (5.2) holds by (5.7) and the assumption $\lim_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$. By the initial choice of the grid, (5.3) is satisfied (due to (3.12) in the proof of Proposition 3.1), and (5.4) holds for r large enough. Moreover, still for r large enough, we know that $\deg(u_r/|u_r|, \partial Q_r) = \sigma_0^j$. Thus, we may invoke Proposition 5.2 to obtain

$$\sum_{j=1}^N \liminf_\varepsilon F_1\left(u_\varepsilon(\varepsilon x + x_\varepsilon^j), \frac{Q_\varepsilon^j}{\varepsilon} - x_\varepsilon^j\right) \geq \sum_{j=1}^N \tau_{\sigma_0^j}. \tag{5.8}$$

Now observe that, since $\overline{U}$ is compact, up to subsequences there exist points $x_0^1,\dots,x_0^N \in \overline{U}$ such that $\lim_\varepsilon x_\varepsilon^j = x_0^j$ for $j = 1,\dots,N$. Let $\{x^1,\dots,x^h\} := \{x^1,\dots,x^m\} \cap \overline{U}$ (with $h \leq m$). Since by Proposition 3.1 (iii) $\mathbb{F}_U((\Phi_\varepsilon)_\sharp(\star J' u_\varepsilon) - \alpha_k M) \to 0$, we necessarily have $x_0^j \in \{x^1,\dots,x^h\}$ for all $j = 1,\dots,N$, and, setting $\mathcal{J}_i := \{j \in \{1,\dots,N\} :$

$x_0^j = x^i\}$, there holds $\sum_{j\in\mathcal{J}_i}\sigma_0^j = \sigma^i$ for all $i = 1,\dots,h$. Then Proposition 4.10 yields

$$\sum_{j=1}^{N}\tau_{\sigma_0^j} = \sum_{i=1}^{h}\sum_{j\in\mathcal{J}_i}\tau_{\sigma_0^j} \geq \sum_{i=1}^{h}\tau_{\sum_{j\in\mathcal{J}_i}\sigma_0^j} = \sum_{i=1}^{h}\tau_{\sigma^i}. \tag{5.9}$$

The thesis follows from (5.7), (5.8), and (5.9) by taking into account the arbitrariness of U. □

We now turn to the proof of Proposition 5.2. What makes it delicate is the lack of compactness of condition (5.5), as the boundary of Q_r, where the degree is prescribed, is moving to infinity. To overcome this difficulty, we first need to suitably extend u_r on a cube Q_η larger than Q_r (for technical reasons which will be clear in the proof of Proposition 5.2); then, we build a sequence of grids, of fixed step ρ independent of r, such that the functions u_r remain in modulus sufficiently far from zero on their $(k-1)$-skeletons; see Lemma 5.3 below. These grids play a crucial role in the proof of Proposition 5.2, since they allow one to individuate for every r a finite number of cubes which carry the degree. Such cubes can be gathered into some separate "clusters," which possibly go to infinity with r but move away from each other. The scope is "catching" such clusters one by one, in order to recover a condition of prescribed degree at infinity. The concluding argument will be again the subadditivity of τ_d proved in section 4.

LEMMA 5.3. *Let $Q_r = [-r,r]^k$, and let $u_r \in W^{1,p}(Q_r;\mathbb{R}^k)$ satisfy assumptions* (5.2), (5.3), *and* (5.4). *For every r, choose $\eta = \eta(r)$ so that $\eta(r) > r$, and $\eta(r)/r \to 1$ as $r \to +\infty$. Then we may construct extensions u_η of u_r to $Q_\eta = [-\eta,\eta]^k$ with the following properties:*

$$|u_\eta| \geq \frac{1}{2} \quad \text{on } \partial Q_\eta, \tag{5.10}$$

$$\lim_{r\to+\infty} F_1(u_\eta, Q_\eta \setminus Q_r) = 0. \tag{5.11}$$

Further, if $\rho > 0$ is sufficiently large, we may position a sequence of grids $\mathcal{G}_r^\rho$ of fixed step ρ (independent of r) such that

$$|u_\eta| \geq 1/2 \text{ on } R_{r,k-1}\cap Q_\eta, \tag{5.12}$$

$$\deg\left(\frac{u_\eta}{|u_\eta|}, \partial(Q_r^i \cap (Q_\eta \setminus Q_r))\right) = 0 \quad \text{for } r \text{ large enough}, \tag{5.13}$$

$$\sup_r \sum_i \left|\deg\left(\frac{u_\eta}{|u_\eta|}, \partial(Q_r^i \cap Q_\eta)\right)\right| < +\infty, \tag{5.14}$$

where in (5.12) *$R_{r,k-1}$ denotes the $(k-1)$-skeleton of $\mathcal{G}_r^\rho$, in* (5.13) *Q_r^i is any cube of $\mathcal{G}_r^\rho$ which intersects $Q_\eta \setminus Q_r$, and in* (5.14) *the sum is extended over all the cubes Q_r^i of $\mathcal{G}_r^\rho$ which intersect Q_η.*

Proof. Thanks to assumptions (5.3) and (5.4) and to the hypothesis on the asymptotic behavior of $\eta(r)$ for $r \to +\infty$, extensions u_η of u_r to Q_η satisfying (5.10) and (5.11) can be constructed essentially by projection on ∂Q_r (we omit the explicit computations which are tedious but not difficult).

If u_r satisfy in addition assumption (5.2), due to property (5.11) we also have $\sup_r F_1(u_\eta, Q_\eta) < +\infty$. Then, by inequality (3.6) in Lemma 3.2 (applied with $n = 0$), for any $\rho > 0$ we may position a sequence of grids $\mathcal{G}_r^\rho$ of step ρ so that $\rho\int_{R_{r,k-1}\cap Q_\eta} e_1(u_\eta) \leq C$, where C is a positive constant independent of r and ρ. We claim that if ρ is

large enough, such a sequence of grids fulfills (5.12), (5.13), and (5.14). To obtain (5.12) it is enough to show that, for λ and Ψ as in Lemma 3.3, and denoting by Q a generic $(k-1)$-cell of $R_{r,k-1}$, the functions u_η satisfy

$$\lim_{\rho\to+\infty} \fint_{Q\cap Q_\eta} \psi(|u_\eta|) = 0 \quad \text{and} \quad \lim_{\rho\to+\infty} \operatorname{osc}(\Psi(|u_\eta|), Q\cap Q_\eta) = 0$$

(uniformly in r and Q).

These equalities can be verified (provided $\lambda < k/p$), in a similar way as already done in Proposition 3.1 and Lemma 4.5. So we avoid repeating the proof, and we pass to show (5.13) and (5.14). Since $\lim_r F_1(u_\eta, Q_\eta \setminus Q_r) = 0$ and $\sup_r F_1(u_\eta, Q_\eta) < +\infty$, it suffices to prove that

$$\sum_i \left| \deg\left(\frac{u_\eta}{|u_\eta|}, \partial(Q_r^i \cap (Q_\eta \setminus Q_r))\right)\right| \le K F_1(u_\eta, Q_\eta \setminus Q_r), \tag{5.15}$$

$$\sum_i \left| \deg\left(\frac{u_\eta}{|u_\eta|}, \partial(Q_r^i \cap Q_\eta)\right)\right| \le K F_1(u_\eta, Q_\eta), \tag{5.16}$$

where the sums are extended, respectively, over the indices i such that Q_r^i intersects $Q_\eta \setminus Q_r$ and Q_η. (We stress that the degrees in the left-hand sides of (5.15) and (5.16) are well defined by (5.4), (5.10), and (5.12).) We prove only (5.15), omitting the proof of (5.16) which runs in the exact same way. Set for brevity $P_r^i := Q_r^i \cap (Q_\eta \setminus Q_r)$ whenever such a set is nonempty. Using (3.1) and the area formula (2.7), we get

$$\begin{aligned}
\int_{Q_\eta\setminus Q_r} e_1(u_\eta) &\ge K \int_{Q_\eta\setminus Q_r} W^{1-k/p}(u_\eta)|Ju_\eta| \\
&= K \sum_i \int_{P_r^i} W^{1-k/p}(u_\eta)|Ju_\eta| \\
&= K \sum_i \int_{\mathbb{R}^k} W^{1-k/p}(y)\, \sharp\{u_\eta^{-1}(y)\cap P_r^i\} \\
&\ge K \sum_i \int_{B^k_{1/2}} W^{1-k/p}(y)\, |\deg(u_\eta, P_r^i, \mathbb{R}^k, y)|.
\end{aligned} \tag{5.17}$$

For every i, since $u_\eta \in C^0(\overline{P}_r^i; \mathbb{R}^k)$ and $|u_\eta| \ge 1/2$ on ∂P_r^i, we may find a continuous homotopy $\varphi(x,t) : \overline{P}_r^i \times [0,1] \to \mathbb{R}^k$ such that $\varphi(\cdot,0) = u_\eta(\cdot)$, $\varphi(\cdot,1) = w_\eta(\cdot)$ with $w_\eta = u_\eta/|u_\eta|$ on ∂P_r^i, and $\varphi(\cdot,t) \ge 1/2$ on ∂P_r^i for all $t \in [0,1]$. Therefore, for all $y \in B^k_{1/2}$, we have

$$\deg(u_\eta, P_r^i, \mathbb{R}^k, y) = \deg(w_\eta, P_r^i, \mathbb{R}^k, y) = \deg(w_\eta, \partial P_r^i) = \deg(u_\eta/|u_\eta|, \partial P_r^i),$$

where the first equality is due to the homotopy invariance of the degree (cf. [14, Theorem 2.3]) and the second equality holds because $w_\eta(\partial P_r^i) \subseteq S^{k-1}$ (cf. section 2.4). Replacing the above identity in the last line of (5.17), we get (5.15). ∎

Proof of Proposition 5.2. Let $\eta = \eta(r)$ satisfy the assumptions of Lemma 5.3 and the additional condition $\lim_{r\to+\infty} \big(\eta(r) - r\big) = +\infty$. For such $\eta(r)$, let u_η and $\mathcal{G}_r^\rho$ be, respectively, extensions of u_r to Q_η and grids of step ρ as in the statement of Lemma 5.3. By (5.11), it is enough to show that $\liminf F_1(u_\eta, Q_\eta) \ge \tau_d$.

For any cube Q_r^i of $\mathcal{G}_r^\rho$ which intersects Q_η, we set $d_r^i := \deg(u_\eta/|u_\eta|, \partial(Q_r^i \cap Q_\eta))$. By (5.14), there exists a finite number N (independent of r) of cubes Q_r^i such that $d_r^i \neq 0$, and we have $\sup_r \sum_{i=1}^N |d_r^i| < +\infty$. Hence, up to subsequences, $d_r^i \equiv d^i$, with d^i independent of r for $i = 1, \dots, N$.

By (5.13), for all $i = 1, \dots, N$ and r large enough, it must be $Q_r^i \cap Q_r \neq \emptyset$, $Q_r^i \subset Q_\eta$, and $d^i = \deg(u_r/|u_r|, \partial(Q_r^i \cap Q_r))$. Then, by (5.5), we have $\sum_{i=1}^N d^i = d$.

Now, in order to "follow" the location of the cubes Q_r^i as $r \to +\infty$, let us introduce an equivalence relation on the set of indices $\{1, \dots, N\}$. For $i = 1, \dots, N$, let x_r^i denote the center of Q_r^i. We say that i is equivalent to j if $\sup_r |x_r^i - x_r^j| < +\infty$. Possibly passing to subsequences, we may assume that, whenever i and j are *not* equivalent, there holds $\lim_r |x_r^i - x_r^j| = +\infty$. Then, we can associate to each equivalence class $[i]$ a sequence of sets $E_r^{[i]}$, each one contained in Q_η and delimited by $\mathcal{G}_r^\rho$, enjoying the following properties:

$$\text{for every } r,\ E_r^{[i]} \cap E_r^{[j]} = \emptyset \text{ whenever } [i] \neq [j]; \tag{5.18}$$

$$\text{for every } r,\ E_r^{[i]} \text{ contains } Q_r^j \text{ if and only if } j \in [i]; \tag{5.19}$$

$$\text{for any } j \in [i],\ \text{there holds } \bigcup_r (E_r^{[i]} - x_r^j) = \mathbb{R}^k. \tag{5.20}$$

Notice in particular that it is possible to satisfy (5.20) thanks to the assumption $\lim_{r\to+\infty} (\eta(r) - r) = +\infty$, which ensures that $\lim_{r\to+\infty} \operatorname{dist}(x_r^j, \partial Q_\eta) = +\infty$.

Notice also that if $d_r^{[i]}$ denotes the degree of $u_\eta/|u_\eta|$ on $\partial E_r^{[i]}$, by (5.19) $d_r^{[i]}$ is just the sum of all the d^j such that $j \in [i]$. In particular, $d_r^{[i]} =: d^{[i]}$ is independent of r, and we have

$$\sum_{[i]} d^{[i]} = \sum_{[i]} \Big(\sum_{j\in[i]} d^j \Big) = \sum_{j=1}^N d^j = d. \tag{5.21}$$

We can now begin to estimate the energy of u_η on Q_η. By (5.18), we have

$$\liminf_r F_1(u_\eta, Q_\eta) \geq \sum_{[i]} \liminf_r F_1(u_\eta, E_r^{[i]}) = \sum_{[i]} \liminf_r F_1\big(u_\eta(x + x_r^i), E_r^{[i]} - x_r^i\big),$$

where for every class $[i]$ we have chosen a representative i, which will stay fixed in the remainder of the proof. Taking (5.20) into account, we infer that, up to subsequences, there exist functions $u^i \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k)$ such that $u_\eta(x + x_r^i) \to u^i(x)$ weakly in $W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k)$, and strongly in $L^\infty_{\mathrm{loc}}(\mathbb{R}^k, \mathbb{R}^k)$. Since, for any open set A, $F_1(\cdot, A)$ is weakly lower semicontinuous on $W^{1,p}(A; \mathbb{R}^k)$, we have

$$\liminf_r F_1\big(u_\eta(x + x_r^i), E_r^{[i]} - x_r^i\big) \geq \sup_{A\subset\subset\mathbb{R}^k} \liminf_r F_1\big(u_\eta(x + x_r^i), A\big) \geq F_1(u^i, \mathbb{R}^k).$$

Hence, $\liminf_r F_1(u_\eta, Q_\eta) \geq \sum_{[i]} F_1(u^i, \mathbb{R}^k)$. Suppose for a moment that we know that $\deg(u^i, \infty) = d^{[i]}$. Then, using Definition 4.8, Propositions 4.9 and 4.10, and (5.21), we get

$$\sum_{[i]} F_1(u^i, \mathbb{R}^k) \geq \sum_{[i]} \overline{\tau}_{d^{[i]}} = \sum_{[i]} \tau_{d^{[i]}} \geq \tau_{\sum_{[i]} d^{[i]}} = \tau_d,$$

and Proposition 5.2 is proved.

In order to compute $\deg(u^i,\infty)$, we first observe that $\mathcal{G}^\rho_r - x^i_r$ is a grid independent of r that we may denote by $\mathcal{G}^\rho$ (indeed, it has fixed step ρ and it is centered at the origin). In view of (5.19), we may find a fixed cube $Q^{[i]}$, *independent of* r, centered at the origin and delimited by $\mathcal{G}^\rho$, with the following crucial property: for every r, $Q^{[i]}$ contains $(Q^j_r - x^i_r)$ if and only if $j \in [i]$. Then, we take a sequence of cubes $Q^{[i]}_n$, whose sides diverge as $n \to +\infty$, all centered at the origin, delimited by $\mathcal{G}^\rho$, and containing $Q^{[i]}$. For fixed n and r large enough, we have $Q^{[i]}_n \subseteq E^{[i]}_r - x^i_r$; hence $Q^{[i]}_n$ contains $(Q^j_r - x^i_r)$ if and only if $j \in [i]$, which implies $\deg\big(u_\eta(x+x^i_r)/|u_\eta(x+x^i_r)|,\, \partial Q^{[i]}_n\big) = d^{[i]}$. Passing to the limit as $r \to +\infty$ gives $\deg(u^i, \partial Q^{[i]}_n) = d^{[i]}$. Finally, letting $n \to +\infty$ and recalling Remark 4.7 (ii), we obtain $\deg(u^i,\infty) = d^{[i]}$ as required. ∎

5.2. Case $n > 0$. In order to obtain the proof of the Γ-lim inf inequality in Theorem 1.1 (i) when $n > 0$, we may exploit Theorem 5.1 by using a slicing technique. For this purpose, we need to recall some basic facts about the slices of currents by projections: they are collected in section 5.3 below. For more details on these topics and their proofs, we refer the reader to [13, sections 4.3.1 and 4.3.2] or [16, section 2.2.5].

5.3. Slicing currents by projections. Fix an arbitrary orthonormal basis $\{e_1,\dots,e_n,e_{n+1},\dots,e_{n+k}\}$ for $\mathbb{R}^{n+k}$. We set $\mathbb{R}^n_y := \mathrm{span}\{e_1,\dots,e_n\}$ and $\mathbb{R}^k_z := \mathrm{span}\{e_{n+1},\dots,e_{n+k}\}$. Given any point $x \in \mathbb{R}^{n+k}$, we identify it with a pair $(y,z) \in \mathbb{R}^n_y \times \mathbb{R}^k_z$, by writing $y_i = x\cdot e_i$ for $1 \le i \le n$ and $z_i = x \cdot e_i$ for $n+1 \le i \le n+k$. We let $dy = dy_1 \wedge \dots dy_n$ be the standard volume form on $\mathbb{R}^n_y$, and we denote by $P^\sharp(dy)$ its pull back by the orthogonal projection $P : \mathbb{R}^{n+k} \to \mathbb{R}^n_y$.

If T is a normal m-current on $\mathbb{R}^{n+k}$, with $m \ge n$, we denote by $T \llcorner P^\sharp(dy)$ the $(m-n)$-current defined by the identity $T \llcorner P^\sharp(dy)(\omega) := T\big(P^\sharp(dy) \wedge \omega\big)$ for all $\omega \in \mathcal{D}^{m-n}(\mathbb{R}^{n+k})$. Then, for $y \in \mathbb{R}^n_y$, the *slices* of T by P in y are $(m-n)$-currents $\{\langle T,P,y\rangle\}_y$ characterized by the following integral representation identity:

$$T \llcorner P^\sharp(dy)(\omega) = \int_{\mathbb{R}^n_y} \langle T,P,y\rangle(\omega)\, d\mathcal{H}^n(y) \qquad \forall \omega \in \mathcal{D}^{m-n}(\mathbb{R}^{n+k}). \tag{5.22}$$

In case T is rectifiable with integer multiplicity, the same holds for $\langle T,P,y\rangle$ for $\mathcal{H}^n$-a.e. $y \in \mathbb{R}^n_y$. Moreover, the support, the boundary, and the mass of $\langle T,P,y\rangle$ satisfy, respectively,

$$\mathrm{spt}\langle T,P,y\rangle \subset P^{-1}(y) \cap \mathrm{spt}\mathrm{T}, \tag{5.23}$$

$$\partial\langle T,P,y\rangle = (-1)^n \langle \partial T,P,y\rangle \qquad (\text{if } m > n), \tag{5.24}$$

$$\int_{\mathbb{R}^n_y} \|\langle T,P,y\rangle\|\, d\mathcal{H}^n(y) \le \|T\|. \tag{5.25}$$

Finally let us stress that, in the case $m = n$ with which we shall be mainly concerned, $T \llcorner P^\sharp(dy)$ is just a signed measure. In particular, the slices $\{\langle T,P,y\rangle\}_y$ may be seen through (5.22) as the disintegration of the measure $T \llcorner P^\sharp(dy)$ with respect to $\mathcal{H}^n \llcorner \mathbb{R}^n_y$ (see, e.g., [16, section 1.1.5]).

We can now proceed to establish some features about the behavior under the slicing of the Jacobians and of the flat norm. In the next two lemmas, we keep the notation adopted above, decomposing $\mathbb{R}^{n+k}$ as $\mathbb{R}^n_y \times \mathbb{R}^k_z$, and indicating by P the projector from $\mathbb{R}^{n+k}$ onto $\mathbb{R}^n_y$.

LEMMA 5.4. *Let a cube $Q = Q' \times Q'' \subset \mathbb{R}^n_y \times \mathbb{R}^k_z$ and a smooth function $u : Q \to \mathbb{R}^k$ be given. For $y \in Q'$, define $u^y : Q'' \to \mathbb{R}^k$ by $u^y(z) := u(y,z)$. Then, for every test function $\omega \in \mathcal{D}(Q)$, there holds*

$$\langle \star Ju, P, y\rangle(\omega) = \star Ju^y(\omega(y,\cdot)) \qquad \text{for } \mathcal{H}^n\text{-a.e. } y \in Q'. \tag{5.26}$$

Proof. Recall that, by definition, $\star Ju^y$ is the absolutely continuous measure on Q'' with density $\det \nabla_z u(y,z)$, where ∇_z denotes partial differentiation with respect to the components $z_{n+1}, \dots, z_{n+k}$ of x. On the other hand, we have $\star Ju = \sum_\alpha \det \nabla_\alpha u(x) \star dx_\alpha$, where the sum is extended over all k-multi-indices $\alpha = (i_1, \dots, i_k)$ with $1 \le i_1 < \cdots < i_k \le n+k$, $\nabla_\alpha u$ is the $(k \times k)$-matrix with columns $\nabla_{i_1} u, \dots, \nabla_{i_k} u$, and dx_α is the k-covector $dx_{i_1} \wedge \cdots \wedge dx_{i_k}$. In particular, we observe that $\star dx_\alpha \cdot P^\sharp(dy)$ equals 1 if $\alpha = (n+1, \dots, n+k)$, and 0 otherwise. Therefore, for every test function $\omega \in \mathcal{D}(Q)$, we have

$$\begin{aligned}(\star Ju \llcorner P^\sharp(dy))(\omega) = \star Ju(\omega P^\sharp(dy)) &= \sum_\alpha \int_Q \omega(x) \det \nabla_\alpha u(x) \star dx_\alpha \cdot P^\sharp(dy) \\ &= \int_Q \omega(y,z) \det \nabla_z u(y,z) = \int_{Q'} \star Ju^y(\omega(y,\cdot))\, d\mathcal{H}^n(y),\end{aligned}$$

where in the last equality we have applied Fubini's theorem. Hence, recalling the disintegration identity (5.22), we deduce the validity of (5.26). ∎

LEMMA 5.5. *Let $\{T_\varepsilon\}$ be a sequence of boundaries in $\mathcal{D}_n(\Omega)$, with $\mathbb{F}_\Omega(T_\varepsilon) \to 0$. Then, possibly passing to a subsequence, there holds*

$$\mathbb{F}_{P^{-1}(y)\cap\Omega}(\langle T_\varepsilon, P, y\rangle) \to 0 \qquad \text{for } \mathcal{H}^n\text{-a.e. } y.$$

Proof. By the definition of flat norm, there exists a sequence of currents $\{S_\varepsilon\} \subset \mathcal{D}_{n+1}(\Omega)$ such that $T_\varepsilon = \partial S_\varepsilon$, and $\|S_\varepsilon\| \le \mathbb{F}_\Omega(T_\varepsilon) + \varepsilon$. By (5.24), it holds that $\partial\langle S_\varepsilon, P, y\rangle = (-1)^n \langle T_\varepsilon, P, y\rangle$ for $\mathcal{H}^n$-a.e. y. Recalling (5.23) and applying again the definition of a flat norm, we infer that

$$\mathbb{F}_{P^{-1}(y)\cap\Omega}(\langle T_\varepsilon, P, y\rangle) \le \|\langle S_\varepsilon, P, y\rangle\| \qquad \text{for } \mathcal{H}^n\text{-a.e. } y.$$

Since $\|S_\varepsilon\| \to 0$, by (5.25) we can pass to a subsequence (not relabeled) such that $\|\langle S_\varepsilon, P, y\rangle\| \to 0$ for $\mathcal{H}^n$-a.e. y, and the statement follows. ∎

Given an n-dimensional integral current $M = (M, \xi, \sigma)$, let $\widehat{M}$ be the current obtained by just changing the multiplicity of M from $\sigma(x)$ into $\tau_{\sigma(x)}$, namely,

$$M = (M, \xi, \sigma) \;\Rightarrow\; \widehat{M} := (M, \xi, \tau_\sigma). \tag{5.27}$$

We are ready to give the following proof.

Proof of the Γ-lim inf *inequality in Theorem* 1.1 (i). Let $\{u_\varepsilon\}$ and $M = (M, \xi, \sigma)$ be as in the assumptions; it is not restrictive to suppose in addition that u_ε are smooth. With the notation introduced at the beginning of this section, decompose $\mathbb{R}^{n+k}$ as $\mathbb{R}^n_y \times \mathbb{R}^k_z$, and let P denote the orthogonal projection from $\mathbb{R}^{n+k}$ onto $\mathbb{R}^n_y$. Then, for the n-current $\widehat{M}$ defined according to (5.27), consider the signed measure $\widehat{M} \llcorner P^\sharp(dy)$. Also, let Q be a closed cube contained into Ω. We claim that the following inequality holds true and implies (1.3):

$$\liminf_\varepsilon F_\varepsilon(u_\varepsilon, Q) \ge |\widehat{M} \llcorner P^\sharp(dy)|(Q). \tag{5.28}$$

Let us first show how (1.3) follows from (5.28) and then turn back to the proof of (5.28). Assume that (5.28) is satisfied. Then, it continues to hold if the cube Q is replaced by an arbitrary open set $A \subset\subset \Omega$. Indeed, by the Vitali covering theorem, we can find a collection $\{Q_i\}$ of pairwise disjoint closed cubes which cover A up to a $|\widehat{M} \llcorner P^\sharp(dy)|$-negligible set. We then have

$$\liminf_\varepsilon F_\varepsilon(u_\varepsilon, A) \geq \sum_i \liminf_\varepsilon F_\varepsilon(u_\varepsilon, Q_i) \geq \sum_i |\widehat{M} \llcorner P^\sharp(dy)|(Q_i) = |\widehat{M} \llcorner P^\sharp(dy)|(A).$$

Recalling that the basis $\{e_1, \dots, e_{n+k}\}$ chosen at the beginning of this section was arbitrary, we infer that, for any simple unit $w \in \Lambda^n(\mathbb{R}^{n+k})$, there holds

$$\liminf_\varepsilon F_\varepsilon(u_\varepsilon, A) \geq \int_{A \cap M} \tau_{\sigma(x)} w \cdot \xi(x)\, d\mathcal{H}^n(x). \tag{5.29}$$

Applying (5.29) to finitely many pairwise disjoint open sets $A_i \subset\subset \Omega$ and simple unit n-covectors w_i, and then summing over i, we get

$$\liminf_\varepsilon F_\varepsilon(u_\varepsilon) \geq \sum_i \int_{A_i \cap M} \tau_{\sigma(x)} w_i \cdot \xi(x)\, d\mathcal{H}^n(x). \tag{5.30}$$

Finally, the supremum of the right-hand side of (5.30) over all possible choices of A_i and w_i furnishes the total mass of the variation measure $|\widehat{M}|$, that is, $\int_M \tau_{\sigma(x)}\, d\mathcal{H}^n(x)$. We have thus obtained (1.3). It remains to prove (5.28).

We write Q as $Q' \times Q'' \subset \mathbb{R}^n_y \times \mathbb{R}^k_z$. For $y \in Q'$, similarly as in Lemma 5.4, we define u_ε^y on Q'' by $u_\varepsilon^y(z) := u_\varepsilon(y, z)$. We pass to a subsequence (not relabeled) such that $\liminf_\varepsilon F_\varepsilon(u_\varepsilon, Q) = \lim_\varepsilon F_\varepsilon(u_\varepsilon, Q)$ and, using Lemma 5.5,

$$\mathbb{F}_{P^{-1}(y) \cap \Omega}(\langle \star J u_\varepsilon, P, y\rangle - \langle M, P, y\rangle) \to 0 \qquad \text{for } \mathcal{H}^n\text{-a.e. } y. \tag{5.31}$$

For the sake of clarity, in the remainder of the proof we denote by G_ε the functionals F_ε when $n = 0$. We have

$$\lim_\varepsilon F_\varepsilon(u_\varepsilon, Q) \geq \liminf_\varepsilon \int_{Q'} G_\varepsilon(u_\varepsilon^y, Q'')\, d\mathcal{H}^n \geq \int_{Q'} \liminf_\varepsilon G_\varepsilon(u_\varepsilon^y, Q'')\, d\mathcal{H}^n, \tag{5.32}$$

where the first inequality holds since $|\nabla u_\varepsilon^y(z)| \leq |\nabla u_\varepsilon(y, z)|$, and the second one is just the Fatou lemma. Now, we want to apply Theorem 5.1 in order to estimate from below the last integral in (5.32). For $\mathcal{H}^n$-a.e. $y \in Q'$, we may identify $\star J u_\varepsilon^y$ with $\langle \star J u_\varepsilon, P, y\rangle$ according to (5.26) in Lemma 5.4. Hence, writing the slices $\langle M, P, y\rangle$ under the form $\sum_i \sigma(y, z_i)\delta(y, z_i)$, we deduce from (5.31) that $\mathbb{F}_{Q''}\big(\star J u_\varepsilon^y - \sum_i \sigma(y, z_i)\delta(y, z_i)\big) \to 0$ (for a finite collection of points $z_i = z_i(y) \in Q''$). Then Theorem 5.1 yields

$$\liminf_\varepsilon G_\varepsilon(u_\varepsilon^y, Q'') \geq \sum_i \tau_{\sigma(y, z_i)} = \langle \widehat{M}, P, y\rangle(Q'') \qquad \text{for } \mathcal{H}^n\text{-a.e. } y \in Q'. \tag{5.33}$$

Integrating (5.33) over Q', it follows from (5.32) that

$$\lim_\varepsilon F_\varepsilon(u_\varepsilon, Q) \geq \int_{Q'} \langle \widehat{M}, P, y\rangle(Q'')\, d\mathcal{H}^n(y) = |\widehat{M} \llcorner P^\sharp(dy)|(Q). \qquad \square$$

6. Γ-lim sup inequality. The proof of the Γ-lim sup inequality in Theorem 1.1 will be achieved in two steps. The first one consists in reducing our attention to proving the statement when the limit boundary M is polyhedral. This goal is obtained by means of the ad hoc density result stated in Proposition 6.1 below (which is an adaptation of Proposition 8.6 in [3]), combined with a standard diagonal argument. For convenience, given an n-dimensional integral current $M = (M, \xi, \sigma)$, and an $\mathcal{H}^n$-measurable set A in $\mathbb{R}^{n+k}$, we consider the modified current $\widehat{M}$ defined by (5.27), and we introduce the energy $\mathcal{E}(M, A)$ given by

$$\mathcal{E}(M, A) := |\widehat{M}|(A) = \int_{M\cap A} \tau_{\sigma(x)}\, d\mathcal{H}^n(x).$$

PROPOSITION 6.1. *For every n-dimensional integral boundary M in Ω, there exists a sequence $\{M_i\}$ of n-dimensional polyhedral boundaries in Ω such that $|M_i|(\partial\Omega) = 0$, $\mathbb{F}_\Omega(M_i - M) \to 0$, and $\mathcal{E}(M_i, \Omega) \to \mathcal{E}(M, \Omega)$.*

Proof. First, we remark that M agrees in Ω with the boundary of an integral $(n+1)$-current N, with compact support K, such that $|\partial N|(\partial\Omega) = 0$; see Proposition 8.6 (i) in [3]. This implies, in particular, that $\mathcal{E}(M, \partial\Omega) = 0$ (indeed, by Proposition 4.10, $\mathcal{E}(M, A)$ is controlled by the total variation of M over A; namely, there exists a positive constant C, independent of M and A, such that $\mathcal{E}(M, A) \le C\,|M|(A)$). By the strong approximation theorem (see, e.g., [13, sections 4.2.20 and 4.2.21]), for every $\varepsilon > 0$, there exist an integral polyhedral current P and a diffeomorphism $f : \mathbb{R}^{n+k} \to \mathbb{R}^{n+k}$ such that

$$|f(x) - x| < \varepsilon \quad \forall x \in \mathbb{R}^{n+k}, \quad \mathrm{Lip}(f) < 1 + \varepsilon, \quad \mathrm{Lip}(f^{-1}) < 1 + \varepsilon, \tag{6.1}$$

$$\|P - f_\sharp N\| + \|\partial P - \partial f_\sharp N\| < \varepsilon, \tag{6.2}$$

$$\mathcal{F}_K(P - N) < o(1), \tag{6.3}$$

where $\mathcal{F}_K$ is the usual flat norm for integral currents over the compact set K [32, section 31], and $o(1)$ denotes an infinitesimal as $\varepsilon \to 0$. We deduce that

$$\mathcal{E}(\partial P - f_\sharp M, \Omega) \le C\,|\partial P - f_\sharp M|(\Omega) < \varepsilon, \tag{6.4}$$

where the first inequality follows from the definition of $\mathcal{E}$ using Proposition 4.10, while the second one is a consequence of (6.2). On the other hand, if Ω_ε denotes the ε-neighborhood of Ω, we have

$$\mathcal{E}(f_\sharp M, \Omega) \le (\mathrm{Lip} f)^n \mathcal{E}(M, f^{-1}(\Omega)) \le (1+\varepsilon)^n\, \mathcal{E}(M, \Omega_\varepsilon) = \mathcal{E}(M, \Omega) + o(1), \tag{6.5}$$

where the first inequality follows from (2.2) and the identity $\widehat{f_\sharp M} = f_\sharp \widehat{M}$; in the second inequality we have used (6.1), while the third equality is consequence of the null energy condition on the boundary $\mathcal{E}(M, \partial\Omega) = 0$. In a similar way, one has

$$\mathcal{E}(M, \Omega) \le \mathcal{E}(f_\sharp M, \Omega) + o(1). \tag{6.6}$$

By (6.3), (6.4), (6.5), and (6.6), we may find a sequence $\{P_i\}$ of integral polyhedral currents such that $\mathbb{F}_\Omega(\partial P_i - M) \to 0$ and $\mathcal{E}(\partial P_i, \Omega) \to \mathcal{E}(M, \Omega)$. The proof is then achieved by setting $M_i = \partial P_i$, after possibly translating the currents P_i so that they also satisfy the additional condition $|\partial P_i|(\partial\Omega) = 0$ (which can always be done thanks to Fubini's theorem). ∎

Let us turn to the second and main step in the proof of the Γ-lim sup inequality. We assume from now on that M agrees in Ω with the boundary of a polyhedral

$(n+1)$-current N, with $|\partial N|(\partial\Omega) = 0$. Our goal is to exhibit a recovery sequence $\{u_\varepsilon\}$ satisfying (1.4). We denote by S the union of all $(n-1)$-simplexes of N. Moreover, given an n-face F of ∂N, we identify the n-plane spanned by F with $\mathbb{R}^n$, and we write $x \in \mathbb{R}^{n+k}$ as $x = (y, z) \in \mathbb{R}^n \times \mathbb{R}^k$; then, for $\delta, \gamma > 0$, we set

$$U(F, \delta, \gamma) := \big\{x = (y, z) \ : \ y \in F \text{ and } |z| < \min\{\delta, \gamma \operatorname{dist}(y, \partial F)\}\big\}.$$

A crucial ingredient in the construction of the sequence $\{u_\varepsilon\}$ is the following existence result for S^{k-1}-valued functions with prescribed singularities proved in [2]: one can construct a map $u \in W^{1,k-1}_{\mathrm{loc}}(\mathbb{R}^{n+k}; S^{k-1}) \cap \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}^{n+k} \setminus (M \cup S); \mathbb{R}^k)$ such that $\star Ju = \alpha_k M$, and

$$|\nabla u(x)| \le \frac{C}{\operatorname{dist}(x, M \cup S)}. \tag{6.7}$$

Moreover, one can ask that, for every n-face F of ∂N, and for some $\delta, \gamma > 0$,

$$u(x) = \phi_\sigma\left(\frac{z}{|z|}\right) \qquad \forall x = (y, z) \in U(F, \delta, \gamma), \tag{6.8}$$

where σ is the multiplicity of ∂N on F, and $\phi_\sigma : S^{k-1} \to S^{k-1}$ is a prescribed map of degree σ. Condition (6.8) leads in a natural way to Definition 4.1 of the constant τ_σ. So, we fix $h \in \mathbb{N}$ and, according to (4.1), we consider for every $\sigma \in \mathbb{Z}$ a function $\psi_\sigma \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^k; \mathbb{R}^k)$ with

$$\operatorname{tr}(\psi_\sigma, \infty) = \phi_\sigma \qquad \text{and} \qquad F_1(\psi_\sigma, \mathbb{R}^k) < \tau_\sigma + \frac{1}{h}. \tag{6.9}$$

The idea is to glue together the trace conditions in (6.8) and (6.9), and to this aim we need a careful choice of some parameters. Precisely, we take $\delta = \delta(\varepsilon)$ and a fixed $\gamma > 0$ such that (6.8) holds and $\varepsilon \ll \delta(\varepsilon) \ll \varepsilon^{k/(k+1)}$. Moreover, we put $r = r(\varepsilon) := \delta(\varepsilon)\sqrt{1 + \gamma^{-2}}$. Then we set $S_\varepsilon := \big\{x \in \mathbb{R}^{n+k} \ : \ \operatorname{dist}(x, S) < r(\varepsilon)\big\}$, and $U_\varepsilon := \bigcup_i U(F_i, \delta, \gamma)$, where the union is extended over all the n-faces F_i of ∂N (with the convention that, whenever $i \neq j$, $F_i \cap F_j$ is either empty or a common face between F_i and F_j).

We are now ready to construct a recovery sequence $\{u_\varepsilon\}$, whose energy concentrates in a transition zone near M, precisely in $\Omega \cap (U_\varepsilon \setminus S_\varepsilon)$. This kind of qualitative behavior is shared with the scalar case $p > k = 1$, and it is in contrast with the Ginzburg–Landau case $p = k > 1$, when the set of energy concentration around M has a "circular crown" shaped section (compare the proof of Proposition 6.2 below, respectively, with section 3.5 in [1] and with Lemma 4.2 in [3]).

PROPOSITION 6.2. *Assume that M is the restriction to Ω of a polyhedral boundary ∂N, with $|\partial N|(\partial\Omega) = 0$, and adopt the notation fixed above. Set*

$$v_\varepsilon(x) = \begin{cases} u(x) & \text{if } x \notin U_\varepsilon, \\ \psi_{\sigma_i}\Big(\frac{\delta z}{\varepsilon g_i(y)}\Big) & \text{if } x = (y, z) \in U(F_i, \delta, \gamma), \end{cases}$$

where σ_i is the multiplicity of ∂N on F_i, and $g_i(y) := \min\{\delta, \gamma \operatorname{dist}(y, \partial F_i)\}$.

Then, if $u_\varepsilon(x) := \min\big\{\operatorname{dist}(x, S)/r, 1\big\}\, v_\varepsilon(x)$, we have $\mathbb{F}_\Omega(\star Ju_\varepsilon - \alpha_k M) \to 0$, and the Γ-lim sup inequality (1.4) is satisfied.

Proof. First we observe that, by (6.8), (6.9), and the assumption $\varepsilon \ll \delta(\varepsilon)$, for ε small enough both $u(x)$ and $\psi_{\sigma_i}\big(\delta z/\varepsilon g_i(y)\big)$ equal $\phi_{\sigma_i}(z/|z|)$ for $x = (y, z) \in$

$\partial U(F_i,\delta,\gamma)$. Notice also that u_ε are equibounded in $L^\infty(\Omega;\mathbb{R}^k)$; therefore, by the choice of u and inequality (2.6), we have $\mathbb{F}_\Omega(\star J u_\varepsilon - \alpha_k M) \to 0$ provided $\int_\Omega |\nabla u_\varepsilon - \nabla u|^{k-1} \to 0$. Let us introduce the following subdomains of Ω:

$$V_\varepsilon^1 = \Omega \cap (U_\varepsilon \setminus S_\varepsilon), \qquad V_\varepsilon^2 = \Omega \setminus (S_\varepsilon \cup U_\varepsilon),$$
$$V_\varepsilon^3 = \Omega \cap (S_\varepsilon \setminus U_\varepsilon), \qquad V_\varepsilon^4 = \Omega \cap S_\varepsilon \cap U_\varepsilon.$$

For $i = 1,\dots,4$, we are going to estimate separately on each V_ε^i the energy amount of u_ε and their $W^{1,k-1}$-convergence.

- On V_ε^1, $u_\varepsilon(x)$ equals $\psi_{\sigma_i}(z/\varepsilon)$ for all $x = (y,z) \in U(F_i,\delta,\gamma)$. Moreover, the projection of $V_\varepsilon^1 \cap U(F_i,\delta,\gamma)$ on the affine plane spanned by F_i is contained into $F_i \cap \Omega_\delta$, Ω_δ being the closed δ-neighborhood of Ω. Below, for the sake of clarity, we indicate by G_ε (respectively, G_1) the functionals F_ε (respectively, F_1) when $n = 0$.

Energy estimate. Using (5.1) and (6.9), we get

$$\begin{aligned} F_\varepsilon(u_\varepsilon, V_\varepsilon^1) &= \sum_i F_\varepsilon\left(\psi_{\sigma_i}\left(\frac{z}{\varepsilon}\right),\, \Omega \cap (U(F_i,\delta,\gamma) \setminus S_\varepsilon)\right) \\ &\le \sum_i \mathcal{H}^n(F_i \cap \Omega_\delta)\, G_\varepsilon\left(\psi_{\sigma_i}\left(\frac{z}{\varepsilon}\right),\, |z| < \delta\right) \\ &= \sum_i \mathcal{H}^n(F_i \cap \Omega_\delta)\, G_1\left(\psi_{\sigma_i}(z),\, |z| < \frac{\delta}{\varepsilon}\right) \\ &\le \sum_i \mathcal{H}^n(F_i \cap \Omega_\delta)(\tau_{\sigma_i} + 1/h). \end{aligned}$$

$W^{1,k-1}$-convergence. By the Hölder inequality and the assumption $\delta(\varepsilon) \ll \varepsilon^{k/(k+1)}$, we have

$$\begin{aligned} \int_{V_\varepsilon^1} |\nabla u_\varepsilon|^{k-1} &\le \sum_i \mathcal{H}^n(F_i \cap \Omega_\delta) \int_{|z|<\delta} \varepsilon^{-k+1} \left|\nabla \psi_{\sigma_i}\left(\frac{z}{\varepsilon}\right)\right|^{k-1} d\mathcal{H}^k(z) \\ &= \sum_i \mathcal{H}^n(F_i \cap \Omega_\delta) \int_{|z|<\delta/\varepsilon} \varepsilon |\nabla \psi_{\sigma_i}(z)|^{k-1}\, d\mathcal{H}^k(z) \\ &\le C\varepsilon \left(\frac{\delta}{\varepsilon}\right)^{k\left(1-\frac{k-1}{p}\right)} \ll C\varepsilon^{1-\frac{k(p-k+1)}{(k+1)p}} \to 0. \end{aligned}$$

- On V_ε^2, $u_\varepsilon(x) = u(x)$. We remark that there exists some positive constant C independent of ε such that outside U_ε (so in particular for $x \in V_\varepsilon^2$), there holds

$$\operatorname{dist}(x, M \cup S) \ge C \operatorname{dist}(x,S). \tag{6.10}$$

Energy estimate. Recalling that u takes values into S^{k-1}, and using (6.7), (6.10), the coarea formula (2.8) (with $D := \sup\{\operatorname{dist}(x,S) : x \in \Omega\}$), and the assumption $r(\varepsilon) \sim \delta(\varepsilon) \gg \varepsilon$, we get

$$\begin{aligned} F_\varepsilon(u_\varepsilon, V_\varepsilon^2) &= \varepsilon^{p-k} \int_{V_\varepsilon^2} |\nabla u|^p \le C\varepsilon^{p-k} \int_{\Omega \setminus S_\varepsilon} \frac{1}{\operatorname{dist}(x,S)^p} \\ &\le C\varepsilon^{p-k} \int_r^D \frac{1}{t^p} \mathcal{H}^{n+k-1}\big(\{x \in \mathbb{R}^{n+k} : \operatorname{dist}(x,S) = t\}\big)\, dt \\ &\le C\varepsilon^{p-k}\big(D^{-p+k+1} - r^{-p+k+1}\big) \to 0. \end{aligned}$$

$W^{1,k-1}$-convergence. We trivially have $\int_{V_\varepsilon^2} |\nabla u_\varepsilon - \nabla u|^{k-1} \equiv 0$.

- On V_ε^3, $u_\varepsilon(x) = r^{-1}\,\mathrm{dist}(x,S)u(x)$. Then, since (6.10) holds on V_ε^3, we have

$$|\nabla u_\varepsilon(x)| \le Cr^{-1} \qquad \forall x \in V_\varepsilon^3. \tag{6.11}$$

Energy estimate. We have $F_\varepsilon(u_\varepsilon, V_\varepsilon^3) = \mathrm{I} + \mathrm{II}$, where, using inequality (6.11) and the assumption $r(\varepsilon) \sim \delta(\varepsilon)$, with $\varepsilon \ll \delta(\varepsilon) \ll \varepsilon^{k/k+1}$,

$$\mathrm{I} = \varepsilon^{p-k} \int_{V_\varepsilon^3} |\nabla u_\varepsilon|^p \le C\varepsilon^{p-k} r^{-p+k+1} \to 0,$$
$$\mathrm{II} = \varepsilon^{-k} \int_{V_\varepsilon^3} W\big(r^{-1}\mathrm{dist}(x,S)u\big) \le C\varepsilon^{-k} r^{k+1} \to 0.$$

$W^{1,k-1}$-convergence. Again by (6.11), we obtain $\int_{V_\varepsilon^3} |\nabla u_\varepsilon|^{k-1} \le Cr^2 \to 0$.

- On $V_\varepsilon^4, u_\varepsilon(x) = r^{-1}\mathrm{dist}(x,S)\psi_{\sigma_i}\big(\delta z/\varepsilon g_i(y)\big)$ for all $x = (y,z) \in U(F_i,\delta,\gamma)$.

For $x = (y,z) \in V_\varepsilon^4$, one can check that $\mathrm{dist}(x,S) \le C\mathrm{dist}(y,S)$, and

$$\left|\nabla\left(\psi_{\sigma_i}\left(\frac{\delta z}{\varepsilon g_i(y)}\right)\right)\right| \le \frac{C\delta}{\varepsilon g_i(y)}\left|\nabla\psi_{\sigma_i}\left(\frac{\delta z}{\varepsilon g_i(y)}\right)\right| \qquad \text{if } x = (y,z) \in U(F_i,\delta,\gamma).$$

It follows that, on $V_\varepsilon^4 \cap U(F_i,\delta,\gamma)$, $|\nabla u_\varepsilon|$ satisfy the inequality

$$|\nabla u_\varepsilon(x)| \le C\left\{\frac{1}{r} + \frac{\mathrm{dist}(y,S)}{r}\frac{\delta}{\varepsilon g_i(y)}\left|\nabla\psi_{\sigma_i}\left(\frac{\delta z}{\varepsilon g_i(y)}\right)\right|\right\}. \tag{6.12}$$

Energy estimate. Using (6.12), we have $F_\varepsilon(u_\varepsilon, V_\varepsilon^4) \le C(\mathrm{I}+\mathrm{II}) + \mathrm{III}$, with

$$\mathrm{I} = \int_{V_\varepsilon^4} \varepsilon^{p-k} r^{-p} \le C\varepsilon^{p-k} r^{-p+k+1} \to 0,$$
$$\begin{aligned}\mathrm{II} &= \sum_i \int_{U(F_i,\delta,\gamma)\cap S_\varepsilon\cap\Omega} \varepsilon^{p-k} r^{-p}\mathrm{dist}(y,S)^p \left(\frac{\delta}{\varepsilon g_i(y)}\right)^p \left|\nabla\psi_{\sigma_i}\left(\frac{\delta z}{\varepsilon g_i(y)}\right)\right|^p \\ &\le \sum_i \left\{\int_{\mathbb{R}^k} |\nabla\psi_{\sigma_i}|^p\right\} \int_{F_i\cap S_\varepsilon\cap\Omega_\delta} \varepsilon^{p-k} r^{-p}\,\mathrm{dist}(y,S)^p \left(\frac{\delta}{\varepsilon g_i(y)}\right)^{p-k} d\mathcal{H}^n(y) \\ &\le C\sum_i \int_{F_i\cap S_\varepsilon\cap\Omega_\delta} \frac{\mathrm{dist}(y,S)^p\delta^{-k}}{g_i(y)^{p-k}}\, d\mathcal{H}^n(y) \le C\sum_i \mathcal{H}^n(F_i\cap S_\varepsilon\cap\Omega_\delta) \to 0,\end{aligned}$$
$$\mathrm{III} = \int_{V_\varepsilon^4} \varepsilon^{-k} W(u_\varepsilon) \le C\varepsilon^{-k} r^{k+1} \to 0.$$

$W^{1,k-1}$-convergence. Using again (6.12), we have $\int_{V_\varepsilon^4} |\nabla u_\varepsilon|^{k-1} \le C(\mathrm{I}+\mathrm{II})$, with

$$\mathrm{I} = \int_{V_\varepsilon^4} r^{-k+1} \le Cr^2 \to 0,$$
$$\begin{aligned}\mathrm{II} &= \sum_i \int_{U(F_i,\delta,\gamma)\cap S_\varepsilon\cap\Omega} r^{-k+1}\,\mathrm{dist}(y,S)^{k-1} \left(\frac{\delta}{\varepsilon g_i(y)}\right)^{k-1} \left|\nabla\psi_{\sigma_i}\left(\frac{\delta z}{\varepsilon g_i(y)}\right)\right|^{k-1} \\ &\le C\sum_i \left\{\int_{|z|<\delta/\varepsilon} |\nabla\psi_{\sigma_i}|^{k-1}\right\} \int_{F_i\cap S_\varepsilon\cap\Omega_\delta} r^{-k+1}\,\mathrm{dist}(y,S)^{k-1} \frac{\varepsilon g_i(y)}{\delta}\, d\mathcal{H}^n(y) \\ &\le C\varepsilon\Big(\frac{\delta}{\varepsilon}\Big)^{k(1-\frac{k-1}{p})} \sum_i \mathcal{H}^n(F_i\cap S_\varepsilon\cap\Omega_\delta) \ll C\varepsilon^{1-\frac{k(p-k+1)}{(k+1)p}} \to 0.\end{aligned}$$

By the estimates obtained on V_ε^i for $i = 1, \dots, 4$, one has $\int_\Omega |\nabla u_\varepsilon - \nabla u|^{k-1} \to 0$, and

$$\begin{aligned}\limsup_\varepsilon F_\varepsilon(u_\varepsilon) &\le \limsup_\varepsilon \sum_i \mathcal{H}^n(F_i \cap \Omega_\delta)(\tau_{\sigma_i} + 1/h) \\ &= \limsup_\varepsilon \left\{ \int_{\partial N \cap \Omega_\delta} \tau_{\sigma(x)}\, d\mathcal{H}^n(x) + (1/h)\mathcal{H}^n(\partial N \cap \Omega_\delta) \right\} \\ &= \int_{M \cap \overline{\Omega}} \tau_{\sigma(x)}\, d\mathcal{H}^n(x) + C/h,\end{aligned}$$

The statement follows from the assumption $|M|(\partial\Omega) = 0$, and letting $h \to +\infty$. ∎

Acknowledgments. The origins of this research are traceable to the work by Alberti, Baldo, and Orlandi; see Remark (iv) after Corollary 1.2 in [3]. In particular, we warmly thank Giovanni Alberti for the stimulating discussions shared on these topics.

REFERENCES

[1] G. Alberti, *Variational models for phase transitions. An approach via Γ-convergence*, in Differential Equations and Calculus of Variations. Topics on Geometrical Evolutions Problems and Degree Theory (Pisa, 1996), G. Buttazzo et al., eds., Springer-Verlag, Berlin, 2000, pp. 95–114.

[2] G. Alberti, S. Baldo, and G. Orlandi, *Functions with prescribed singularities*, J. Eur. Math. Soc. (JEMS), 5 (2003), pp. 275–311.

[3] G. Alberti, S. Baldo, and G. Orlandi, *Variational convergence for functionals of Ginzburg-Landau type*, Indiana Univ. Math. J., 54 (2005), pp. 1411–1472.

[4] F. Bethuel, J. Bourgain, H. Brezis, and G. Orlandi, *$W^{1,p}$ estimates for solutions to the Ginzburg-Landau equation with boundary data in $H^{1/2}$*, C. R. Acad. Sci. Paris, Sér. I Math., 333 (2001), pp. 1069–1076.

[5] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices*, Progr. Nonlinear Differential Equations Appl. 13, Birkhäuser Boston, Boston, 1994.

[6] A. Braides, *Γ-Convergence for Beginners*, Oxford Lecture Ser. Math. Appl. 22, Oxford University Press, New York, 2002.

[7] H. Brezis and P. Mironescu, *Sur une conjecture de E. De Giorgi relative à l'energie de Ginzburg-Landau*, C. R. Acad. Sci. Paris, Sér. I Math., 319 (1994), pp. 167–170.

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

[9] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, 1993.

[10] E. De Giorgi and T. Franzoni, *Su un tipo di convergenza variazionale*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend., 58 (1958), pp. 258–267.

[11] N. Desenzani, *Variational Convergence of Ginzburg-Landau Functionals with Supercritical Growth*, Ph.D. thesis, Università di Milano, Milan, Italy, 2004.

[12] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

[13] H. Federer, *Geometric Measure Theory*, Grundlehren Math. Wiss. 153, Springer-Verlag, New York, 1969.

[14] I. Fonseca and W. Gangbo, *Degree Theory in Analysis and Its Applications*, Oxford Lecture Ser. Math. Appl. 2, Clarendon, Oxford, UK, 1995.

[15] I. Fonseca and C. Mantegazza, *Second order singular perturbation models for phase transitions*, SIAM J. Math. Anal., 31 (2000), pp. 1121–1143.

[16] M. Giaquinta, G. Modica, and J. Souček, *Cartesian Currents in the Calculus of Variations* I, Ergeb. Math. Grenzgeb. (3) 37, Springer-Verlag, Berlin, 1998.

[17] V. Ginzburg and L. Landau, *On the theory of superconductivity*, Zh. Èksper. Teoret. Fiz., 20 (1950), pp. 1064–1082.

[18] V. Ginzburg and L. Pitaevskii, *On the theory of superfluidity*, Soviet Physics JETP, 34 (1958), pp. 858–861.

[19] Z.-C. Han and Y. Y. Li, *Degenerate elliptic systems and applications to Ginzburg-Landau type equations*. I, Calc. Var. Partial Differential Equations, 4 (1996), pp. 171–202.

[20] Z.-C. Han and Y. Y. Li, *Erratum: Degenerate elliptic systems and applications to Ginzburg-Landau type equations*. I, Calc. Var. Partial Differential Equations, 4 (1996), p. 497.

[21] R. Hardt and F. Lin, *Mappings minimizing the L^p-norm of the gradient*, Comm. Pure Appl. Math., 40 (1987), pp. 555–588.

[22] M. W. Hirsch, *Differential Topology*, Grad. Texts in Math. 33, Springer-Verlag, New York, 1994.

[23] M. C. Hong, *Asymptotic behaviour for minimizers of a Ginzburg-Landau-type functional in higher dimension associated with n-harmonic maps*, Adv. Differential Equations, 1 (1996), pp. 611–634.

[24] R. L. Jerrard, *Lower bounds for generalized Ginzburg–Landau functionals*, SIAM J. Math. Anal., 30 (1999), pp. 721–746.

[25] R. L. Jerrard and H. M. Soner, *Functions of bounded higher variation*, Indiana Univ. Math. J., 51 (2002), pp. 645–677.

[26] R. L. Jerrard and H. M. Soner, *The Jacobian and the Ginzburg-Landau energy*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 151–191.

[27] Y. Lei, Z. Wu, and H. Yuan, *Radial minimizers of a Ginzburg-Landau functional*, Electron. J. Differential Equations, 30 (1999), pp. 1–21.

[28] F. Lin and T. Rivière, *Complex Ginzburg-Landau equations in high dimensions and codimension two area minimizing currents*, J. Eur. Math. Soc. (JEMS), 1 (1999), pp. 237–311.

[29] F. Lin and T. Rivière, Erratum, *Complex Ginzburg-Landau equations in high dimensions and codimension two area minimizing currents*, J. Eur. Math. Soc. (JEMS), 2 (2000), pp. 87–91.

[30] L. Modica and S. Mortola, *Un esempio di Γ-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[31] E. Sandier, *Ginzburg-Landau minimizers from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ and minimal connections*, Indiana Univ. Math. J., 50 (2001), pp. 1087–1844.

[32] L. Simon, *Lectures on Geometric Measure Theory*, Proceedings of the Centre for Mathematical Analysis 3, Australian National University, Canberra, Australia, 1983.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 414–433

THE Γ-CONVERGENCE OF A SHARP INTERFACE THIN FILM MODEL WITH NONCONVEX ELASTIC ENERGY*

PAVEL BĚLÍK† AND MITCHELL LUSKIN‡

Abstract. We give results for the Γ-limit of a scaled elastic energy of a film as the thickness $h > 0$ converges to zero. The elastic energy density models materials with multiple phases or variants and is thus nonconvex. The model includes an interfacial energy that allows sharp interfaces between the phases and variants and is proportional to the total variation of the deformation gradient.

Key words. thin film, Γ-convergence, surface energy, bounded variation, martensite

AMS subject classifications. 49J45, 65N15, 65N30, 73C50, 73G05, 73K20, 73V05

DOI. 10.1137/050622596

1. Introduction.

Thin films of martensitic crystals are the subject of increasing scientific and technological interest [6, 14, 20]. Dimensionally reduced models that replace the three-dimensional (3-D) bulk energy with a two-dimensional (2-D) thin film energy can make the design of applications more tractable and the computation of the deformation more efficient. New challenges arise in the derivation of thin film energies for martensitic crystals since the presence of multiple phases and variants requires that the elastic energy density be nonconvex [4, 23, 27] and since an interfacial energy that allows sharp interfaces is often useful for accurate modeling [10, 11, 12]. Related work on the general problem of rigorously deriving dimensionally reduced energy functionals has been given in [1, 3, 6, 15, 17, 25, 28].

We present results for the Γ-limit [7, 24] of the scaled elastic energy of a thin film with deformation $\tilde{u} : \Omega_h \to \mathbb{R}^3$ defined on a reference domain of thickness $h > 0$ given by $\Omega_h = S \times (-h/2,\, h/2)$ for $S \subset \mathbb{R}^2$ and subject to boundary conditions

$$\tilde{u}(x_1, x_2, x_3) = y_0(x_1, x_2) + b_0(x_1, x_2)x_3 \quad \text{for } (x_1, x_2, x_3) \in \gamma \times (-h/2,\, h/2) \tag{1.1}$$

so the film adheres on a part of its lateral boundary given by

$$\gamma \times (-h/2,\, h/2) \subset \partial S \times (-h/2,\, h/2).$$

The elastic energy of the film is given by

$$\mathcal{E}_h(\tilde{u}) = \kappa \int_{\Omega_h} |D(\nabla \tilde{u})| + \int_{\Omega_h} \phi(\nabla \tilde{u}(x), x)\, dx, \tag{1.2}$$

where the term $\kappa \int_{\Omega_h} |D(\nabla \tilde{u})|$ for $\kappa > 0$ models the interfacial energy between phases and variants (the total variation of the deformation gradient is precisely defined in section 2), and the term $\int_{\Omega_h} \phi(\nabla \tilde{u}(x), x)\, dx$ models the elastic energy of the film.

*Received by the editors January 12, 2005; accepted for publication (in revised form) December 19, 2005; published electronically May 12, 2006. This work was supported in part by the National Science Foundation under grant DMS-0304326, the Institute for Mathematics and its Applications, and by the Minnesota Supercomputer Institute.
http://www.siam.org/journals/sima/38-2/62259.html
†Department of Mathematics, University of St. Thomas, 2115 Summit Avenue, St. Paul, MN 55105 (pbelik@stthomas.edu).
‡School of Mathematics, University of Minnesota, 206 Church Street SE, Minneapolis, MN 55455 (luskin@math.umn.edu).

Since we are interested in modeling and computing the deformation of thin films that undergo structural phase transformation, the energy density, $\phi(F,x)$, is generally a nonconvex function of the deformation gradient, $F \in \mathbb{R}^{3\times 3}$. The explicit dependence of the energy density, $\phi(F,x)$, on $x \in \Omega_h$ allows the modeling of alloys with compositional fluctuation [11, 12, 14, 19]. We rescale the deformations $\tilde{u} : \Omega_h \to \mathbb{R}^3$ to deformations on a fixed domain of thickness one, $u : \Omega_1 \to \mathbb{R}^3$, by

$$u(z_1, z_2, z_3) = \tilde{u}(z_1, z_2, hz_3) \quad \text{for } z = (z_1, z_2, z_3) \in \Omega_1,$$

and we determine and analyze the Γ-limit of the rescaled energy

$$\mathcal{E}_1^{(h)}(u) = \frac{1}{h}\,\mathcal{E}_h(\tilde{u})$$

subject to rescaled boundary conditions

$$u(z_1, z_2, z_3) = y_0(z_1, z_2) + b_0(z_1, z_2)hz_3 \quad \text{for } (z_1, z_2, z_3) \in \gamma \times (-1/2,\, 1/2). \tag{1.3}$$

We analyze the Γ-limit of $\mathcal{E}_1^{(h)}(u)$ with respect to two related definitions of convergence for deformations. For the first definition, we prove that the Γ-limit of $\mathcal{E}_1^{(h)}(u)$ is given by

$$\mathcal{E}^{(0)}(y,b) = \kappa\left[\int_S |D(\nabla y|\sqrt{2}\,b)| + \sqrt{2}\int_\gamma |b - b_0|\right] + \int_S \phi(\nabla y(\hat{z})|b(\hat{z}), \hat{z}, 0)\,d\hat{z} \tag{1.4}$$

for $y : S \to \mathbb{R}^3$ such that $y = y_0$ on $\gamma \subset \partial S$ and $b : S \to \mathbb{R}^3$. The matrix-valued function $(\nabla y|b) : S \to \mathbb{R}^{3\times 3}$ in the thin film limit (1.4) models the thin film deformation gradient. We also identify $(\hat{z}, 0) \in \mathbb{R}^3$ with $\hat{z} \in S$. For the second definition, we prove that the Γ-limit of $\mathcal{E}_1^{(h)}(u)$ is given by

$$\mathcal{E}_1^{(0)}(u) = \begin{cases} \min\limits_b \mathcal{E}^{(0)}(u_M, b) & \text{if } u_{,3} = 0 \text{ a.e. in } \Omega_1, \\ +\infty & \text{otherwise,} \end{cases}$$

where u_M is the deformation of the midplane, $u_M(z_1, z_2) = u(z_1, z_2, 0)$. For both definitions of convergence of deformations, we give compactness results and show that the uniform coerciveness of the energy functionals $\mathcal{E}_1^{(h)}(u)$ allows us to prove that subsequences of energy-minimizing deformations of $\mathcal{E}_1^{(h)}(u)$ converge to minimizers of the Γ-limit as $h \to 0$.

We have used the thin film energy (1.4) to compute the quasi-static evolution of a martensitic thin film subject to a varying temperature field [8, 9]. In these computations, we use continuation methods for which the film need only be in a local minimum. We think that the results in this paper, especially the Γ-convergence described in Theorem 5.3, justify the use of the thin film energy (1.4) in this context because the result (5.10) guarantees that any admissible (y, b) defined on S can be used to construct an admissible $\tilde{u}_h$ defined on Ω_h that is a "smoothed" version of the deformation

$$y(x_1, x_2) + b(x_1, x_2)x_3 \quad \text{for } (x_1, x_2, x_3) \in \Omega_h,$$

such that $\mathcal{E}^{(0)}(y, b)$ is approximated by $\frac{1}{h}\mathcal{E}_h(\tilde{u}_h)$.

The energy density, $\phi(F,x)$, in models for crystals which undergo a structural phase transformation is not quasi-convex [4, 5, 21, 22, 23, 26, 27]. The Γ-limit with

respect to weak $W^{1,p}$ convergence of a scaled elastic energy that does not include interfacial energy will generally thus involve the quasi-convexification of the elastic energy density [7, 15]. However, the interfacial energy $\kappa \int_{\Omega_h} |D(\nabla \tilde{u})|$ in our model allows us to obtain sequences of deformations with gradients that converge strongly and use the strong continuity of the scaled elastic energy $h^{-1} \int_{\Omega_h} \phi(\nabla \tilde{u}(x), x)\, dx$. A related result has been obtained in [6] for a diffuse interfacial energy $\kappa \int_{\Omega_h} |\nabla^2 \tilde{u}|^2\, dx$.

The loading of a martensitic thin film can be computed by using the Γ-limit of the energy

$$\begin{aligned} \hat{\mathcal{E}}_h(\tilde{u}) &= \kappa \int_{\Omega_h} |D(\nabla \tilde{u})| + \int_{\Omega_h} \phi(\nabla \tilde{u}(x), x)\, dx - \int_{\partial \Omega_h} (Tn) \cdot \tilde{u} \\ &= \kappa \int_{\Omega_h} |D(\nabla \tilde{u})| + \int_{\Omega_h} \phi(\nabla \tilde{u}(x), x)\, dx - \int_{\Omega_h} T \cdot \nabla \tilde{u}, \end{aligned} \tag{1.5}$$

where the dead load is Tn for a constant $T \in \mathbb{R}^{3\times 3}$ at points on the boundary $\partial\Omega_h$ with unit exterior normal vector n. The Γ-limit is shown in this paper to be

$$\hat{\mathcal{E}}_1^{(0)}(u) = \begin{cases} \min\limits_b \hat{\mathcal{E}}^{(0)}(u_M, b) & \text{if } u_{,3} = 0 \text{ a.e. in } \Omega_1, \\ +\infty & \text{otherwise,} \end{cases}$$

with

$$\hat{\mathcal{E}}^{(0)}(y, b) = \kappa \int_S |D(\nabla y |\sqrt{2}\, b)| + \int_S \phi(\nabla y(\hat{z}) | b(\hat{z}), \hat{z}, 0)\, d\hat{z} - \int_S T \cdot (\nabla y | b).$$

In section 2, we recall the total variation of functions of bounded variation and give a few needed properties. In section 3, we describe the assumed properties of the elastic energy density, and in section 4 we recall the definition of Γ-convergence. The main results and analysis for the Γ-limit of the film model with Dirichlet boundary conditions are given in section 5, and the main results and analysis for the Γ-limit of the film model with loading boundary conditions are given in section 6.

Our results in this paper extend the analysis given in [10] by proving the Γ-convergence of the scaled energy functional for the adhering boundary condition (1.1). We also extend the class of energy densities, $\phi(F, x)$, to allow compositional variation, and we extend the class of boundary conditions that can be analyzed by giving results for the Γ-limit of the scaled energy functional with dead loads (1.5).

2. Functions of bounded variation. We will assume that $S \subset \mathbb{R}^2$ is a bounded domain with a Lipschitz continuous boundary, ∂S, and denote the reference undistorted configuration of the thin film of the martensitic material by Ω_h, $0 < h \leq 1$, where

$$\Omega_h = S \times (-h/2,\, h/2).$$

The deformations of the thin film are given by functions $\tilde{u} : \Omega_h \to \mathbb{R}^3$ with gradient $\nabla \tilde{u} : \Omega_h \to \mathbb{R}^{3\times 3}$. We use the notation $\tilde{u}_{i,j} = \partial \tilde{u}_i / \partial x_j$, and we denote the columns of $\nabla \tilde{u}$ by $\tilde{u}_{,i}$, $i = 1, 2, 3$. The "planar" gradient of $\tilde{u}$, denoted by $\nabla_P \tilde{u} : \Omega_h \to \mathbb{R}^{3\times 2}$, has columns given by $\tilde{u}_{,1}$ and $\tilde{u}_{,2}$.

Given an open set $\Omega \subset \mathbb{R}^3$ and a function $v \in L^1(\Omega; \mathbb{R})$, we define the total

variation of v [16, 18] by

$$\int_\Omega |Dv| = \sup\Big\{ \int_\Omega \sum_{k=1,2,3} v(x)\psi_{k,k}(x)\,dx : \psi \in \mathcal{C}_0^\infty(\Omega;\mathbb{R}^3),\ |\psi(x)| \le 1 \ \text{ for all } x \in \Omega \Big\}$$

and say $v \in BV(\Omega)$ if $\int_\Omega |Dv| < +\infty$. We recall that $\mathcal{C}_0^\infty(\Omega;\mathbb{R}^3)$ denotes the space of infinitely differentiable functions compactly supported in Ω, whose range is $\mathbb{R}^3$, and we note that $|\psi(x)|$ denotes the usual Euclidean norm, that is, the square root of the sum of the squares of all the components of $\psi(x)$.

For a matrix-valued function $v \in L^1(\Omega;\mathbb{R}^{m\times p})$, we define

$$\int_\Omega |Dv| = \sup\Big\{ \sum_{\substack{i=1,\dots,m\\ j=1,\dots,p\\ k=1,2,3}} \int_\Omega v_{ij}(x)\psi_{ijk,k}(x)\,dx : \psi \in \mathcal{C}_0^\infty(\Omega;\mathbb{R}^{m\times p\times 3}),\ |\psi(x)| \le 1 \text{ for all } x \in \Omega \Big\} \tag{2.1}$$

and say $v \in BV(\Omega)$ if $\int_\Omega |Dv| < +\infty$. We again assume that $|\psi(x)|$ denotes the square root of the sum of the squares of all the components of $\psi(x)$, which is often called the Frobenius norm. Finally, we define the "planar" variation

$$\int_\Omega |D_P v| = \sup\Big\{ \sum_{\substack{i=1,\dots,m\\ j=1,\dots,p\\ k=1,2}} \int_\Omega v_{ij}(x)\psi_{ijk,k}(x)\,dx : \psi \in \mathcal{C}_0^\infty(\Omega;\mathbb{R}^{m\times p\times 2}),\ |\psi(x)| \le 1 \text{ for all } x \in \Omega \Big\}.$$

For a matrix-valued function $v \in L^1(S;\mathbb{R}^{m\times p})$ we similarly define

$$\int_S |Dv| = \sup\Big\{ \sum_{\substack{i=1,\dots,m\\ j=1,\dots,p\\ k=1,2}} \int_S v_{ij}(x)\psi_{ijk,k}(x)\,dx : \psi \in \mathcal{C}_0^\infty(S;\mathbb{R}^{m\times p\times 2}),\ |\psi(x)| \le 1 \text{ for all } x \in S \Big\}.$$

We remark that if $v \in BV(\Omega_1)$ is independent of z_3, then, abusing the notation slightly, we have

$$\int_{\Omega_1} |Dv| = \int_{\Omega_1} |D_P v| = \int_S |Dv|.$$

The notation $BV_q(\Omega)$ will denote the space $BV(\Omega) \cap L^q(\Omega)$.

For $A \in \mathbb{R}^{m\times p}$ and $B \in \mathbb{R}^{m\times q}$, we denote by $(A|B) \in \mathbb{R}^{m\times(p+q)}$ the matrix whose first p columns are those of A and whose last q columns are those of B. For

$v \in L^1(\Omega_1;\mathbb{R}^{m\times p})$ and $b \in L^1(\Omega_1;\mathbb{R}^m)$, we will use the identity

$$\int_{\Omega_1} |D(v|\sqrt{2}b)| = \int_{\Omega_1} |D(v|b|b)|. \tag{2.2}$$

We will use the following extension of the classical result on the lower semicontinuity of the BV seminorm [16, 18] to functions with fixed trace [10].

THEOREM 2.1. *If $w_j, b_j \in BV(\Omega_1)$ for $j \in \mathbb{N}$ and $w, b \in BV(\Omega_1)$ satisfy*

$$\lim_{j\to\infty} \|w_j - w\|_{L^1(\Omega_1)} = 0 \quad \text{and} \quad \lim_{j\to\infty} \|b_j - b\|_{L^1(\Omega_1)} = 0,$$

and $b_j = b_0$ on $\Gamma_1 = \gamma \times (-\frac{1}{2}, \frac{1}{2})$ for fixed $b_0 \in BV(\Omega_1)$, then

$$\int_{\Omega_1} |D_P(w|\sqrt{2}\, b)| + \sqrt{2}\int_{\Gamma_1} |b - b_0| \le \liminf_{j\to\infty} \int_{\Omega_1} |D_P(w_j|\sqrt{2}\, b_j)|.$$

We will also use the following extension of the classical result on the approximation by smooth functions in the BV seminorm [16, 18] to functions with fixed trace [10].

THEOREM 2.2. *Let $1 \le q < +\infty$, let $b_0 \in W^{1,q}(S)$ be such that $\nabla b_0 \in BV(S)$, let $b \in BV_q(S)$, and let $w \in BV(S)$. Then there exists a family $\{b_\varepsilon : \varepsilon > 0\} \subset W^{1,q}(S)$ with $\nabla b_\varepsilon \in BV(S)$ such that $b_\varepsilon = b_0$ on γ for every $\varepsilon > 0$, and*

$$\lim_{\varepsilon\to 0} \|b_\varepsilon - b\|_{L^q(S)} = 0,$$

$$\lim_{\varepsilon\to 0} \int_S |D(w|\sqrt{2}\, b_\varepsilon)| = \int_S |D(w|\sqrt{2}\, b)| + \sqrt{2}\int_\gamma |b - b_0|.$$

3. The elastic energy density ϕ. We will assume that the energy density $\phi : \mathbb{R}^{3\times 3} \times \Omega_1 \to \mathbb{R}$ satisfies the Carathéodory condition [24]:

1. $\phi(F, \hat{z}, z_3)$ is continuous in $(F, z_3) \in \mathbb{R}^{3\times 3} \times (-1/2, 1/2)$ for almost every $\hat{z} \in S$,
2. $\phi(F, \hat{z}, z_3)$ is measurable in $\hat{z} \in S$ for every $(F, z_3) \in \mathbb{R}^{3\times 3} \times (-1/2, 1/2)$,

and satisfies the growth condition

$$c_1|F|^p - c_2 \le \phi(F, z) \le c_3(|F|^p + 1) \quad \text{for all } F \in \mathbb{R}^{3\times 3} \text{ and } z \in \Omega_1, \tag{3.1}$$

where c_1, c_2, and c_3 are fixed positive constants and $3 < p < +\infty$.

We can obtain this energy density ϕ from a free energy density $\hat{\phi}(F, \theta, c)$, where $\theta(z)$ is a given temperature and $c(z)$ is a given order parameter such as alloy composition, by $\phi(F, z) = \hat{\phi}(F, \theta(z), c(z))$. In what follows, we will usually not denote the explicit dependence of ϕ on z. Notice that ϕ is bounded below and its absolute value satisfies the growth property

$$|\phi(F, z)| \le c_3|F|^p + \max\{c_2, c_3\} \quad \text{for all } F \in \mathbb{R}^{3\times 3} \text{ and } z \in \Omega_1.$$

4. The Γ-limit. We now give a definition of Γ-convergence [3, 7, 24] that allows the domain $\tilde{\mathcal{A}}$ of the approximating functionals $\mathcal{F}_h$ to be different than the domain $\mathcal{A}$ of the Γ-limit $\mathcal{F}$.

DEFINITION 4.1. *Let $\mathcal{A}$ and $\tilde{A}$ be spaces such that the convergence of elements of $\tilde{\mathcal{A}}$ to an element of $\mathcal{A}$ is defined. We say that the family of functionals $\{\mathcal{F}_h : \tilde{\mathcal{A}} \to \mathbb{R} \cup \{+\infty\}$ for $h > 0\}$ Γ-converges to $\mathcal{F} : \mathcal{A} \to \mathbb{R} \cup \{+\infty\}$ as $h \to 0$ if the following*

two conditions are satisfied:

Lower bound: *given any* $u \in \mathcal{A}$ *and any family* $\{u_h \in \tilde{\mathcal{A}} : \ h > 0\}$ *such that* $u_h \to u$ *as* $h \to 0$*, we have*

$$\mathcal{F}(u) \le \liminf_{h\to 0} \mathcal{F}_h(u_h);$$

Upper bound: *given any* $u \in \mathcal{A}$*, there exists a family* $\{u_h \in \tilde{\mathcal{A}} : \ h > 0\}$ *such that* $u_h \to u$ *as* $h \to 0$ *and*

$$\mathcal{F}(u) \ge \limsup_{h\to 0} \mathcal{F}_h(u_h),$$

or equivalently, in view of the lower bound above,

$$\mathcal{F}(u) = \lim_{h\to 0} \mathcal{F}_h(u_h).$$

We note that the first condition above (Lower bound) guarantees that $\mathcal{F}$ is below the Γ-limit (if it exists), and the second condition (Upper bound) guarantees that $\mathcal{F}$ is above the Γ-limit (if it exists). If $\mathcal{F}$ satisfies both conditions, then $\mathcal{F}$ is the Γ-limit.

5. Γ-limit of the film model with Dirichlet boundary conditions. In this section, we assume that the film adheres to a rigid material on its lateral surface

$$\Gamma_h = \gamma \times (-h/2,\, h/2),$$

where we assume that $\gamma \neq \emptyset$ is a finite union of connected $\mathcal{C}^{1,1}$ open subsets of ∂S. Let $y_0, b_0 \in W^{1,p}(S;\mathbb{R}^3)$ be such that $\nabla y_0, \nabla b_0 \in BV(S)$ and define the boundary condition

$$\tilde{u}_0(x_1,x_2,x_3) = y_0(x_1,x_2) + b_0(x_1,x_2)x_3 \quad \text{for } (x_1,x_2,x_3) \in \Omega_h. \tag{5.1}$$

We then define the space $\mathcal{A}_h$ of admissible deformations of the domain Ω_h by

$$\mathcal{A}_h = \left\{\tilde{u} \in W^{1,p}(\Omega_h;\mathbb{R}^3) : \ \nabla\tilde{u} \in BV(\Omega_h),\ \tilde{u} = \tilde{u}_0 \text{ on } \Gamma_h\right\}.$$

We note that due to the growth condition (3.1), we have that

$$\mathcal{A}_h = \left\{\tilde{u} : \Omega_h \to \mathbb{R}^3 : \ \mathcal{E}_h(\tilde{u}) < +\infty,\ \tilde{u} = \tilde{u}_0 \text{ on } \Gamma_h\right\}.$$

Also, since $p > 3$, it follows from the Sobolev embedding theorem [2] that $\mathcal{A}_h \subset \mathcal{C}(\bar{\Omega}_h)$. This ensures that there is no tear in the deformed configurations $\tilde{u}(\Omega_h)$ for $\tilde{u} \in \mathcal{A}_h$.

We are interested in studying the thin film limit of the energies

$$\mathcal{E}_h(\tilde{u}) = \kappa \int_{\Omega_h} |D(\nabla\tilde{u})| + \int_{\Omega_h} \phi(\nabla\tilde{u}(x),x)\,dx, \qquad \tilde{u} \in \mathcal{A}_h, \tag{5.2}$$

where the constant $\kappa > 0$ is a measure of interfacial energy per unit area. We rescale the deformations $\tilde{u} : \Omega_h \to \mathbb{R}^3$ to deformations on a fixed domain of thickness one, $u : \Omega_1 \to \mathbb{R}^3$, via

$$u(z_1,z_2,z_3) = \tilde{u}(z_1,z_2,hz_3) \quad \text{for } z = (z_1,z_2,z_3) \in \Omega_1, \tag{5.3}$$

and we then study the Γ-convergence as $h \to 0$ of the rescaled energy

$$\mathcal{E}_1^{(h)}(u) = \frac{1}{h}\,\mathcal{E}_h(\tilde{u}) \tag{5.4}$$

for u defined in the space of admissible deformations

$$\mathcal{A}_1 = \left\{ u \in W^{1,p}(\Omega_1;\mathbb{R}^3) :\ \nabla u \in BV(\Omega_1),\ u = u_0 \text{ on } \Gamma_1 \right\},$$

where u_0 is defined by (5.1) and (5.3) to be

$$u_0(z_1,z_2,z_3) = y_0(z_1,z_2) + b_0(z_1,z_2)hz_3 \quad \text{for } (z_1,z_2,z_3) \in \Omega_1.$$

We will first define a topology for the convergence of $u_h \in \mathcal{A}_1$ to $(y,b) \in \mathcal{A}_0$, where

$$\mathcal{A}_0 = \{(y,b) \in W^{1,p}(S;\mathbb{R}^3) \times L^p(S;\mathbb{R}^3) :\ \nabla y, b \in BV(S),\ y = y_0 \text{ on } \gamma\}, \tag{5.5}$$

and we will then show that the Γ-limit of $\mathcal{E}_1^{(h)}$ is given by $\mathcal{E}^{(0)}$, where

$$\mathcal{E}^{(0)}(y,b) = \kappa \left[\int_S |D(\nabla y|\sqrt{2}\, b)| + \sqrt{2} \int_\gamma |b - b_0| \right] + \int_S \phi(\nabla y(\hat{z})|b(\hat{z}),\hat{z},0)\, d\hat{z}. \tag{5.6}$$

Here we use Definition 4.1 with $\mathcal{A} = \mathcal{A}_1$ and $\tilde{\mathcal{A}} = \mathcal{A}_0$. We note that above and in what follows we will often use the notation

$$\int_S \phi(\nabla y|b) = \int_S \phi(\nabla y(\hat{z})|b(\hat{z}),\hat{z},0)\, d\hat{z}.$$

In a second approach, we will set $\mathcal{A} = \tilde{\mathcal{A}} = \mathcal{A}_1$ with the topology on the space $\mathcal{A}_1$ given by weak $W^{1,p}$ convergence, and we will prove that a related functional $\mathcal{E}_1^{(0)}$ is the Γ-limit of $\mathcal{E}_1^{(h)}$ as $h \to 0$. The relation between $\mathcal{E}^{(0)}$ and $\mathcal{E}_1^{(0)}$ will become clear.

We now consider the Γ-convergence of $\mathcal{E}_1^{(h)}$ to $\mathcal{E}^{(0)}$ as $h \to 0$. We start by introducing a notion of the convergence of 3-D deformations $\{u_h\} \subset \mathcal{A}_1$ to a 2-D deformation $(y,b) \in \mathcal{A}_0$ as $h \to 0$.

DEFINITION 5.1. *We shall say that a family* $\{u_h \in \mathcal{A}_1 :\ h > 0\}$ *converges to* $(y,b) \in \mathcal{A}_0$ *if the following conditions are satisfied for* $\hat{y}(z_1,z_2,z_3) = y(z_1,z_2)$ *and* $\hat{b}(z_1,z_2,z_3) = b(z_1,z_2)$*:*

$$\left.\begin{array}{llllll} u_h \rightharpoonup \hat{y} & \text{in } W^{1,p}(\Omega_1;\mathbb{R}^3) & \text{and} & h^{-1}u_{h,3} \rightharpoonup \hat{b} & \text{in } L^p(\Omega_1;\mathbb{R}^3) \\ u_h \to \hat{y} & \text{in } W^{1,1}(\Omega_1;\mathbb{R}^3) & \text{and} & h^{-1}u_{h,3} \to \hat{b} & \text{in } L^1(\Omega_1;\mathbb{R}^3) \end{array}\right\} \quad \text{as } h \to 0.$$

We shall use this definition of convergence when proving the Γ-convergence of the functionals $\mathcal{E}_1^{(h)}$ to $\mathcal{E}^{(0)}$ since it allows the compactness property of Lemma 5.2 for sequences of deformations

$$\{u_{h_n} \in \mathcal{A}_1 : n = 1,\dots, \text{ and } h_n \to 0 \text{ as } n \to \infty\}$$

with uniformly bounded energy $\mathcal{E}_1^{(h_n)}(u_{h_n}) \le C$ for all $n \ge 1$. This compactness property can then be used with the Γ-convergence of the functionals $\mathcal{E}_1^{(h)}$ to $\mathcal{E}^{(0)}$ to give a proof of the convergence of minimizers of $\mathcal{E}_1^{(h)}$ to minimizers of the Γ-limit $\mathcal{E}^{(0)}$ (see Corollary 5.4 following the proof of Theorem 5.3). We will see from the proof of Theorem 5.3 that $\mathcal{E}^{(0)}$ is also the Γ-limit of $\mathcal{E}_1^{(h)}$ if we use the strong convergence

$$u_h \to \hat{y} \quad \text{in } W^{1,p}(\Omega_1;\mathbb{R}^3) \qquad \text{and} \qquad h^{-1}u_{h,3} \to \hat{b} \quad \text{in } L^p(\Omega_1;\mathbb{R}^3)$$

in Definition 5.1, but we do not have a compactness property for this topology since $BV(\Omega_1)$ is not compactly embedded in $L^p(\Omega_1)$ if $p \geq \frac{3}{2}$ [18].

LEMMA 5.2. *Suppose that* $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots$ *and* $h_n \to 0$ *as* $n \to \infty\}$ *is a sequence of deformations with uniformly bounded energy* $\mathcal{E}_1^{(h_n)}(u_{h_n}) \leq C$ *for all* $n \geq 1$. *Then there exists a further subsequence, also denoted by* $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots\}$, *and* $(y, b) \in \mathcal{A}_0$ *such that* $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots\}$ *converges to* $(y, b) \in \mathcal{A}_0$ *in the sense of Definition* 5.1. *We may take a further subsequence such that the convergence is also almost everywhere in* Ω_1.

Proof. We have from the definition of the total variation for matrix valued functions (2.1) that

$$\begin{aligned}
&\frac{1}{h_n}\int_{\Omega_{h_n}} |D(\nabla \tilde{u}_{h_n})| \\
&\quad = \sup\Bigg\{ \sum_{\substack{i=1,2,3\\ j,k=1,2}} \int_{\Omega_1} (u_{h_n})_{i,j}\psi_{ijk,k} + \sum_{\substack{i=1,2,3\\ j=1,2}} \int_{\Omega_1} {h_n}^{-1}(u_{h_n})_{i,j}\psi_{ij3,3} \\
&\qquad + \sum_{\substack{i=1,2,3\\ k=1,2}} \int_{\Omega_1} {h_n}^{-1}(u_{h_n})_{i,3}\psi_{i3k,k} + \sum_{i=1,2,3} \int_{\Omega_1} {h_n}^{-2}(u_{h_n})_{i,3}\psi_{i33,3} : \\
&\qquad\qquad \psi \in \mathcal{C}_0^\infty(\Omega_1),\ |\psi(z)| \leq 1 \ \text{ for all } z \in \Omega_1 \Bigg\}.
\end{aligned} \tag{5.7}$$

Since $\mathcal{E}_1^{(h_n)}(u_{h_n}) \leq C$ for all $n \geq 1$, we have by the growth condition (3.1) that

$$\|u_{h_n}\|_{W^{1,p}(\Omega_1;\mathbb{R}^3)} \leq C, \qquad \|h_n^{-1}u_{h_n,3}\|_{L^p(\Omega_1;\mathbb{R}^3)} \leq C, \tag{5.8}$$

and we have by (5.7) that

$$\begin{aligned}
&\int_{\Omega_1} |D(\nabla u_{h_n})| \leq C, \qquad \int_{\Omega_1} |D(h_n^{-1}u_{h_n,3})| \leq C, \\
&\sup\Bigg\{ \sum_{i=1,2,3} \int_{\Omega_1} h_n^{-2}(u_{h_n})_{i,3}\psi_{i33,3} : \\
&\qquad \psi \in \mathcal{C}_0^\infty(\Omega_1),\ |\psi(z)| \leq 1 \ \text{ for all } z \in \Omega_1 \Bigg\} \leq C
\end{aligned} \tag{5.9}$$

for all $n \geq 1$. It then follows from the compactness of the BV spaces [18] and the trace theorem [2] that there exists $\hat{u} \in W^{1,p}(\Omega_1;\mathbb{R}^3)$ such that $\nabla\hat{u} \in BV(\Omega_1)$ and $\hat{u} = y_0$ on $\gamma \times (-\frac{1}{2}, \frac{1}{2})$ and that there exists $\hat{b} \in BV_p(\Omega_1)$ such that for a further subsequence of $\{u_{h_n}\}$, not relabeled, we have that

$$\left.\begin{array}{llll} u_{h_n} \rightharpoonup \hat{u} & \text{in } W^{1,p}(\Omega_1;\mathbb{R}^3) & \text{and} & h_n^{-1}u_{h_n,3} \rightharpoonup \hat{b} \quad \text{in } L^p(\Omega_1;\mathbb{R}^3) \\ u_{h_n} \to \hat{u} & \text{in } W^{1,1}(\Omega_1;\mathbb{R}^3) & \text{and} & h_n^{-1}u_{h_n,3} \to \hat{b} \quad \text{in } L^1(\Omega_1;\mathbb{R}^3) \end{array}\right\} \ \text{as } n \to \infty,$$

and the convergence is also almost everywhere in Ω_1. In addition, from (5.8) and (5.9) it follows that $\hat{u}$ and $\hat{b}$ are independent of z_3, so we can set $y(z_1, z_2) = \hat{u}(z_1, z_2, z_3)$ and $b(z_1, z_2) = \hat{b}(z_1, z_2, z_3)$ to prove the lemma. ∎

We have the following Γ-convergence theorem.

THEOREM 5.3. *The functional* $\mathcal{E}^{(0)} : \mathcal{A}_0 \to \mathbb{R}$ *is the* Γ*-limit of the functionals* $\mathcal{E}_1^{(h)} : \mathcal{A}_1 \to \mathbb{R}$ *with respect to the convergence from Definition* 5.1*; that is,*

Lower bound: *given any* $(y, b) \in \mathcal{A}_0$ *and any family* $\{u_h \in \mathcal{A}_1 : \ h > 0\}$ *that converges to* (y, b)*, we have*

$$\mathcal{E}^{(0)}(y, b) \le \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h);$$

Upper bound: *given any* $(y, b) \in \mathcal{A}_0$*, there exists a family* $\{u_h \in \mathcal{A}_1 : \ h > 0\}$ *that converges to* (y, b) *such that*

$$\mathcal{E}^{(0)}(y, b) \ge \limsup_{h\to 0} \mathcal{E}_1^{(h)}(u_h),$$

or equivalently, in view of the lower bound above,

$$\mathcal{E}^{(0)}(y, b) = \lim_{h\to 0} \mathcal{E}_1^{(h)}(u_h). \tag{5.10}$$

Proof (lower bound). To prove the lower bound, let $(y, b) \in \mathcal{A}_0$ and let $\{u_h \in \mathcal{A}_1 : \ h > 0\}$ converge to (y, b) in the sense of Definition 5.1. Consider a subsequence $\{u_{h_n}\}_{n=1}^{\infty}$ such that

$$\lim_{n\to\infty} \mathcal{E}_1^{(h_n)}(u_{h_n}) = \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h)$$

and such that $\nabla_P u_{h_n} \to \nabla_P \hat{y}$ and $h_n^{-1} u_{h_n,3} \to \hat{b}$ almost everywhere in Ω_1 as $n \to \infty$. It follows from the identity (2.2) and

$$\int_\Omega v_{i,j}\psi_{ij3,3} = \int_\Omega v_{i,3}\psi_{ij3,j} \qquad \text{for all } v \in W^{1,1}(\Omega; \mathbb{R}^3) \text{ and } \psi_{ij3} \in \mathcal{C}_0^\infty(\Omega)$$

that

$$\begin{aligned}
&\int_{\Omega_1} |D_P(\nabla_P u_{h_n} | \sqrt{2}\, h_n^{-1} u_{h_n,3})| \\
&\quad = \int_{\Omega_1} |D_P(\nabla_P u_{h_n} | h_n^{-1} u_{h_n,3} | h_n^{-1} u_{h_n,3})| \\
&\quad = \sup \Bigg\{ \sum_{\substack{i=1,2,3\\ j,k=1,2}} \int_{\Omega_1} (u_{h_n})_{i,j} \psi_{ijk,k} + \sum_{\substack{i=1,2,3\\ j=1,2}} \int_{\Omega_1} {h_n}^{-1} (u_{h_n})_{i,3} \psi_{ij3,j} \\
&\qquad + \sum_{\substack{i=1,2,3\\ k=1,2}} \int_{\Omega_1} {h_n}^{-1} (u_{h_n})_{i,3} \psi_{i3k,k} : \\
&\qquad\qquad \psi \in \mathcal{C}_0^\infty(\Omega_1), \ |\psi(z)| \le 1 \ \text{ for all } z \in \Omega_1 \Bigg\} \\
&\quad \le \sup \Bigg\{ \sum_{\substack{i=1,2,3\\ j,k=1,2}} \int_{\Omega_1} (u_{h_n})_{i,j} \psi_{ijk,k} + \sum_{\substack{i=1,2,3\\ j=1,2}} \int_{\Omega_1} {h_n}^{-1} (u_{h_n})_{i,j} \psi_{ij3,3}
\end{aligned}$$

$$+\sum_{\substack{i=1,2,3\\k=1,2}}\int_{\Omega_1} {h_n}^{-1}(u_{h_n})_{i,3}\psi_{i3k,k}+\sum_{i=1,2,3}\int_{\Omega_1} {h_n}^{-2}(u_{h_n})_{i,3}\psi_{i33,3}:$$

$$\psi\in\mathcal{C}_0^\infty(\Omega_1),\ |\psi(z)|\le 1\ \text{ for all } z\in\Omega_1\Big\} \tag{5.11}$$

$$=\frac{1}{h_n}\int_{\Omega_{h_n}}|D(\nabla\tilde{u}_{h_n})|.$$

So, by using Theorem 2.1 on (5.11) and Fatou's lemma to control the ϕ term, we obtain that

$$\begin{aligned}\mathcal{E}^{(0)}(y,b)&=\kappa\left[\int_{\Omega_1}\left|D_P(\nabla_P\hat{y}|\sqrt{2}\,\hat{b})\right|+\sqrt{2}\int_{\Gamma_1}|\hat{b}-b_0|\right]+\int_{\Omega_1}\phi(\nabla_P\hat{y}|\hat{b},\hat{z},0)\,dz\\&\le\liminf_{n\to\infty}\mathcal{E}_1^{(h_n)}(u_{h_n})\\&=\lim_{n\to\infty}\mathcal{E}_1^{(h_n)}(u_{h_n})\\&=\liminf_{h\to0}\mathcal{E}_1^{(h)}(u_h),\end{aligned}$$

and this establishes the first part of the theorem. We note that above and in what follows we use the convention $z=(\hat{z},z_3)$ for $\hat{z}\in S$ and $z_3\in(-1/2,1/2)$.

Upper bound. To prove the upper bound, we would like to consider deformations of the form $y(z_1,z_2)+hz_3b(z_1,z_2)$; however, such deformations do not belong to $\mathcal{A}_1$ because b does not belong to $W^{1,p}(S;\mathbb{R}^3)$ and ∇b does not belong to $BV(S)$. We can overcome this problem by using Theorem 2.2: since $b_0\in W^{1,p}(S;\mathbb{R}^3)$ and $\nabla b_0\in BV(S)$, there exists a family of functions $b_\varepsilon\in W^{1,p}(S;\mathbb{R}^3)$ with $\nabla b_\varepsilon\in BV(S)$ such that $b_\varepsilon=b_0$ on γ for every $\varepsilon>0$, $b_\varepsilon\to b$ almost everywhere in S and in $L^p(S)$ as $\varepsilon\to0$, and

$$\lim_{\varepsilon\to0}\int_S|D(\nabla y|\sqrt{2}\,b_\varepsilon)|=\int_S|D(\nabla y|\sqrt{2}\,b)|+\sqrt{2}\int_\gamma|b-b_0|. \tag{5.12}$$

We construct the functions

$$u_h^\varepsilon(z_1,z_2,z_3)=y(z_1,z_2)+hz_3b_\varepsilon(z_1,z_2)\in\mathcal{A}_1\quad\text{ for } 0<h\le1.$$

Now $\nabla_Pu_h^\varepsilon=\nabla_Py+hz_3\nabla_Pb_\varepsilon\to\nabla_Py$ in $L^p(\Omega_1)$ and almost everywhere in Ω_1 as $h\to0$, so we can obtain by using the growth condition (3.1) for ϕ, the Carathéodory property of ϕ given in section 3, and the dominated convergence theorem that

$$\begin{aligned}\frac{1}{h}\int_{\Omega_h}\phi(\nabla\tilde{u}_h^\varepsilon(x),x)\,dx&=\int_{\Omega_1}\phi(\nabla_Pu_h^\varepsilon|h^{-1}u_{h,3}^\varepsilon,\hat{z},hz_3)\,dz\\&=\int_{\Omega_1}\phi(\nabla_Pu_h^\varepsilon|b_\varepsilon,\hat{z},hz_3)\,dz\to\int_{\Omega_1}\phi(\nabla_Py|b_\varepsilon,\hat{z},0)\,dz\end{aligned}$$

as $h\to0$. By the same argument,

$$\int_{\Omega_1}\phi(\nabla_Py|b_\varepsilon,\hat{z},0)\,dz=\int_S\phi(\nabla y|b_\varepsilon)\to\int_S\phi(\nabla y|b)\quad\text{ as }\varepsilon\to0,$$

so

$$\lim_{\varepsilon\to0}\lim_{h\to0}\frac{1}{h}\int_{\Omega_h}\phi(\nabla\tilde{u}_h^\varepsilon(x),x)\,dx=\int_S\phi(\nabla y|b). \tag{5.13}$$

We now have since b_ε is independent of z_3 that

$$\begin{aligned}
&\frac{1}{h}\int_{\Omega_h} |D(\tilde{u}_h^\varepsilon)| \\
&\quad = \sup\Bigg\{ \sum_{\substack{i=1,2,3\\ j,k=1,2}} \int_{\Omega_1} (u_h^\varepsilon)_{i,j}\psi_{ijk,k} + \sum_{\substack{i=1,2,3\\ j=1,2}} \int_{\Omega_1} h^{-1}(u_h^\varepsilon)_{i,3}\psi_{ij3,j} \\
&\qquad + \sum_{\substack{i=1,2,3\\ k=1,2}} \int_{\Omega_1} h^{-1}(u_h^\varepsilon)_{i,3}\psi_{i3k,k} + \sum_{i=1,2,3} \int_{\Omega_1} h^{-2}(u_h^\varepsilon)_{i,3}\psi_{i33,3} : \\
&\qquad\qquad \psi \in \mathcal{C}_0^\infty(\Omega),\ |\psi(z)| \le 1 \ \text{ for all } z \in \Omega_1 \Bigg\} \qquad (5.14)\\
&\quad = \sup\Bigg\{ \sum_{\substack{i=1,2,3\\ j,k=1,2}} \int_{\Omega_1} (u_h^\varepsilon)_{i,j}\psi_{ijk,k} + \sum_{\substack{i=1,2,3\\ j=1,2}} \int_{\Omega_1} b_\varepsilon\psi_{ij3,j} \\
&\qquad + \sum_{\substack{i=1,2,3\\ k=1,2}} \int_{\Omega_1} b_\varepsilon\psi_{i3k,k} : \psi \in \mathcal{C}_0^\infty(\Omega_1),\ |\psi(z)| \le 1 \ \text{ for all } z \in \Omega_1 \Bigg\} \\
&\quad = \int_{\Omega_1} \left| D_P(\nabla_P u_h^\varepsilon | \sqrt{2}\, b_\varepsilon) \right|.
\end{aligned}$$

Since $\nabla b_\varepsilon \in BV(S)$ and y and b_ε are independent of z_3, we have that

$$\begin{aligned}
&\lim_{h\to 0} \int_{\Omega_1} \left| D_P(\nabla_P u_h^\varepsilon | \sqrt{2}\, b_\varepsilon) \right| \\
&\qquad = \lim_{h\to 0} \int_{\Omega_1} |D_P(\nabla_P(y + hz_3 b_\varepsilon)|\sqrt{2}\, b_\varepsilon)| = \int_S |D(\nabla y|\sqrt{2}\, b_\varepsilon)|. \qquad (5.15)
\end{aligned}$$

It then follows from (5.14), (5.15), and (5.12) that

$$\begin{aligned}
\lim_{\varepsilon\to 0}\lim_{h\to 0} \frac{1}{h}\int_{\Omega_h} |D(\tilde{u}_h^\varepsilon)| &= \lim_{\varepsilon\to 0}\lim_{h\to 0} \int_{\Omega_1} |D_P(\nabla_P u_h^\varepsilon|\sqrt{2} b_\varepsilon)| \\
&= \int_S |D(\nabla y|\sqrt{2}\, b)| + \sqrt{2}\int_\gamma |b - b_0|. \qquad (5.16)
\end{aligned}$$

We can then conclude from (5.13) and (5.16) that

$$\lim_{\varepsilon\to 0}\lim_{h\to 0} \mathcal{E}_1^{(h)}(u_h^\varepsilon) = \mathcal{E}^{(0)}(y, b). \qquad (5.17)$$

We note that in view of (5.17), it is clear that for any $\eta > 0$ there exists $\varepsilon > 0$ and $h_0 > 0$ such that

$$|\mathcal{E}_1^{(h)}(u_h^\varepsilon) - \mathcal{E}^{(0)}(y, b)| < \eta \quad \text{ for all } 0 < h \le h_0. \qquad \square$$

COROLLARY 5.4. *For every sequence $\{u_h \in \mathcal{A}_1 : h \to 0\}$ of minimizers of $\mathcal{E}_1^{(h)}$, there exists a subsequence $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots$ and $h_n \to 0$ as $n \to \infty\}$ and a minimizer $(y, b) \in \mathcal{A}_0$ of $\mathcal{E}^{(0)}$ such that $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots\}$ converges to $(y, b) \in \mathcal{A}_0$ in the sense of Definition* 5.1.

Proof. We first note that $\mathcal{E}_1^{(h)}(u_0)$ is bounded independent of $h > 0$. We can thus prove the existence of minimizers, $u_h \in \mathcal{A}_1$, of the functional $\mathcal{E}_1^{(h)}$ for fixed $h > 0$ by using the bounds (5.8) and (5.9), the compactness and lower-semicontinuity of the BV spaces [18], and Fatou's lemma.

Since $\mathcal{E}_1^{(h)}(u_0)$ is bounded independent of $h > 0$, we have the uniform bound

$$\mathcal{E}_1^{(h)}(u_h) \le \mathcal{E}_1^{(h)}(u_0) \le C.$$

We let $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots \text{ and } h_n \to 0 \text{ as } n \to \infty\}$ be a subsequence such that

$$\lim_{n\to\infty} \mathcal{E}_1^{(h_n)}(u_{h_n}) = \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h).$$

We can conclude from Lemma 5.2 that there exists a further subsequence (not relabeled), $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots \text{ and } h_n \to 0 \text{ as } n \to \infty\}$, and $(y, b) \in \mathcal{A}_0$ such that $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots\}$ converges to $(y, b) \in \mathcal{A}_0$ in the sense of Definition 5.1. It follows from the lower bound in Theorem 5.3 that

$$\mathcal{E}^{(0)}(y, b) \le \lim_{n\to\infty} \mathcal{E}_1^{(h_n)}(u_{h_n}) = \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h).$$

Since $u_h \in \mathcal{A}_1$ are minimizers of $\mathcal{E}_1^{(h)}$, we can conclude from the upper bound in Theorem 5.3 that $\limsup_{h\to 0} \mathcal{E}_1^{(h)}(u_h) \le \mathcal{E}^{(0)}(y, b)$, so $\lim_{h\to 0} \mathcal{E}_1^{(h)}(u_h)$ exists and

$$\mathcal{E}^{(0)}(y, b) = \lim_{h\to 0} \mathcal{E}_1^{(h)}(u_h).$$

We can now conclude from the upper bound in Theorem 5.3 that $(y, b) \in \mathcal{A}_0$ is a minimizer of $\mathcal{E}^{(0)}$. ∎

We next address the question of Γ-convergence of $\mathcal{E}_1^{(h)}$ with respect to the weak convergence in $\mathcal{A}_1$. We start by considering the problem of minimizing $\mathcal{E}^{(0)}(y, b)$ with respect to b.

LEMMA 5.5. *Let $y \in W^{1,p}(S; \mathbb{R}^3)$ be such that $\nabla y \in BV(S)$ and $y = y_0$ on γ. Then*

$$\inf_{b\in BV_p(S;\mathbb{R}^3)} \mathcal{E}^{(0)}(y, b) = \inf_{\substack{b\in BV_p(S;\mathbb{R}^3)\\ b=b_0 \text{ on } \gamma}} \mathcal{E}^{(0)}(y, b). \tag{5.18}$$

Proof. It is clear that the left-hand side is less than or equal to the right-hand side since the infimum is taken over a larger space.

To show the opposite inequality, it is enough to show that for any $b \in BV_p(S; \mathbb{R}^3)$ the energy $\mathcal{E}^{(0)}(y, b)$ can be arbitrarily closely approximated by energies $\mathcal{E}^{(0)}(y, b)$ with $b \in BV_p(S; \mathbb{R}^3)$ such that $b = b_0$ on γ. However, this follows by applying Theorem 2.2 with $q = p$ and showing that the elastic energy term $\int_S \phi(\nabla y|b_\varepsilon)$ again converges to $\int_S \phi(\nabla y|b)$ as in the proof of the second part of Theorem 5.3. ∎

We next have that the infimum on the left-hand side of (5.18) in Lemma 5.5 is attained for any y.

LEMMA 5.6. *Let $y \in W^{1,p}(S; \mathbb{R}^3)$ be such that $\nabla y \in BV(S)$ and $y = y_0$ on γ. Then there exists a function $\tilde{b} \in BV_p(S; \mathbb{R}^3)$ such that*

$$\mathcal{E}^{(0)}(y, \tilde{b}) = \inf_{b\in BV_p(S;\mathbb{R}^3)} \mathcal{E}^{(0)}(y, b).$$

Proof. Since $\mathcal{E}^{(0)}$ is bounded below, we can consider a minimizing sequence $\{b_j\}_{j=1}^{\infty} \subset BV_p(S;\mathbb{R}^3)$; in view of Lemma 5.5, we can also assume that $b_j = b_0$ on γ for all $j \in \mathbb{N}$. Since the variations of the b_j and their L^p-norms (and thus also the L^1-norms) lie in a compact subset of $\mathbb{R}$, we can use the compactness of $BV(S;\mathbb{R}^3)$ [18] and retrieve a subsequence, not relabeled, which converges to a function $\tilde{b} \in BV_p(S;\mathbb{R}^3)$ strongly in $L^1(S;\mathbb{R}^3)$, weakly in $L^p(S;\mathbb{R}^3)$, and almost everywhere in S. In addition, by applying Theorem 2.1, we have

$$\int_S |D(\nabla y|\sqrt{2}\,\tilde{b})| + \sqrt{2}\int_\gamma |\tilde{b} - b_0| \le \liminf_{j\to\infty} \int_S |D(\nabla y|\sqrt{2}\,b_j)|.$$

Similarly, applying Fatou's lemma to $\phi(\nabla y|b_j)$ gives

$$\int_S \phi(\nabla y|\tilde{b}) \le \liminf_{j\to\infty} \int_S \phi(\nabla y|b_j),$$

and therefore

$$\begin{aligned}\mathcal{E}^{(0)}(y,\tilde{b}) &\le \liminf_{j\to\infty} \mathcal{E}^{(0)}(y,b_j)\\ &= \inf_{b\in BV_p(S;\mathbb{R}^3)} \mathcal{E}^{(0)}(y,b). \qquad \square\end{aligned}$$

We are now in the position to find the Γ-limit of the functionals $\mathcal{E}_1^{(h)} = \frac{1}{h}\,\mathcal{E}_h(\tilde{u})$. Given a continuous $u \in \mathcal{A}_1$, we can define the deformation of the midplane

$$u_M(z_1,z_2) = u(z_1,z_2,0)$$

and a functional

$$\mathcal{E}_1^{(0)}(u) = \begin{cases} \min\limits_{b\in BV_p(S;\mathbb{R}^3)} \mathcal{E}^{(0)}(u_M,b) & \text{if } u_{,3} = 0 \text{ a.e. in } \Omega_1,\\ +\infty & \text{otherwise.}\end{cases}$$

In what follows, C will denote a generic positive constant independent of h, which can change from line to line.

THEOREM 5.7. *The functional $\mathcal{E}_1^{(0)} : \mathcal{A}_1 \to \mathbb{R}\cup\{+\infty\}$ is the Γ-limit of the functionals $\mathcal{E}_1^{(h)} : \mathcal{A}_1 \to \mathbb{R}$ as $h \to 0$ with respect to the weak $W^{1,p}(\Omega_1;\mathbb{R}^3)$ convergence in $\mathcal{A}_1$; that is,*

Lower bound: *given any $u \in \mathcal{A}_1$ and any family $\{u_h \in \mathcal{A}_1 : h > 0\}$ such that $u_h \rightharpoonup u$ in $W^{1,p}(\Omega_1;\mathbb{R}^3)$ as $h \to 0$, we have*

$$\mathcal{E}_1^{(0)}(u) \le \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h);$$

Upper bound: *given any $u \in \mathcal{A}_1$, there exists a family $\{u_h \in \mathcal{A}_1 : h > 0\}$ such that $u_h \rightharpoonup u$ in $W^{1,p}(\Omega_1;\mathbb{R}^3)$ as $h \to 0$ and*

$$\mathcal{E}_1^{(0)}(u) \ge \limsup_{h\to 0} \mathcal{E}_1^{(h)}(u_h),$$

or equivalently, in view of the lower bound above,

$$\mathcal{E}_1^{(0)}(u) = \lim_{h\to 0} \mathcal{E}_1^{(h)}(u_h).$$

Proof (lower bound). Let $u \in \mathcal{A}_1$ and let $\{u_h : h > 0\} \subset \mathcal{A}_1$ be such that $u_h \rightharpoonup u$ in $W^{1,p}(\Omega_1)$ as $h \to 0$. If $\liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h) = +\infty$, then

$$\mathcal{E}_1^{(0)}(u) \le \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h)$$

is trivially satisfied.

On the other hand, if $\liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h) < +\infty$, then we can first consider a subsequence $\{u_{h_n}\}_{n=1}^{\infty}$ such that

$$\lim_{n\to\infty} \mathcal{E}_1^{(h_n)}(u_{h_n}) = \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h).$$

Since then $\mathcal{E}_1^{(h_n)}(u_{h_n}) \le C$ for all $n \ge 1$, we have by Lemma 5.2 that there exists $b \in BV_p(\Omega_1;\mathbb{R}^3)$ such that for a further subsequence of $\{u_{h_n}\}$, not relabeled, we have that

$$\left.\begin{array}{llllll} u_{h_n} \rightharpoonup u & \text{in } W^{1,p}(\Omega_1;\mathbb{R}^3) & \text{and} & h_n^{-1}u_{h_n,3} \rightharpoonup b & \text{in } L^p(\Omega_1;\mathbb{R}^3) \\ u_{h_n} \to u & \text{in } W^{1,1}(\Omega_1;\mathbb{R}^3) & \text{and} & h_n^{-1}u_{h_n,3} \to b & \text{in } L^1(\Omega_1;\mathbb{R}^3) \end{array}\right\} \quad \text{as } n \to \infty,$$

and the convergence is also almost everywhere in Ω_1. It also follows from Lemma 5.2 that u and b are independent of z_3. Therefore, by the definition of $\mathcal{E}_1^{(0)}(u)$, we have for $u_M(z_1,z_2) = u(z_1,z_2,0)$ and $b_M(z_1,z_2) = b(z_1,z_2,0)$ that

$$\mathcal{E}_1^{(0)}(u) \le \mathcal{E}^{(0)}(u_M, b_M). \tag{5.19}$$

Using (5.11), Theorem 2.1, and Fatou's lemma to control the ϕ term, we have that

$$\begin{aligned} \mathcal{E}^{(0)}(u_M,b_M) &= \kappa\left[\int_{\Omega_1} |D_P(\nabla_P u|\sqrt{2}\,b)| + \sqrt{2}\int_{\Gamma_1} |b-b_0|\right] + \int_{\Omega_1} \phi(\nabla_P u|b,\hat{z},0)\,d\hat{z} \\ &\le \liminf_{n\to\infty} \mathcal{E}_1^{(h_n)}(u_{h_n}) \\ &= \lim_{n\to\infty} \mathcal{E}_1^{(h_n)}(u_{h_n}) \\ &= \liminf_{h\to 0} \mathcal{E}_1^{(h)}(u_h). \end{aligned}$$

Combining the above result with (5.19) completes the first part of the proof.

Upper bound. If $u \in \mathcal{A}_1$ is not independent of z_3, then $\mathcal{E}_1^{(0)}(u) = +\infty$ and

$$\mathcal{E}_1^{(0)}(u) \ge \limsup_{h\to 0} \mathcal{E}_1^{(h)}(u_h)$$

holds for any family $\{u_h \in \mathcal{A}_1 : h > 0\}$ such that $u_h \rightharpoonup u$ in $W^{1,p}(\Omega_1)$ as $h \to 0$.

On the other hand, if $u_{,3} = 0$ a.e. in Ω_1, then by Lemma 5.6 there exists $\tilde{b} \in BV_p(S;\mathbb{R}^3)$ such that $\mathcal{E}_1^{(0)}(u) = \mathcal{E}^{(0)}(u_M,\tilde{b})$. Using the upper bound of Theorem 5.3, there exists a family $\{u_h \in \mathcal{A}_1 : h > 0\}$ such that

$$u_h \rightharpoonup u \quad \text{in } W^{1,p}(\Omega_1) \text{ as } h \to 0$$

and

$$\begin{aligned} \lim_{h\to 0} \mathcal{E}_1^{(h)}(u_h) &= \mathcal{E}^{(0)}(u_M,\tilde{b}) \\ &= \mathcal{E}_1^{(0)}(u). \qquad \square \end{aligned}$$

We note that the family $\{u_h \in \mathcal{A}_1 : \ h > 0\}$ constructed for the upper bound in Theorem 5.7 actually converges strongly; that is, $u_h \to u$ in $W^{1,p}(\Omega_1)$ as $h \to 0$. We thus have that the functional $\mathcal{E}_1^{(0)} : \mathcal{A}_1 \to \mathbb{R} \cup \{+\infty\}$ is also the Γ-limit of the functionals $\mathcal{E}_1^{(h)} : \mathcal{A}_1 \to \mathbb{R}$ as $h \to 0$ with respect to the strong $W^{1,p}(\Omega_1;\mathbb{R}^3)$ convergence in $\mathcal{A}_1$.

We can obtain the following result on the convergence of minimizers of $\mathcal{E}_1^{(h)}$ to minimizers of $\mathcal{E}_1^{(0)}$ by an argument analogous to that of Corollary 5.4.

COROLLARY 5.8. *For every sequence $\{u_h \in \mathcal{A}_1 : h \to 0\}$ of minimizers of $\mathcal{E}_1^{(h)}$, there exists a subsequence $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots$ and $h_n \to 0$ as $n \to \infty\}$ and a minimizer $u \in \mathcal{A}_1$ of $\mathcal{E}_1^{(0)}$ such that $\{u_{h_n} \in \mathcal{A}_1 : n = 1, \dots\}$ converges to $u \in \mathcal{A}_1$ with respect to weak $W^{1,p}(\Omega_1;\mathbb{R}^3)$ convergence in $\mathcal{A}_1$.*

6. Γ-limit of the dead-loaded film model. We now assume that the film is subject to a dead load Tn on its boundary, $\partial\Omega_h$, with unit exterior normal n, where $T \in \mathbb{R}^{3\times 3}$ is independent of $x \in \Omega_h$. In this case, the energy of the three-dimensional thin film is given by

$$\begin{aligned}\hat{\mathcal{E}}_h(\tilde{u}) &= \kappa \int_{\Omega_h} |D(\nabla \tilde{u})| + \int_{\Omega_h} \phi(\nabla \tilde{u}(x), x)\, dx - \int_{\partial\Omega_h} (Tn)\cdot \tilde{u} \\ &= \kappa \int_{\Omega_h} |D(\nabla \tilde{u})| + \int_{\Omega_h} \phi(\nabla \tilde{u}(x), x)\, dx - \int_{\Omega_h} T \cdot \nabla \tilde{u}.\end{aligned}$$

If the elastic energy density ϕ satisfies the growth condition (3.1), then we can define

$$\hat{\phi}(F, x) = \phi(F, x) - T \cdot F$$

and $\hat{\phi}$ still satisfies (3.1) for some different positive constants, which we still denote by c_1, c_2, and c_3.

In this case, we define the space $\hat{\mathcal{A}}_h$ of admissible deformations of the domain Ω_h by

$$\hat{\mathcal{A}}_h = \left\{ \tilde{u} \in W^{1,p}(\Omega_h;\mathbb{R}^3) : \ \nabla \tilde{u} \in BV(\Omega_h), \ \int_{\Omega_h} \tilde{u} = 0 \right\}.$$

The energies of the deformations $\tilde{u} \in \hat{\mathcal{A}}_h$ of films are again given by (5.2) with $\phi(F)$ replaced by $\phi(F) - T \cdot F$. As before, due to the growth condition (3.1), we have

$$\hat{\mathcal{A}}_h = \left\{ \tilde{u} : \Omega_h \to \mathbb{R}^3 : \ \hat{\mathcal{E}}_h(\tilde{u}) < +\infty, \ \int_{\Omega_h} \tilde{u} = 0 \right\} \subset \mathcal{C}(\hat{\Omega}_h).$$

The proof of the convergence to a Γ-limit for the problem of the dead-loaded film is similar to the proof for the film constrained on part of the boundary. We start with the rescaled energy $\hat{\mathcal{E}}_1^{(h)} : \hat{\mathcal{A}}_1 \to \mathbb{R}$ (defined by (5.2)–(5.4)), where the space of admissible deformations, $\hat{\mathcal{A}}_1$, is defined by

$$\hat{\mathcal{A}}_1 = \left\{ u \in W^{1,p}(\Omega_1;\mathbb{R}^3) : \ \nabla u \in BV(\Omega_1), \int_{\Omega_1} u = 0 \right\}.$$

We then show that the Γ-limit of $\hat{\mathcal{E}}_1^{(h)} : \hat{\mathcal{A}}_1 \to \mathbb{R}$ is given by

$$\hat{\mathcal{E}}^{(0)}(y, b) = \kappa \int_S |D(\nabla y|\sqrt{2}\, b)| + \int_S \phi(\nabla y(\hat{z})|b(\hat{z}), \hat{z}, 0)\, d\hat{z} - \int_S T \cdot (\nabla y|b) \quad \text{for } (y, b) \in \hat{\mathcal{A}}_0,$$

where the space of admissible deformations is given by

$$\hat{\mathcal{A}}_0 = \left\{ (y,b) \in W^{1,p}(S;\mathbb{R}^3) \times L^p(S;\mathbb{R}^3) : \ \nabla y, b \in BV(S), \ \int_S y = 0 \right\}.$$

The proof of the following compactness result for sequences

$$\{u_{h_n} \in \hat{\mathcal{A}}_1 : n = 1, \dots \text{ and } h_n \to 0 \text{ as } n \to \infty\}$$

is analogous to that of Lemma 5.2.

LEMMA 6.1. *Suppose that* $\{u_{h_n} \in \hat{\mathcal{A}}_1 : n = 1, \dots$ *and* $h_n \to 0$ *as* $n \to \infty\}$ *is a sequence of deformations with uniformly bounded energy* $\hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) \le C$ *for all* $n \ge 1$. *Then there exists a further subsequence, also denoted by* $\{u_{h_n} \in \hat{\mathcal{A}}_1 : n = 1, \dots\}$, *and* $(y,b) \in \hat{\mathcal{A}}_0$ *such that* $\{u_{h_n} \in \hat{\mathcal{A}}_1 : n = 1, \dots\}$ *converges to* $(y,b) \in \hat{\mathcal{A}}_0$ *in the sense of Definition* 5.1 *(with the spaces* $\mathcal{A}_1$ *and* $\mathcal{A}_0$ *replaced by* $\hat{\mathcal{A}}_1$ *and* $\hat{\mathcal{A}}_0$, *respectively). We may take a further subsequence such that the convergence is also almost everywhere in* Ω_1.

We have the following Γ-convergence theorem.

THEOREM 6.2. *The functional* $\hat{\mathcal{E}}^{(0)} : \hat{\mathcal{A}}_0 \to \mathbb{R}$ *is the* Γ*-limit of the functionals* $\hat{\mathcal{E}}_1^{(h)} : \hat{\mathcal{A}}_1 \to \mathbb{R}$ *with respect to the convergence from Definition* 5.1.

Proof (lower bound). Let $(y,b) \in \hat{\mathcal{A}}_0$ and let $\{u_h \in \hat{\mathcal{A}}_1 : \ h > 0\}$ converge to (y,b) in the sense of Definition 5.1. Consider a subsequence $\{u_{h_n}\}_{n=1}^{\infty}$ such that

$$\lim_{n\to\infty} \hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) = \liminf_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h)$$

and such that $\nabla_P u_{h_n} \to \nabla_P \hat{y}$ and $h_n^{-1} u_{h_n,3} \to \hat{b}$ almost everywhere in Ω_1 as $n \to \infty$. Using (5.11), the lower semicontinuity of the total variation, and Fatou's lemma to control the ϕ term, we have that

$$\begin{aligned}
\hat{\mathcal{E}}^{(0)}(y,b) &= \kappa \int_{\Omega_1} |D_P(\nabla_P \hat{y} | \sqrt{2}\, \hat{b})| + \int_{\Omega_1} \phi(\nabla_P \hat{y} | \hat{b}, \hat{z}, 0)\, dz - \int_S T \cdot (\nabla_P \hat{y} | \hat{b}) \\
&\le \liminf_{n\to\infty} \hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) \\
&= \lim_{n\to\infty} \hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) \\
&= \liminf_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h),
\end{aligned}$$

which establishes the first part of the theorem.

Upper bound. To prove the upper bound, one should again consider deformations of the form $y(z_1,z_2) + hz_3 b(z_1,z_2)$; as before, such deformations do not belong to $\hat{\mathcal{A}}_1$, because b does not belong to $W^{1,p}(S;\mathbb{R}^3)$. However, we can find a family of functions $b_\varepsilon \in \mathcal{C}^\infty(\hat{S};\mathbb{R}^3) \subset W^{1,p}(S;\mathbb{R}^3)$ such that $b_\varepsilon \to b$ almost everywhere in S and in $L^p(S)$ as $\varepsilon \to 0$ [13], and

$$\lim_{\varepsilon\to 0} \int_S |D(\nabla y | \sqrt{2}\, b_\varepsilon)| = \int_S |D(\nabla y | \sqrt{2}\, b)|. \tag{6.1}$$

Consider now the functions

$$w_h^\varepsilon(z_1,z_2,z_3) = y(z_1,z_2) + hz_3 b_\varepsilon(z_1,z_2) hz_3 \in \mathcal{A}_1 \quad \text{for } 0 < h \le 1,$$

and their mean-zero translations

$$u_h^\varepsilon = w_h^\varepsilon - \frac{1}{|\Omega_1|}\int_{\Omega_1} w_h^\varepsilon \, dz \in \hat{\mathcal{A}}_1.$$

We can now apply the same argument as in the proof of the upper bound in Theorem 5.3 to conclude that

$$\lim_{\varepsilon\to 0}\lim_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h^\varepsilon) = \hat{\mathcal{E}}^{(0)}(y,b),$$

from which it is clear that for any $\eta > 0$ there exist $\varepsilon > 0$ and $h_0 > 0$ such that

$$|\hat{\mathcal{E}}_1^{(h)}(u_h^\varepsilon) - \hat{\mathcal{E}}^{(0)}(y,b)| < \eta \quad \text{for all } 0 < h \le h_0. \qquad \square$$

We can obtain a result on the convergence of minimizers of $\mathcal{E}_1^{(h)}(u)$ to minimizers of $\mathcal{E}^{(0)}(y,b)$ by an argument analogous to that of Corollary 5.8.

The Γ-limit of $\hat{\mathcal{E}}_1^{(h)} : \hat{\mathcal{A}}_1 \to \mathbb{R}$ can again be obtained by minimizing out b in the energy $\hat{\mathcal{E}}^{(0)}(y,b)$. The existence of a minimizing $\tilde{b}$ can be shown by using the direct method of the calculus of variations as in Lemma 5.6.

LEMMA 6.3. *Let $y \in W^{1,p}(S;\mathbb{R}^3)$ be such that $\nabla y \in BV(S)$ and $\int_S y = 0$. Then there exists a function $\tilde{b} \in BV_p(S;\mathbb{R}^3)$ such that*

$$\hat{\mathcal{E}}^{(0)}(y,\tilde{b}) = \inf_{b\in BV_p(S;\mathbb{R}^3)} \hat{\mathcal{E}}^{(0)}(y,b).$$

Proof. Since $\hat{\mathcal{E}}^{(0)}$ is bounded below, we can consider a minimizing sequence $\{b_j\}_{j=1}^\infty \subset BV_p(S;\mathbb{R}^3)$. Since the variations of the b_j and their L^p-norms (and thus also the L^1-norms) lie in a compact subset of $\mathbb{R}$, we can use the compactness of $BV(S;\mathbb{R}^3)$ and retrieve a subsequence, not relabeled, which converges to a function $\tilde{b} \in BV_p(S;\mathbb{R}^3)$ strongly in $L^1(S;\mathbb{R}^3)$, weakly in $L^p(S;\mathbb{R}^3)$, and almost everywhere in S. From the lower semicontinuity of the total variation, we have

$$\int_S |D(\nabla y|\sqrt{2}\,\tilde{b}| \le \liminf_{j\to\infty} \int_S |D(\nabla y|\sqrt{2}\,b_j|.$$

Similarly, applying Fatou's lemma to $\phi(\nabla y|b_j)$ gives

$$\int_S \left[\phi(\nabla y|\tilde{b}) - T\cdot(\nabla y|\tilde{b})\right] \le \liminf_{j\to\infty} \int_S \left[\phi(\nabla y|b_j) - T\cdot(\nabla y|b_j)\right],$$

and therefore

$$\begin{aligned}\hat{\mathcal{E}}^{(0)}(y,\tilde{b}) &\le \liminf_{j\to\infty} \hat{\mathcal{E}}^{(0)}(y,b_j)\\ &= \inf_{b\in BV_p(S;\mathbb{R}^3)} \hat{\mathcal{E}}^{(0)}(y,\tilde{b}). \qquad \square\end{aligned}$$

Next, we define a functional

$$\hat{\mathcal{E}}_1^{(0)}(u) = \begin{cases} \min\limits_{b\in BV_p(S;\mathbb{R}^3)} \hat{\mathcal{E}}^{(0)}(u_M,b) & \text{if } u_{,3} = 0 \text{ a.e. in } \Omega_1,\\ +\infty & \text{otherwise.}\end{cases}$$

THEOREM 6.4. *The functional* $\hat{\mathcal{E}}_1^{(0)} : \hat{\mathcal{A}}_1 \to \mathbb{R} \cup \{+\infty\}$ *is the* Γ*-limit of the functionals* $\hat{\mathcal{E}}_1^{(h)} : \hat{\mathcal{A}}_1 \to \mathbb{R}$ *as* $h \to 0$ *with respect to the weak* $W^{1,p}(\Omega_1;\mathbb{R}^3)$ *convergence in* $\hat{\mathcal{A}}_1$.

Proof (lower bound). The proof is similar to the proof of Theorem 5.7. Let $u \in \hat{\mathcal{A}}_1$ and let $\{u_h : h > 0\} \subset \hat{\mathcal{A}}_1$ be such that $u_h \rightharpoonup u$ in $W^{1,p}(\Omega_1)$ as $h \to 0$. If $\liminf_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h) = +\infty$, then

$$\hat{\mathcal{E}}_1^{(0)}(u) \le \liminf_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h)$$

is trivially satisfied.

On the other hand, if $\liminf_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h) < +\infty$, then we can first consider a subsequence $\{u_{h_n}\}_{n=1}^{\infty} \subset \hat{\mathcal{A}}_1$ such that

$$\lim_{n\to\infty} \hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) = \liminf_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h).$$

Since then $\hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) \le C$ for all $n \ge 1$, we have by Lemma 6.1 that there exists $b \in BV_p(\Omega_1;\mathbb{R}^3)$ such that for a further subsequence of $\{u_{h_n}\}$, not relabeled, we have that

$$\left.\begin{array}{llllll} u_{h_n} \rightharpoonup u & \text{in } W^{1,p}(\Omega_1;\mathbb{R}^3) & \text{and} & h_n^{-1}u_{h_n,3} \rightharpoonup b & \text{in } L^p(\Omega_1;\mathbb{R}^3) \\ u_{h_n} \to u & \text{in } W^{1,1}(\Omega_1;\mathbb{R}^3) & \text{and} & h_n^{-1}u_{h_n,3} \to b & \text{in } L^1(\Omega_1;\mathbb{R}^3) \end{array}\right\} \quad \text{as } n \to \infty,$$

and the convergence is also almost everywhere in Ω_1. It also follows from Lemma 6.1 that u and b are independent of z_3. Therefore, we have for $u_M(z_1,z_2) = u(z_1,z_2,0)$ and $b_M(z_1,z_2) = b(z_1,z_2,0)$ that

$$\hat{\mathcal{E}}_1^{(0)}(u) \le \hat{\mathcal{E}}^{(0)}(u_M, b_M). \tag{6.2}$$

Using (5.11), the lower semicontinuity of the total variation, and Fatou's lemma to control the ϕ term, we have that

$$\begin{aligned} \hat{\mathcal{E}}^{(0)}(u_M,b_M) &= \kappa \int_{\Omega_1} |D_P(\nabla_P u|\sqrt{2}\, b)| + \int_{\Omega_1} \phi(\nabla_P u|b,\hat{z},0)\,dz - \int_{\Omega_1} T\cdot(\nabla_P u|b) \\ &\le \liminf_{n\to\infty} \hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) \\ &= \lim_{n\to\infty} \hat{\mathcal{E}}_1^{(h_n)}(u_{h_n}) \\ &= \liminf_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h). \end{aligned}$$

Combining the above result with (6.2) completes the first part of the proof.

Upper bound. If $u \in \hat{\mathcal{A}}_1$ is not independent of z_3, then $\hat{\mathcal{E}}_1^{(0)}(u) = +\infty$; taking $u_h = u$ for all $h > 0$ produces a family in $\hat{\mathcal{A}}_1$ such that $\lim_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h) = \hat{\mathcal{E}}_1^{(0)}(u) = +\infty$.

On the other hand, if $u_{,3} = 0$ almost everywhere in Ω_1, then by Lemma 6.3 there exists $\tilde{b} \in BV_p(S;\mathbb{R}^3)$ such that $\hat{\mathcal{E}}_1^{(0)}(u) = \hat{\mathcal{E}}^{(0)}(u_M,\tilde{b})$. Using the upper bound of Theorem 6.2, there exists a family $\{u_h \in \hat{\mathcal{A}}_1 : h > 0\}$ such that

$$u_h \rightharpoonup u \quad \text{in } W^{1,p}(\Omega_1) \text{ as } h \to 0$$

and

$$\lim_{h\to 0} \hat{\mathcal{E}}_1^{(h)}(u_h) = \hat{\mathcal{E}}^{(0)}(u_M, \tilde{b}) = \hat{\mathcal{E}}_1^{(0)}(u). \qquad \square$$

We note that we can obtain a result on the convergence of minimizers of $\mathcal{E}_1^{(h)}$ to minimizers of $\mathcal{E}_1^{(0)}$ by an argument analogous to that of Corollary 5.8.

REFERENCES

[1] E. Acerbi, G. Buttazzo, and D. Percivale, *A variational definition of the strain energy for an elastic string*, J. Elasticity, 25 (1991), pp. 137–148.

[2] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[3] G. Anzellotti, S. Baldo, and D. Percivale, *Dimension reduction in variational problems, asymptotic development in* Γ*-convergence and thin structures in elasticity*, Asymptotic Anal., 9 (1994), pp. 61–100.

[4] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[5] P. Bauman and D. Phillips, *A nonconvex variational problem related to change of phase*, Appl. Math. Optim., 21 (1990), pp. 113–138.

[6] K. Bhattacharya and R. D. James, *A theory of thin films of martensitic materials with applications to microactuators*, J. Mech. Phys. Solids, 47 (1999), pp. 531–576.

[7] A. Braides, Γ*-convergence for Beginners*, Oxford University Press, Oxford, UK, 2002.

[8] P. Bělík, T. Brule, and M. Luskin, *On the numerical modeling of deformations of pressurized martensitic thin films*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 525–548.

[9] P. Bělík and M. Luskin, *A computational model for the indentation and phase transformation of a martensitic thin film*, J. Mech. Phys. Solids, 50 (2002), pp. 1789–1815.

[10] P. Bělík and M. Luskin, *A total-variation surface energy model for thin films of martensitic crystals*, Interfaces Free Bound., 4 (2002), pp. 71–88.

[11] P. Bělík and M. Luskin, *A computational model for martensitic thin films with compositional fluctuation*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1585–1598.

[12] P. Bělík and M. Luskin, *Computational modeling of softening in a structural phase transformation*, Multiscale Model. Simul., 3 (2004), pp. 764–781.

[13] E. Casas, K. Kunisch, and C. Pola, *Regularization by functions of bounded variation and applications to image enhancement*, Appl. Math. Optim., 40 (1999), pp. 229–257.

[14] J. W. Dong, J. Q. Xie, J. Lu, C. Adelmann, C. J. Palmstrøm, J. Cui, Q. Pan, T. W. Shield, R. D. James, and S. McKernan, *Shape memory and ferromagnetic shape memory effects in single-crystal* Ni_2MnGa *thin films*, J. Appl. Phys., 95 (2004), pp. 2593–2600.

[15] H. Le Dret and A. Raoult, *The nonlinear membrane model as variational limit of nonlinear three-dimensional elasticity*, J. Math. Pures Appl. (9), 73 (1995), pp. 549–578.

[16] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[17] I. Fonseca and G. Francfort, $3D$*-*$2D$ *asymptotic analysis of an optimal design problem for thin films*, J. Reine Angew. Math, 505 (1998), pp. 173–202.

[18] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Verlag, Basel, 1984.

[19] S. Kartha, J. A. Krumhansl, J. P. Sethna, and L. K. Wickham, *Disorder-driven pretransitional tweed in martensitic transformations*, Phys. Rev. B, 52 (1995), pp. 803–822.

[20] P. Krulevitch, A. P. Lee, P. B. Ramsey, J. C. Trevino, J. Hamilton, and M. A. Northrup, *Thin film shape memory microactuators*, J. MEMS, 5 (1996), pp. 270–282.

[21] M. Kružík, *Numerical approach to double well problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1833–1849.

[22] B. Li, *Finite element analysis of a class of stress-free martensitic microstructures*, Math. Comp., 72 (2003), pp. 1675–1688 (electronic).

[23] M. Luskin, *On the computation of crystalline microstructure*, Acta Numer., 5 (1996), pp. 191–257.

[24] G. Dal Maso, *An introduction to* Γ*-convergence*, Birkhäuser Boston, Boston, 1993.

[25] L. Modica and S. Mortola, *Il limite nella* Γ*-convergenza di una famiglia di funzionali ellittici*, Boll. Un. Mat. Ital. A (5), 14 (1977), pp. 526–529.

[26] P. Pedregal, *Variational Methods in Nonlinear Elasticity*, SIAM, Philadelphia, 2000.
[27] M. Pitteri and G. Zanzotto, *Continuum models for phase transitions and twinning in crystals*, Chapman and Hall/CRC, Boca Raton, FL, 2003.
[28] Y. C. Shu, *Heterogeneous thin films of martensitic materials*, Arch. Ration. Mech. Anal., 153 (2000), pp. 39–90.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 434–451

STABLE DETERMINATION OF THE SURFACE IMPEDANCE OF AN OBSTACLE BY FAR FIELD MEASUREMENTS*

E. SINCICH†

Abstract. We deal with the inverse scattering problem of determining the surface impedance of a partially coated obstacle. We prove a stability estimate of logarithmic type for the impedance term by the far field measurements.

Key words. inverse scattering problem, impedance boundary condition, stability

AMS subject classifications. 35J05, 35R30, 35R25, 31B20

DOI. 10.1137/050631513

1. Introduction.

We consider the scattering of an acoustic incident time-harmonic plane wave, at a given wave number $k > 0$ and at a given incident direction $\omega \in \mathbb{S}^2$, by an obstacle $D \subset \mathbb{R}^3$ partially coated by a material with surface impedance λ. Such a problem is modeled by the following mixed boundary value problem for the Helmholtz equation:

$$\begin{cases} \Delta u + k^2 u = 0 & \text{in } \mathbb{R}^3 \setminus \overline{D}, \\ u = 0 & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial \nu} + i\lambda(x) u = 0 & \text{on } \Gamma_I, \end{cases} \tag{1.1}$$

where $u = u^s + \exp(ikx \cdot \omega)$ is the total field, which is given as the sum of the scattered wave u^s and the incident plane waves $\exp(ikx \cdot \omega)$, and where $\Gamma_I,\ \Gamma_D$ are two open and connected portions of the boundary ∂D such that $\partial D = \overline{\Gamma_I \cup \Gamma_D}$.

Moreover, the scattered field u^s is required to satisfy the so-called *Sommerfeld radiation condition*

$$\lim_{r \to \infty} r \left(\frac{\partial u^s}{\partial r} - iku^s \right) = 0, \quad r = \|x\|. \tag{1.2}$$

It is well known that the scattered field u^s has the following asymptotic behavior:

$$u^s(x) = \frac{\exp(ikr)}{r} \left\{ u_\infty(\hat{x}) + O\left(\frac{1}{r}\right) \right\} \tag{1.3}$$

as r tends to ∞, uniformly with respect to $\hat{x} = \frac{x}{\|x\|}$ and where u_∞ is the so-called far field pattern of the scattered wave (see, for instance, [11]).

The *inverse scattering problem* that we examine here consists in the determination of the surface impedance $\lambda(x)$ by the knowledge of the far field pattern, provided some suitable a priori assumptions on the impedance are made.

Such a problem, in two dimensions, has been recently studied by Cakoni and Colton in [7]. The authors have provided a variational method for the determination of the essential supremum of the surface impedance when the far field data are available.

*Received by the editors May 13, 2005; accepted for publication (in revised form) January 10, 2006; published electronically May 12, 2006. This work was supported in part by MIUR grant 2004011204.

http://www.siam.org/journals/sima/38-2/63151.html

†INRIA, Project Apics, 06902 Sophia-Antipolis Cedex, France (esincich@sophia.inria.fr).

In this paper, we shall deal with the stability issue; namely, we will prove a stability estimate of logarithmic type for the surface impedance by the far field measurements under some a priori mild assumptions on the impedance itself.

Let us point out that a stability result for this type of problem has been proved in [15] by Labreuche under the assumption of an analytic boundary. The new feature of the present paper consists in a reduced assumption on the regularity of the boundary; namely, we shall assume that Γ_I is a $C^{1,1}$ portion of ∂D. Thus it turns out that the argument of analytic extension used in [15] cannot be applied.

The stable recovering of the surface impedance needs some a priori mild assumptions on the impedance itself. The additional a priori information that we require on the unknown surface impedance λ is an a priori bound on its Lipschitz continuity; that is, we assume that for a given positive constant Λ, the following holds:

$$\|\lambda\|_{C^{0,1}(\Gamma_I)} \leqslant \Lambda. \tag{1.4}$$

Moreover, we prescribe the following uniform lower bound:

$$\lambda(x) \geqslant \lambda_0 \quad \text{for every } x \in \Gamma_I, \tag{1.5}$$

where λ_0 is a given positive constant.

In order to treat the inverse scattering problem we first need to analyze the *direct* one. In section 3, indeed, following the arguments of potential theory developed in [8], we observe that the direct scattering problem is well posed (see Lemma 3.1). The proof relies on the fact that the mixed boundary value problem (1.1) can be reformulated as a system of boundary integral equations. Moreover, we prove (see Theorem 3.2) that the solution and its first order derivatives are Hölder continuous in a neighborhood of the portion Γ_I, where the impedance takes place. The proof is based on Moser's iteration technique. Finally in Corollary 3.3, we obtain a uniform lower bound for the total field u on sets away from the obstacle.

In section 4, we deal with the inverse scattering problem. The underlying ideas and the main tools that lead to the stability result can be outlined as follows:

(i) As the first step we evaluate how much the error on the far field can affect the values of the field near the scatterer.

(ii) In the second step we are concerned with a stability estimate of the field at the boundary in terms of the near field.

(iii) Finally, as the last step, we obtain a stability result for the impedance λ by the estimate of the field at the boundary.

Let us start the analysis of section 4 by illustrating the arguments introduced in step (iii) of the above list.

By the impedance condition in (1.1) we can formally compute λ as

$$\lambda(x) = \frac{i}{u(x)} \frac{\partial u(x)}{\partial \nu(x)}. \tag{1.6}$$

Since u may vanish in some points of Γ_I, it follows that the quotient in (1.6) may be undetermined. In this respect, we found it necessary to evaluate the local vanishing rate of the solution on the boundary. To establish such a control we shall make use of quantitative estimates of unique continuation. We first obtain, in Lemma 4.5, a *volume doubling inequality* at the boundary, namely

$$\int_{\Gamma_{I,2\rho}(x_0)} |u|^2 \leqslant \text{const.} \int_{\Gamma_{I,\rho}(x_0)} |u|^2, \tag{1.7}$$

where $\Gamma_{I,\rho}(x_0)$ and $\Gamma_{I,2\rho}(x_0)$ are the portions of the balls centered at the boundary point x_0 of radius ρ and 2ρ, respectively, contained in $\mathbb{R}^3 \setminus \overline{D}$ (see (2.10) for a precise definition).

In order to obtain the formula in (1.7), we have adapted the arguments developed in [2] for the more general setting of complex valued solutions which is required by the boundary value problem (1.1).

A further difficulty in dealing with such arguments is due to the fact that the techniques used in [2] apply to a homogeneous Neumann boundary condition. We overcome such a difficulty by performing a suitable change of the independent variable (see Proposition 4.3) that fits our problem under the assumptions required in [2]. Moreover, well-known stability estimates for the Cauchy problem [18] allow us to reformulate the *volume* doubling inequality at the boundary, deriving in Theorem 4.6 a new one on the boundary, that is, a *surface doubling inequality*

$$\int_{\Delta_{I,2\rho}(x_0)} |u|^2 \leqslant \text{const.} \int_{\Delta_{I,\rho}(x_0)} |u|^2, \tag{1.8}$$

where $\Delta_{I,\rho}(x_0)$ and $\Delta_{I,2\rho}(x_0)$ are the portions of the boundary of $\Gamma_{I,\rho}(x_0)$ and $\Gamma_{I,2\rho}(x_0)$, respectively, which have nonempty intersection with ∂D (see (2.11) for a precise definition).

The surface doubling inequality allows us to apply the theory of *Muckenhoupt weights* [9], which, in particular, implies the existence of some exponent $p > 1$ such that $|u|^{-\frac{2}{p-1}}$ is integrable on an inner portion of Γ_I; see Corollary 4.7. This integrability property, as well as the Hölder continuity of the normal derivative, justifies the computation made in (1.6) in the $L^{\frac{2}{p-1}}$ sense.

Let us carry over our analysis by discussing the evaluation introduced in step (i). Such an evaluation, introduced by Isakov [13, 14] and then developed by Bushuyev [6], concerns a stability estimate for the *near field* in terms of the measurements of the *far field* (see Lemma 4.1). It means that if u_1 and u_2 are two acoustic fields corresponding to impedances λ_1 and λ_2 such that their scattering amplitudes, $u_{1,\infty}$ and $u_{2,\infty}$, respectively, are close, i.e.,

$$\|u_{1,\infty} - u_{2,\infty}\|_{L^2(\partial B_1(0))} \leqslant \varepsilon, \tag{1.9}$$

then u_1 and u_2 satisfy

$$\|u_1 - u_2\|_{L^2(B_{R_1+1}(0)\setminus B_{R_1}(0))} \leqslant \text{const. } \varepsilon^{\alpha(\varepsilon)}, \tag{1.10}$$

where $R_1 > 0$ is a suitable radius such that $B_{R_1}(0) \supset \overline{D}$ and $\alpha(\varepsilon)$ is the function introduced in (4.2).

As the last step of this treatment we provide the stability estimate introduced in (ii). The proof is based on arguments of quantitative unique continuation, as the *three spheres inequality*, and leads to the following estimate:

$$\|u_1 - u_2\|_{C^1(\Gamma_I^\rho)} \leqslant \text{const.} |\log (\|u_1 - u_2\|^{-1}_{L^2(B_{R_1+1}(0)\setminus B_{R_1}(0))})|^{-2\theta}, \tag{1.11}$$

where $\theta > 0$ and Γ_I^ρ is a given inner portion of Γ_I (see (2.9) for a precise definition).

By combining the stability estimates listed in (i) and (ii), we obtain a stability result for the total field at the boundary in terms of the measurements of the far field (see Theorem 4.2).

Finally, as a consequence of Theorem 4.2 and Corollary 4.7, let us formulate the main result of the present paper, which consists in a stability estimate of the surface impedance by the far field measurements (see Theorem 2.1). Assuming that (1.9) holds, we have shown that the impedances λ_1, λ_2 agree up to an error

$$\big|\log(\varepsilon)\big|^{-\theta}. \tag{1.12}$$

2. Main assumptions and results.

2.1. Main hypothesis and notation. Assumptions on the domain. We shall assume throughout that D is a bounded domain in $\mathbb{R}^3$, such that $\mathrm{diam} D \leqslant d$, with Lipschitz boundary ∂D with constants r_0, M. More precisely, for every $x_0 \in \partial D$, there exists a rigid transformation of coordinates under which

$$D \cap B_{r_0}(x_0) = \{(x', x_3) : x_3 > \gamma(x')\}, \tag{2.1}$$

where $x \in \mathbb{R}^3$, and $x = (x', x_3)$, with $x' \in \mathbb{R}^2$, $x_3 \in \mathbb{R}$, and

$$\gamma : B'_{r_0}(x_0) \subset \mathbb{R}^2 \to \mathbb{R}$$

satisfying $\gamma(0) = 0$ and

$$\|\gamma\|_{C^{0,1}(B'_{r_0}(x_0))} \leqslant M r_0,$$

where for every integer $k \geqslant 1$ and for every multi-index $0 \leqslant |\beta| \leqslant k$ we denote

$$\begin{aligned} \|\gamma\|_{C^{k,1}(B'_{r_0}(z_0))} = \sum_{j=0}^{k} {r_0}^j \sum_{|\beta|=j} \|D^\beta \gamma\|_{L^\infty(B'_{r_0}(z_0))} \\ + {r_0}^{k+1} \sum_{|\beta|=k} \sup_{\substack{x,y \in B'_{r_0}(z_0) \\ x \neq y}} \frac{|D^\beta \gamma(x) - D^\beta \gamma(y)|}{|x-y|} \end{aligned} \tag{2.2}$$

and $B'_{r_0}(x_0)$ denotes a ball in $\mathbb{R}^2$. Moreover, we assume that the portion of the boundary Γ_I is contained in a surface S_I, which is $C^{1,1}$ smooth with constants r_0, M. More precisely, for any $x_0 \in S_I$, we have that up to a rigid change of coordinates,

$$S_I \cap B_{r_0}(x_0) = \{(x', x_3) : x_3 = \varphi_I(x')\}, \tag{2.3}$$

where

$$\varphi_I : B'_{r_0}(z_0) \subset \mathbb{R}^2 \to \mathbb{R} \tag{2.4}$$

is a $C^{1,1}$ function satisfying $\varphi_I(0) = |\nabla \varphi_I(0)| = 0$ and

$$\|\varphi_I\|_{C^{1,1}(B'_{r_0}(z_0))} \leqslant M r_0. \tag{2.5}$$

For the sake of simplicity we shall assume that $0 \in D$.

Fixing $R > d$, $\rho \in (0, r_0)$, and $x_0 \in \Gamma_I$, let us define the following sets:

$$D^+ = \mathbb{R}^3 \setminus \overline{D}, \tag{2.6}$$

$$D_R^+ = B_R(0) \cap D^+, \tag{2.7}$$

$$D_{R,\rho}^+ = \{x \in \overline{D_R^+} \ : \mathrm{dist}(x, \Gamma_D) > \rho\}, \tag{2.8}$$

$$\Gamma_I^\rho = \partial D_{R,\rho}^+ \cap \Gamma_I, \tag{2.9}$$

$$\Gamma_{I,\rho}(x_0) = B_\rho(x_0) \setminus \overline{D}, \tag{2.10}$$

$$\Delta_{I,\rho}(x_0) = \overline{\Gamma_{I,\rho}(x_0)} \cap \partial D. \tag{2.11}$$

A priori information on the impedance term. We assume that the impedance coefficient λ belongs to $C^{0,1}(\Gamma_I, \mathbb{R})$ and is such that (1.5) and (1.4) hold for given positive constants λ_0, Λ.

From now on we shall refer to the *a priori data* as the following set of quantities: $d, r_0, M, \lambda_0, \Lambda, k, \omega$.

In what follows we shall denote with $\eta(t)$ a positive increasing function defined on $(0, +\infty)$, which satisfies

$$\eta(t) \leqslant C|\log(t)|^{-\theta} \quad \text{for every} \quad 0 < t < 1, \tag{2.12}$$

where $C > 0$, $\theta > 0$ are constants depending on the a priori data only.

2.2. The main result.

THEOREM 2.1 (stability for λ). *Let u_i, $i = 1, 2$, be the weak solutions to the problem (1.1) with $\lambda = \lambda_i$ and $u_\infty = u_{i,\infty}$, respectively. There exists a constant $\varepsilon_0 > 0$ depending on the a priori data only, such that if for some ε, $0 < \varepsilon < \varepsilon_0$, we have*

$$\|u_{1,\infty} - u_{2,\infty}\|_{L^2(\partial B_1(0))} \leqslant \varepsilon, \tag{2.13}$$

then

$$\|\lambda_1 - \lambda_2\|_{L^\infty(\Gamma_I^{r_0})} \leqslant \eta(\varepsilon). \tag{2.14}$$

3. The direct scattering problem. Let us introduce the space

$$H^1_{\mathrm{loc}}(D^+) = \{v \in D^*(D^+) : \, v|_{D_R^+} \in H^1(D_R^+) \quad \text{for every } R > 0 \text{ such that } \overline{D} \subset B_R(0)\},$$

where $D^*(D^+)$ is the space of distribution on D^+.

A weak solution to the problem (1.1) is a function $u = \exp(ik\omega \cdot x) + u^s$, where $u^s \in H^1_{\mathrm{loc}}(D^+)$ is a weak solution to the problem

$$\begin{cases} \Delta u^s + k^2 u^s = 0 & \text{in } D^+, \\ u^s = -\exp(ik\omega \cdot x) & \text{on } \Gamma_D, \\ \dfrac{\partial u^s}{\partial \nu} + i\lambda(x) u^s = -\dfrac{\partial}{\partial \nu} \exp(ik\omega \cdot x) - i\lambda(x) \exp(ik\omega \cdot x) & \text{on } \Gamma_I, \\ \lim\limits_{r \to \infty^r} \left(\dfrac{\partial u^s}{\partial r}(r\hat{x}) - iku^s(r\hat{x}) \right) = 0 & \text{uniformly in } \hat{x}, \end{cases} \tag{3.1}$$

where ν is the inward unit normal to D.

LEMMA 3.1 (well-posedness). *The problem (3.1) has one and only one weak solution u^s. Moreover, for every $R > d$, there exists a constant $C_R > 0$ depending on the a priori data and R only, such that the following holds:*

$$\|u^s\|_{H^1(D_R^+)} \leqslant C_R. \tag{3.2}$$

Proof. For the proof we refer to [8, Theorem 2.5], in which the authors show, among various results, that the exterior mixed boundary value problem (3.1) can be reformulated as a 2×2 system of boundary integral equations. In [8], Theorem 2.5 has been proved in two dimensions for a constant λ; however, it can be verified that the same techniques can be carried over in three dimensions (see, for instance, [10]) and with $\lambda = \lambda(x) \in C^{0,1}(\Gamma_I)$. □

THEOREM 3.2 ($C^{1,\alpha}$ regularity at the boundary). *Let u be the weak solution to* (1.1)*; then there exists a constant $\alpha,\ 0 < \alpha < 1$, such that for every $R > d$ and $\rho \in (0, r_0)$, $u \in C^{1,\alpha}(D^+_{R,\rho})$. Moreover, there exists a constant $C_{R,\rho} > 0$ depending on the a priori data, R, and ρ only, such that*

$$\|u\|_{C^{1,\alpha}(D^+_{R,\rho})} \leqslant C_{R,\rho}. \tag{3.3}$$

Proof. The proof mainly relies on the standard iteration techniques due to Moser (see, for instance, [12]) as well as well-known regularity bounds for the Neumann problem (see, for instance, [3, p. 667]). For the details of the proof we refer the reader to [17]. □

COROLLARY 3.3 (lower bound). *Let u be the weak solution to* (1.1)*; then there exists a radius $R_0 > 0$ depending on the a priori data only, such that*

$$|u(x)| > \frac{1}{2} \quad \textit{for every } x,\ |x| > R_0. \tag{3.4}$$

Proof. Let us choose $R = 4d + 4r_0$. By Theorem 3.2 it follows that there exists a constant $C > 0$ depending on the a priori data only, such that

$$\|u\|_{C^{1,\alpha}\left(D^+_{2R,\frac{r_0}{2}}\right)} \leqslant C. \tag{3.5}$$

By Green's formula for the scattered wave u^s (see, for instance, [11, p. 18]), we have that

$$u^s(x) = \int_{\partial B_R(0)} \left(u^s(y) \frac{\partial \phi(x,y)}{\partial \nu(y)} - \frac{\partial u^s(y)}{\partial \nu(y)} \phi(x,y) \right) \mathrm{d}s(y), \qquad |x| > R, \tag{3.6}$$

where

$$\phi(x,y) = \frac{1}{4\pi} \frac{\exp\left(ik|x-y|\right)}{|x-y|}, \quad x \neq y,$$

is the fundamental solution to the Helmholtz equation in $\mathbb{R}^3$.

Thus, by (3.6) and by (3.5) it follows that

$$|u^s| < \frac{1}{2} \quad \text{for every } x,\ |x| > R_0, \tag{3.7}$$

with $R_0 = (k+1)8R^3C + 2R$. The thesis follows, observing that $|u| \geqslant 1 - |u^s|$. □

4. The inverse scattering problem.

LEMMA 4.1 (from the far field to the near field). *Let $u_i, u_{i,\infty},\ i = 1, 2$, be as in Theorem* 2.1*. Suppose that for some $\varepsilon,\ 0 < \varepsilon < 1$,* (2.13) *holds; then there exist a radius $R_1 > 0$ and a constant $C > 0$, depending on the a priori data only, such that*

$$\|u_1 - u_2\|_{L^2(B_{R_1+1}(0) \setminus B_{R_1}(0))} \leqslant C\varepsilon^{\alpha(\varepsilon)}, \tag{4.1}$$

where $\alpha(\varepsilon)$ is defined as

$$\alpha(\varepsilon) = \frac{1}{1 + \log(\log(\varepsilon^{-1}) + e)}. \tag{4.2}$$

Proof. Let us choose $R = 4d + 4r_0$ and let us denote by $u^s_i,\ i = 1, 2$, the scattered wave of the problem (1.1) with $\lambda = \lambda_i$, respectively. By (3.5) it follows that

$$\|u^s_1 - u^s_2\|_{L^2(\partial B_R(0))} \leqslant C, \tag{4.3}$$

where $C > 0$ is a constant depending on the a priori data only.

Hence choosing $R_1 = 16d + 16r_0$, the thesis follows by the argument in [14] (see also [6]). □

THEOREM 4.2 (stability at the boundary). *Let $u_i, u_{i,\infty},\ i = 1, 2$, be as in Theorem 2.1. We have that there exists $\varepsilon_0 > 0$ depending on the a priori data only, such that if for some $\varepsilon,\ 0 < \varepsilon < \varepsilon_0$, (2.13) holds, then for every $\rho \in (0, r_0)$ we have*

$$\|u_1 - u_2\|_{C^1(\Gamma_I^\rho)} \leqslant \eta(\varepsilon), \tag{4.4}$$

where η is given by (2.12), with a constant $C > 0$ depending on the a priori data and ρ only.

Proof. By the Lipschitz regularity of the boundary ∂D, it follows that for every point $Q \in \partial D$, there exists a rigid transformation of coordinates under which we have $Q = 0$ and the finite cone

$$\mathcal{C} = \left\{ x : |x| < r_0,\ \frac{x \cdot \xi}{|x|} > \cos\theta \right\}$$

with axis in the direction ξ and width 2θ, where $\theta = \arctan \frac{1}{M}$, is such that $\mathcal{C} \subset D^+$.

Let Q be a point such that $Q \in \Gamma_I^{r_0}$ and let Q_0 be a point lying on the axis ξ of the cone with vertex in $Q = 0$ such that $d_0 = \operatorname{dist}(Q_0, 0) < \frac{r_0}{2}$.

Let us define $R_2 = 2R_1 + 2$, where R_1 is the radius introduced in the statement of Lemma 4.1. Proceeding as in Lieberman [16], we consider a regularized distance $\tilde{d}$ from the boundary of ∂D such that $\tilde{d} \in C^2(D^+_{R_2}) \cap C^{0,1}(\overline{D^+_{R_2}})$ (see also [4, Lemma 5.2]).

Let us define for every $\rho > 0$

$$D^\rho = \{x \in D^+_{R_2} : \operatorname{dist}(x, \partial D) > \rho\}, \tag{4.5}$$

$$\tilde{D}^\rho = \{x \in D^+_{R_2} : \tilde{d}(x) > \rho\}. \tag{4.6}$$

It follows that there exists a, $0 < a \leqslant 1$, depending on M only such that for every ρ, $0 < \rho \leqslant a r_0$, $\tilde{D}^\rho$ is connected with boundary of class C^1 and

$$\tilde{c}_1 \rho \leqslant \operatorname{dist}(x, \partial D) \leqslant \tilde{c}_2 \rho \quad \text{for every } \ x \in \partial \tilde{D}^\rho, \tag{4.7}$$

where $\tilde{c}_1, \tilde{c}_2$ are positive constants depending on M only. By (4.7) we deduce that

$$D^{\tilde{c}_2 \rho} \subset \tilde{D}^\rho \subset D^{\tilde{c}_1 \rho}.$$

Let us now define $\rho_0 = \min\{\frac{1}{16}, \frac{r_0}{4}\sin\theta\}$ and let P be a point in the annulus $B_{R_1+1}(0) \setminus B_{R_1}(0)$, such that $B_{4\rho_0}(P) \subset B_{R_1+1}(0) \setminus B_{R_1}(0)$. Furthermore, let γ be a path in $\tilde{D}^{\frac{\rho_0}{\tilde{c}_1}}$ joining P to Q_0 and let us define $\{y_i\}$, $i = 0, \dots, s$, as follows: $y_0 = Q_0$, $y_{i+1} = \gamma(t_i)$, where $t_i = \max\{t \text{ such that } |\gamma(t) - y_i| = 2\rho_0\}$ if $|P - y_i| > 2\rho_0$; otherwise let $i = s$ and stop the process.

Let us introduce the function $U \in H^1_{\mathrm{loc}}(D^+)$ defined as

$$U(x) = u_1(x) - u_2(x). \tag{4.8}$$

Thus, by the three spheres inequality for elliptic systems with Laplacian principal part (see [5, Theorem 3.1]), we infer that

$$\int_{B_{3\rho_0}(y_0)} |U|^2 \leqslant C \left(\int_{B_{\rho_0}(y_0)} |U|^2 \right)^\tau \cdot \left(\int_{B_{4\rho_0}(y_0)} |U|^2 \right)^{1-\tau}, \tag{4.9}$$

where $C > 0, 0 < \tau < 1$ are constants depending on the a priori data only.

Let us observe that $B_{4\rho_0}(y_0) \subset D^+_{R_2}$ and $B_{\rho_0}(y_0) \subset B_{3\rho_0}(y_1)$. Thus by (4.9) and by Lemma 3.1 we deduce that

$$\int_{B_{\rho_0}(y_0)} |U|^2 \leqslant C \left(\int_{B_{3\rho_0}(y_1)} |U|^2 \right)^{\tau} \cdot C^{1-\tau}.$$

An iterated application of the three spheres inequality leads to

$$\int_{B_{\rho_0}(y_0)} |U|^2 \leqslant \left(\int_{B_{\rho_0}(y_s)} |U|^2 \right)^{\tau^s} \cdot C^{1-\tau^s}.$$

Finally, since $B_\rho(y_s) \subset B_{R_1+1}(0) \setminus B_{R_1}(0)$, by (4.1) we obtain that

$$\int_{B_{\rho_0}(y_0)} |U|^2 \leqslant C\big\{\varepsilon^{\alpha(\varepsilon)}\big\}^{\tau^s}.$$

We shall construct a chain of balls $B_{\rho_k}(Q_k)$ centered on the axis of the cone, pairwise tangent to one another and all contained in the cone

$$\mathcal{C}' = \left\{ x : |x| < r_0,\ \frac{x \cdot \xi}{|x|} > \cos\theta' \right\},$$

where $\theta' = \arcsin\left(\frac{\rho_0}{d_0}\right)$. Let $B_{\rho_0}(Q_0)$ be the first of them; the rest are defined by induction such that

$$\begin{aligned} Q_{k+1} &= Q_k - (1+\mu)\rho_k \xi, \\ \rho_{k+1} &= \mu \rho_k, \\ d_{k+1} &= \mu d_k, \\ \mu &= \frac{1-\sin\theta'}{1+\sin\theta'}. \end{aligned}$$

Hence, with this choice, we have $\rho_k = \mu^k \rho_0$ and $B_{\rho_{k+1}}(Q_{k+1}) \subset B_{3\rho_k}(Q_k)$.

Considering the following estimate obtained by a repeated application of the three spheres inequality, we have that

$$\begin{aligned} \|U\|_{L^2(B_{\rho_k}(Q_k))} &\leqslant \|U\|^{\tau}_{L^2(B_{\rho_{k-1}}(Q_{k-1}))} \|U\|^{1-\tau}_{L^2(B_{4\rho_{l-1}}(Q_{k-1}))} \\ &\leqslant C\|U\|^{\tau^k}_{L^2(B_{\rho_0}(Q_0))} \leqslant C\Big\{ \big[\varepsilon^{\alpha(\varepsilon)}\big]^{\tau^s} \Big\}^{\tau^k}. \end{aligned} \tag{4.10}$$

For every r, $0 < r < d_0$, let $k(r)$ be the smallest positive integer such that $d_k \leqslant r$; then, since $d_k = \mu^k d_0$, it follows that

$$\frac{|\log(\frac{r}{d_0})|}{\log\mu} \leqslant k(r) \leqslant \frac{|\log(\frac{r}{d_0})|}{\log\mu} + 1, \tag{4.11}$$

and by (4.10) we deduce that

$$\|U\|_{L^2(B_{\rho_{k(r)}}(Q_k(r)))} \leqslant C\Big\{ \big[\varepsilon^{\alpha(\varepsilon)}\big]^{\tau^s} \Big\}^{\tau^{k(r)}}. \tag{4.12}$$

Let $\bar{x} \in \Gamma_I^{\frac{\rho}{2}}$ with $\rho \in (0, r_0)$ and let $x \in B_{\frac{\rho_{k(r)-1}}{2}}(Q_{k(r)-1})$. By Theorem 3.2 it follows that

$$|U(\bar{x})| \leqslant |U(x)| + C_\rho |x - \bar{x}|^\alpha \leqslant |U(x)| + C_\rho \left(\frac{2}{\mu} r\right)^\alpha,$$

where $C_\rho > 0$ is a constant depending on the a priori data and ρ only.

Integrating this inequality over $B_{\frac{\rho_{k(r)-1}}{2}}(Q_{k(r)-1})$, we have that

$$|U(\bar{x})|^2 \leqslant \frac{2}{\omega_3 \left(\frac{\rho_{k-1}}{2}\right)^3} \int_{B_{\frac{\rho_{k(r)-1}}{2}}\left(Q_{k(r)-1}\right)} |U(x)|^2 \mathrm{d}x + 2C_\rho^2 \left(\frac{4r^2}{\mu^2}\right)^\alpha. \tag{4.13}$$

Since $k(r)$ is the smallest integer such that $d_k \leqslant r$, then $d_{k-1} > r$ and thus (4.13) yields

$$|U(\bar{x})|^2 \leqslant \frac{C}{\left(r \sin \theta'\right)^3} \int_{B_{\rho_{k(r)-1}}(Q_{k(r)-1})} |U(x)|^2 \mathrm{d}x + C_\rho r^{2\alpha}.$$

By (4.12) we deduce that

$$|U(\bar{x})|^2 \leqslant \frac{C}{r^3} \left\{ \left[\varepsilon^{\alpha(\varepsilon)}\right]^{\tau^s} \right\}^{\tau^{k(r)-1}} + C_\rho r^{2\alpha}. \tag{4.14}$$

Moreover, Theorem 3.2 also provides that

$$\left|\frac{\partial U(\bar{x})}{\partial \nu}\right| \leqslant \left|\frac{\partial U(x)}{\partial \nu}\right| + C_\rho \left(\frac{2}{\mu} r\right)^\alpha.$$

The Caccioppoli inequality and the same arguments that lead to (4.14) yield

$$\left|\frac{\partial U(\bar{x})}{\partial \nu}\right|^2 \leqslant \frac{C}{r^5} \left\{ \left[\varepsilon^{\alpha(\varepsilon)}\right]^{\tau^s} \right\}^{\tau^{k(r)-1}} + C_\rho r^{2\alpha}. \tag{4.15}$$

The choice in (4.11) guarantees that

$$\tau^{k(r)-1} \geqslant \left(\frac{r}{d_0}\right)^\nu, \tag{4.16}$$

where $\nu = -\log\left(\frac{1}{\mu}\right) \log \tau$. Hence inserting (4.16) into (4.14) and (4.15) and minimizing their right-hand sides with respect to r, we deduce that

$$\|U(\bar{x})\|_{L^\infty(\Gamma_I^{\frac{\rho}{2}})} \leqslant C_\rho \left(\log\left(\varepsilon^{-\alpha(\varepsilon)}\right)\right)^{-\frac{2\alpha}{\nu+2}}, \tag{4.17}$$

$$\left\|\frac{\partial U(\bar{x})}{\partial \nu}\right\|_{L^\infty(\Gamma_I^{\frac{\rho}{2}})} \leqslant C_\rho \left(\log\left(\varepsilon^{-\alpha(\varepsilon)}\right)\right)^{-\frac{2\alpha}{\nu+2}}, \tag{4.18}$$

where $C_\rho > 0$ is a constant depending on the a priori data and ρ only.

By an interpolation inequality and (3.2) we have

$$\|\nabla_t(U)\|_{L^\infty(\Gamma_{1,\rho})} \leqslant c_\rho \|U\|^\beta_{L^\infty(\Gamma_{1,\frac{\rho}{2}})} C_\rho^{\,1-\beta},$$

where $\beta = \frac{\alpha}{\alpha+1}$ and $c_\rho > 0$ depends on the a priori data and ρ only. After straightforward calculations and by a possible replacing of ε_0 with a smaller one depending on the a priori data only, we have that

$$\|u_1 - u_2\|_{C^1(\Gamma_{1,\rho})} \leq C_\rho |\log(\varepsilon)|^{-\frac{2\alpha\beta}{\nu+2}} \quad \text{for every } \varepsilon,\ 0 < \varepsilon < \varepsilon_0. \tag{4.19}$$

Thus the thesis follows, replacing in (2.12) C with C_ρ and θ with $\frac{2\alpha\beta}{\nu+2}$. □

PROPOSITION 4.3. *There exists a radius $r_1 > 0$ depending on the a priori data only, such that for every $x_0 \in \Gamma_I^{r_0}$, the problem*

$$\begin{cases} \Delta\psi + k^2\psi = 0 & \text{in } \Gamma_{I,r_1}(x_0), \\ \dfrac{\partial\psi}{\partial\nu} + i\lambda(x)\psi = 0 & \text{on } \Delta_{I,r_1}(x_0) \end{cases} \tag{4.20}$$

admits a solution $\psi \in H^1(\Gamma_{I,r_1}(x_0))$ satisfying

$$|\psi(x)| \geqslant 1 \text{ for every } x \in \Gamma_{I,r_1}(x_0). \tag{4.21}$$

Moreover, there exists a constant $\bar{\psi} > 0$ depending on the a priori data only, such that for every $x_0 \in \Gamma_I^{r_0}$,

$$\|\psi\|_{C^1(\Gamma_{I,r_1}(x_0))} \leqslant \bar{\psi}. \tag{4.22}$$

Proof. Let us consider a point $x_0 \in \Gamma_I^{r_0}$. After a translation we may assume that $x_0 = 0$ and, fixing local coordinates, we can represent the boundary as a graph of a $C^{1,1}$ function. Namely, we have that

$$D^+ \cap B_{r_0}(0) = \{(x', x_3) \in B_{r_0}(0) : x_3 < \varphi_I(x')\}, \tag{4.23}$$

where φ_I is the $C^{1,1}$ function satisfying (2.4) and (2.5).

Let $\Phi \in C^{1,1}(B_{\frac{r_0}{4M}}, \mathbb{R}^3)$ be the map defined as

$$\Phi(y', y_3) = (y', y_3 + \varphi_I(y')). \tag{4.24}$$

We have that there exist θ_1, θ_2, $\theta_1 > 1 > \theta_2 > 0$, constants depending on M and r_0 only, such that for every $r \in (0, \frac{r_0}{4M})$, it follows that

$$\Gamma_{I,\theta_2 r}(0) \subset \Phi(B_r^-(0)) \subset \Gamma_{I,\theta_1 r}(0), \tag{4.25}$$

where $B_r^-(0) = \{y \in \mathbb{R}^3 : |y| < r,\ y_3 < 0\}$, and furthermore we have

$$|\det D\Phi| = 1. \tag{4.26}$$

The inverse map $\Phi^{-1} \in C^{1,1}(\Gamma_{I,r_0}(0), \mathbb{R}^3)$ and is defined by

$$\Phi^{-1}(x', x_3) = (x', x_3 - \varphi_I(x')). \tag{4.27}$$

Denoting

$$\sigma(y) = (\sigma_{i,j}(y))_{i,j=1}^3 = (D\Phi^{-1})(\Phi(y)) \cdot (D\Phi^{-1})^T(\Phi(y)), \tag{4.28}$$

$$\lambda'(y) = \lambda(\Phi(y)), \tag{4.29}$$

$$\lambda_0{}' = \lambda'(0), \tag{4.30}$$

where λ is the function introduced in (1.5), we have that

$$\sigma(0) = \mathrm{I}, \tag{4.31}$$

$$\|\sigma_{i,j}\|_{C^{0,1}(\Gamma_{I,r_0})} \leqslant \Sigma \quad \text{for} \quad i,j = 1,2,3, \tag{4.32}$$

$$\frac{1}{2}|\xi|^2 \leqslant \sigma(y)\xi\cdot\xi \leqslant C_1|\xi|^2 \text{ for every } y \in B^-_{(\frac{r_0}{4M})}(0) \text{ and every } \xi \in \mathbb{R}^3, \tag{4.33}$$

$$\|\lambda'\|_{C^{0,1}(B'_{\frac{r_0}{4M}}(0))} \leqslant \Lambda', \tag{4.34}$$

where $\Sigma > 0, C_1 > 0$, and $\Lambda' > 0$ are constants depending on M, r_0, and Λ only.

Claim 4.4. *There exist a radius r_2, $0 < r_2 < \frac{r_0}{4M}$, and a solution $\psi' \in H^1(B^-_{r_2}(0))$ to the problem*

$$\begin{cases} \operatorname{div}(\sigma\nabla\psi') + k^2\psi' = 0 & \text{in} \quad B^-_{r_2}(0), \\ \sigma\nabla\psi'\cdot\nu' + i\lambda'\psi' = 0 & \text{on} \quad B'_{r_2}(0), \end{cases} \tag{4.35}$$

where $\nu' = (0,0,1)$ such that $|\psi'| \geqslant 1$ in $B^-_{r_2}(0)$.

Proof of Claim 4.4. We look for a radius $r_2 > 0$ and for a solution of the form $\psi' = \psi_0 - s$ such that $\psi_0 \in H^1(B^-_{r_2}(0))$ is a weak solution to the problem

$$\begin{cases} \Delta\psi_0 + k^2\psi_0 = 0 & \text{in} \quad B^-_{r_2}(0), \\ \dfrac{\partial\psi_0}{\partial\nu} + i\lambda_0{}'\psi_0 = 0 & \text{on} \quad B'_{r_2}(0), \end{cases} \tag{4.36}$$

satisfying $|\psi_0| \geqslant 2$ in $B^-_{r_2}(0)$.

Furthermore, $s \in H^1(B^-_{r_2}(0))$ is a weak solution to the problem

$$\begin{cases} \operatorname{div}(\sigma\nabla s) + k^2 s = \operatorname{div}((\sigma - I)\nabla\psi_0) & \text{in} \quad B^-_{r_2}(0), \\ \sigma\nabla s\cdot\nu + i\lambda' s = (\sigma - I)\nabla\psi_0\cdot\nu + i(\lambda' - \lambda_0{}')\psi_0 & \text{on} \quad B'_{r_2}(0), \\ s = 0 & \text{on} \quad |y| = r_2, \end{cases} \tag{4.37}$$

such that $s(y) = O(|y|^2)$ near the origin.

We can construct ψ_0 explicitly as follows:

$$\psi_0(y_1,y_2,y_3) = 8\cosh\left(|\lambda_0{}'^2 - k^2|^{\frac{1}{2}}y_1\right)\left[\sin\left(\lambda_0{}'y_3\right) + i\cos\left(\lambda_0{}'y_3\right)\right] \text{ if } k^2 < \lambda_0{}'^2,$$
$$\psi_0(y_1,y_2,y_3) = 8\cos\left(|k^2 - \lambda_0{}'^2|^{\frac{1}{2}}y_1\right)\left[\sin\left(\lambda_0{}'y_3\right) + i\cos\left(\lambda_0{}'y_3\right)\right] \text{ if } k^2 > \lambda_0{}'^2,$$
$$\psi_0(y_1,y_2,y_3) = 8\sin\left(\lambda_0{}'y_3\right) + i8\cos\left(\lambda_0{}'y_3\right) \text{ if } k^2 = \lambda_0{}'^2.$$

Denoting $\tilde{r} = \frac{\pi}{4}\min\{|k^2 - \lambda_0{}'^2|^{-\frac{1}{2}}, (\lambda_0{}')^{-1}\}$, it follows, by straightforward calculations, that $\psi_0 \in H^1(B^-_{\tilde{r}}(0))$ is a weak solution of (4.36) with $r_2 = \tilde{r}$ and $|\psi_0| \geqslant 2$ in $B^-_{\tilde{r}}(0)$.

Let us now look for a solution s to the problem (4.37).

Fixing $r \in (0, \frac{r_0}{8M})$, let us define the space

$$H^1_{0^-}(B^-_r(0)) = \{\eta \in H^1(B^-_r(0)) \text{ such that } \eta(y) = 0 \text{ on } |y| = r\}, \tag{4.38}$$

endowed with the usual $\|\cdot\|_{H^1_0(B_r^-(0))}$ norm. Let us introduce the bilinear form

$$A : H^1_{0-}(B_r^-(0)) \times H^1_{0-}(B_r^-(0)) \to \mathbb{C} \tag{4.39}$$

such that

$$A(\eta_1,\eta_2) = \int_{B_r^-(0)} \sigma \nabla\eta_1 \cdot \nabla\bar{\eta_2} - \int_{B_r^-(0)} k^2 \eta_1 \bar{\eta_2} - \int_{B_r'(0)} i\lambda' \eta_1 \bar{\eta_2} \tag{4.40}$$

and the functional

$$F : H^1_{0-}(B_r^-(0)) \to \mathbb{C} \tag{4.41}$$

such that

$$F(\eta) = \int_{B_r^-(0)} (\sigma - I)\nabla\psi_0 \cdot \nabla\bar{\eta} + i\int_{B_r'(0)} (\lambda' - {\lambda_0}')\psi_0\bar{\eta}\ . \tag{4.42}$$

It immediately follows that A and F are continuous on $H^1_{0-}(B_r^-(0))$ as a bilinear form and a functional, respectively.

Moreover, by standard arguments we can infer that there exists a radius $r_3 > 0$, depending on the a priori data only, such that for every $r \in (0, r_3)$, the bilinear form A is coercive on $H^1_{0-}(B_r^-(0))$. Hence by the Lax–Milgram theorem we deduce that for every $r \in (0, r_3)$, there exists a unique solution $s \in H^1_{0-}(B_r^-(0))$ to the problem (4.37).

By the coercitivity of A, we have that

$$\frac{1}{4}\int_{B_r^-(0)} |\nabla s|^2 \leqslant \left|\int_{B_r^-(0)} (\sigma - I)\nabla\psi_0 \cdot \nabla\bar{s}\right| + \left|\int_{B_r'(0)} (\lambda' - {\lambda_0}')\psi_0\bar{s}\right|. \tag{4.43}$$

By the Schwartz inequality, (4.31), and (4.32), we have that

$$\left|\int_{B_r^-(0)} (\sigma - I)\nabla\psi_0 \cdot \nabla\bar{s}\right| \leqslant 16\Sigma r^2 \int_{B_r^-(0)} |\nabla\psi_0|^2 + \frac{1}{16}\int_{B_r^-(0)} |\nabla s|^2. \tag{4.44}$$

By (4.30), (4.34), and a trace inequality it follows that

$$\left|\int_{B_r'(0)} (\lambda' - {\lambda_0}')\psi_0\bar{s}\right| \leqslant c_2^2 r^4 16\Lambda' \int_{B_r^-(0)} |\nabla\psi_0|^2 + \frac{1}{16} r \int_{B_r^-(0)} |\nabla s|^2. \tag{4.45}$$

Hence inserting (4.44) and (4.45) into (4.43) we obtain that

$$\frac{1}{8}\int_{B_r^-(0)} |\nabla s|^2 \leqslant (16\Sigma + c_2^2 16\Lambda')r^2 \int_{B_r^-(0)} |\nabla\psi_0|^2. \tag{4.46}$$

We have that

$$\frac{1}{8}\int_{B_r^-(0)} |\nabla s|^2 \leqslant \frac{4}{3}\pi(16\Sigma + c_1^2 16\Lambda')r^5 Q, \tag{4.47}$$

where $Q = \sup_{B^-_{\frac{r_0}{8M}}(0)} |\nabla\psi_0|^2$. By standard estimates for solutions of elliptic equations (see, for instance, [12, Chap. 8]) and observing that $Q > 0$ depends on the a priori data only, we can infer that for every $r \in (0, \frac{r_3}{2})$

$$\|s\|_{L^\infty(B_r^-(0))} \leqslant c_4 r^2,$$

where $c_4 > 0$ is a constant depending on the a priori data only.

Hence the claim follows, choosing $r_2 = \min\{\tilde{r}, \frac{r_3}{2}, \frac{1}{\sqrt{c_4}}\}$. □

Let us notice that choosing $r_1 = \theta_2 r_2$ and $\psi(x', x_3) = \psi'(\Phi^{-1}(x', x_3))$, we have that $\psi \in H^1(\Gamma_{I,r_1}(0))$ is a weak solution to the problem (4.20) and is such that $|\psi| \geqslant 1$ in $\Gamma_{I,r_1}(0)$. We conclude the proof observing that (4.22) follows with the same argument as in Theorem 3.2. □

LEMMA 4.5 (volume doubling inequality). *Let u be the solution to the problem (1.1); then there exists a radius $\bar{r} > 0$, such that for every $x_0 \in \Gamma_I^{r_0}$ the following holds:*

$$\int_{\Gamma_{I,\beta r}} |u|^2 \leqslant C\beta^K \int_{\Gamma_{I,r}} |u|^2 \tag{4.48}$$

for every r, β such that $\beta > 1$ and $0 < \beta r < \bar{r}$, where $C > 0, K > 0$ are constants depending on the a priori data only.

Proof. Let $x_0 \in \Gamma_I^{r_0}$ and let r_1 and ψ be, respectively, the radius and the function, introduced in Proposition 4.3. Denoting

$$z = \frac{u}{\psi}, \tag{4.49}$$

it follows that $z \in H^1(\Gamma_{I,r_1}(x_0))$ is a weak solution to the problem

$$\begin{cases} \Delta z + 2\dfrac{\nabla\psi}{\psi} \cdot \nabla z = 0 & \text{in } \Gamma_{I,r_1}(x_0), \\ \dfrac{\partial z}{\partial \nu} = 0 & \text{on } \Delta_{I,r_1}(x_0). \end{cases} \tag{4.50}$$

Proceeding as in Proposition 4.3, we may assume that, up to a rigid transformation of coordinates, $x_0 = 0$ and, by local coordinates, we can locally represent the boundary as a graph of a $C^{1,1}$ function as in (4.23).

Following [2, Theorem 0.8] (see also [4, Proposition 3.5]), we have that there exists a map $\Psi \in C^{1,1}(B_{\rho_2}(0), \mathbb{R}^3)$ such that

$$\Psi(B_{\rho_2}(0)) \subset B_{\rho_1}(0), \tag{4.51}$$

$$\Psi(y', 0) = (y', \varphi_I(y')) \qquad \text{for every } y' \in B'_{\rho_2}(0), \tag{4.52}$$

$$\Gamma_{I,\frac{\rho}{2}} \subset \Psi(B_\rho^-(0)) \subset \Gamma_{I,c_1\rho} \quad \text{for every } \rho \in (0, \rho_2), \tag{4.53}$$

$$\frac{1}{8} \leqslant |\det D\Psi| \leqslant c_2, \tag{4.54}$$

where $\rho_1, 0 < \rho_1 < r_0, \rho_2 > 0, c_1 > 0$, and $c_2 > 0$ are constants depending on r_0, M, and Λ only. Denoting

$$A(y) = |\det D\Psi(y)|(D\Psi^{-1})(\Psi(y))(D\Psi^{-1})^T(\Psi(y)), \tag{4.55}$$

$$B(y) = 2|\det D\Psi(y)|(D\Psi^{-1})(\Psi(y))\frac{\nabla\psi(\Psi(y))}{\psi(\Psi(y))}, \tag{4.56}$$

it follows that $v(y) = z(\Psi(y)) \in H^1(B^-_{\rho_2}(0))$ is a weak solution to the problem

$$\begin{cases} \operatorname{div}(A\nabla v) + B\nabla v = 0 & \text{in } B^-_{\rho_2}(0), \\ A(y',0)\nabla v \cdot \nu' = 0 & \text{on } B'_{\rho_2}(0). \end{cases} \tag{4.57}$$

Hence we are under the assumptions of Theorem 1.3 in [2] and thus we can infer that there exists a radius $\rho_3, 0 < \rho_3 < \rho_2$, depending on the a priori data only, such that

$$\int_{B^-_{\beta\rho}(0)} |v|^2 \leqslant c\beta^K \int_{B^-_{\rho}(0)} |v|^2 \tag{4.58}$$

for every ρ, β such that $\beta > 1$ and $0 < \beta\rho \leqslant \rho_3$, where $c > 0$ is constant depending on the a priori data only, and $K > 0$ depends on the a priori data and increasingly on

$$N(\rho_3) = \frac{\int_{B^-_{\rho_3}(0)} \rho_3^2 |\nabla v|^2 + |v|^2}{\int_{B^-_{\rho_3}(0)} |v|^2}. \tag{4.59}$$

By (4.53) and (4.58), it follows that for every r and $\beta > 1$ such that $0 < r < \beta r < \frac{\rho_3}{2}$,

$$\int_{\Gamma_{I,\beta r}} |z|^2 \leqslant C(2\beta c_1)^K \int_{\Gamma_{I,r}(0)} |z|^2. \tag{4.60}$$

Finally the last inequality, (4.21), and (4.22) imply that

$$\int_{\Gamma_{I,\beta r}} |u|^2 \leqslant C(\beta)^K \int_{\Gamma_{I,r}(0)} |u|^2, \tag{4.61}$$

where $C > 0, K > 0$ are constants depending on a priori data and on $N(\rho_3)$ only. Thus the lemma follows with

$$\bar{r} = \frac{\rho_3}{2}. \tag{4.62}$$

It remains only to majorize the quantity (4.59) by a constant depending on the a priori data only. Let us observe that by (4.53), (4.21), (4.22), and (3.3), we have

$$\int_{B^-_{\rho_3}(0)} |\nabla v|^2 + |v|^2 \leqslant C, \tag{4.63}$$

where $C > 0$ is a constant depending on a priori data only.

On the other hand, we have that, choosing $P_0 = \frac{M}{8\sqrt{1+M^2}\rho_3}\nu$ and $\rho_4 = \frac{1}{32}\frac{M}{\sqrt{1+M^2}}\rho_3$, where ν is the outer unit normal to D at 0, it follows that $B_{\rho_4}(P_0) \subset \Gamma_{I,\frac{\rho_3}{2}}(0)$.

Thus, by (4.53) and (4.22) it follows that

$$\int_{B^-_{\rho_3}(0)} |v|^2 \geqslant C \int_{\Gamma_{I,\frac{\rho_3}{2}}(0)} |u|^2 \geqslant C \int_{B_{\rho_4}(P_0)} |u|^2, \tag{4.64}$$

where $C > 0$ is a constant depending on the a priori data only.

Let us consider a point $Q \in \mathbb{R}^3 \setminus D^+_{2R_0}$ such that

$$B_{4\rho_4}(Q) \subset \mathbb{R}^3 \setminus \overline{D}^+_{2R_0}, \tag{4.65}$$

where R_0 is the radius introduced in Corollary 3.3. Proceeding as in the proof of Theorem 4.2, we cover a path joining P_0 to Q by a chain of balls of radius ρ_4 pairwise tangent to each other. Hence, by an iterated use of the three spheres inequality, we have that the following holds:

$$\|u\|_{L^2(B_{\rho_4(P_0)})} \geqslant C, \tag{4.66}$$

where $C > 0$ is a constant depending on a priori data only. Hence, by (4.63) and (4.66), we can majorize $N(\rho_3)$ by a constant depending on the a priori data only and thus the lemma follows. □

THEOREM 4.6 (surface doubling inequality). *Let u be the solution to the problem (1.1); then there exists a constant $C > 0$ depending on the a priori data only, such that for every $x_0 \in \Gamma_I^{r_0}$ and for every $r \in (0, \frac{\bar{r}}{4})$, the following holds:*

$$\int_{\Delta_{I,2r}(x_0)} |u|^2 \, d\sigma \leqslant C \int_{\Delta_{I,r}(x_0)} |u|^2 \, d\sigma. \tag{4.67}$$

Proof. Let $x_0 \in \Gamma_I^{r_0}$ and let $z \in H^1(\Gamma_{I,r_1}(x_0))$ and $\bar{r}$ be, respectively, the solution to the problem (4.50) defined by (4.49) and the radius introduced in (4.62). By a regularity estimate at the boundary (see, for instance, [4, p. 777]), we have that, for any $r \in (0, \frac{\bar{r}}{4})$, the following holds:

$$\int_{\Delta_{I,r}(x_0)} |\nabla_t z|^2 \leqslant C \left(\frac{1}{r} \int_{\Gamma_{I,2r}(x_0)} |\nabla z|^2 \right)^{1-\gamma} \left(\frac{1}{r^2} \int_{\Delta_{I,r}(x_0)} |z|^2 \right)^{\gamma}, \tag{4.68}$$

where $C > 0$ and $0 < \gamma < 1$ are constants depending on the a priori data only and $\nabla_t z$ represents the tangential gradient.

Moreover, by a well-known estimate of stability for the Cauchy problem (see, for instance, [18]), we have that

$$\begin{aligned} \int_{\Gamma_{I,\frac{r}{2}}(x_0)} |z|^2 \leqslant Cr & \left(\int_{\Delta_{I,r}(x_0)} |z|^2 + r^2 \int_{\Delta_{I,r}(x_0)} |\nabla_t z|^2 \right)^{1-\delta} \\ & \cdot \left(\int_{\Delta_{I,r}(x_0)} |z|^2 + r^2 \int_{\Delta_{I,r}(x_0)} |\nabla_t z|^2 + r \int_{\Gamma_{I,r}(x_0)} |\nabla z|^2 \right)^{\delta}, \end{aligned} \tag{4.69}$$

where $C > 0$ and $0 < \delta < 1$ are constants depending on the a priori data only.

Hence by the Young inequality, (4.68), and (4.69), we obtain

$$\int_{\Gamma_{I,\frac{r}{2}}(x_0)} |z|^2 \leqslant \frac{Cr}{\varepsilon^{\frac{\gamma^2+1-\gamma}{\gamma(1-\gamma)}}} \int_{\Delta_{I,r}(x_0)} |z|^2 + C\varepsilon r^2 \int_{\Gamma_{I,2r}(x_0)} |\nabla z|^2,$$

where $C > 0$ is a constant depending on the a priori data only. By the Caccioppoli inequality, (4.21), and (4.22) we can infer that

$$\int_{\Gamma_{I,\frac{r}{2}}(x_0)} |u|^2 \leqslant \frac{Cr}{\varepsilon^{\frac{\gamma^2+1-\gamma}{\gamma(1-\gamma)}}} \int_{\Delta_{I,2r}(x_0)} |u|^2 + C\varepsilon \int_{\Gamma_{I,8r}(x_0)} |u|^2,$$

where $C > 0$ is a constant depending on the a priori data only.

By (4.48) it follows that

$$\int_{\Gamma_{I,\frac{r}{2}}(x_0)} |u|^2 \leqslant \frac{Cr}{\varepsilon^{\frac{\gamma^2+1-\gamma}{\gamma(1-\gamma)}}} \int_{\Delta_{I,r}(x_0)} |u|^2 + C(8)^K \varepsilon \int_{\Gamma_{I,\frac{r}{2}}(x_0)} |u|^2, \tag{4.70}$$

where $C > 0$ is a constant depending on the a priori data only.

Hence, choosing ε in (4.70) such that $\varepsilon = \frac{1}{2C(8)^K}$ and applying again (4.48) on the left-hand side of (4.70), we obtain that

$$\int_{\Gamma_{I,2r}(x_0)} |u|^2 \leqslant Cr \int_{\Delta_{I,r}(x_0)} |u|^2, \tag{4.71}$$

where $C > 0$ is a constant depending on the a priori data only.

Moreover, by a standard Dirichlet trace inequality, we have that

$$\int_{\Delta_{I,2r}(x_0)} |u|^2 \leqslant C \int_{\Delta_{I,r}(x_0)} |u|^2, \tag{4.72}$$

where $C > 0$ is a constant depending on the a priori data only. □

Corollary 4.7 (A_p property on the boundary). *Let u be the solution to the problem (1.1); then there exist constants $p > 1, A > 0$ depending on the a priori data only, such that for every $x_0 \in \Gamma_I^{r_0}$ and every $r \in (0, \frac{\bar{r}}{4})$, the following holds:*

$$\left(\frac{1}{|\Delta_{I,r}(x_0)|} \int_{\Delta_{I,r}(x_0)} |u|^2 d\sigma\right) \left(\frac{1}{|\Delta_{I,r}(x_0)|} \int_{\Delta_{I,r}(x_0)} |u|^{-\frac{2}{p-1}} d\sigma\right)^{p-1} \leqslant A. \tag{4.73}$$

Proof. Let $x_0 \in \Gamma_I^{r_0}$ and let $r \in (0, \frac{\bar{r}}{4})$; then by a trace inequality (see, for instance, [1, Chap. 5]), it follows that

$$\|u\|_{L^4(\Delta_{I,r}(x_0))} \leqslant C \|u\|_{H^1(\Gamma_{I,r}(x_0))}, \tag{4.74}$$

where $C > 0$ is a constant depending on the a priori data only. By the Caccioppoli inequality and the doubling inequality (4.48) we deduce that

$$\|u\|_{L^4(\Delta_{I,r}(x_0))} \leqslant \frac{C}{r} \|u\|_{L^2(\Gamma_{I,r}(x_0))}, \tag{4.75}$$

where $C > 0$ is a constant depending on the a priori data only. Combining (4.71) and (4.75) we have that

$$\|u\|_{L^4(\Delta_{I,r}(x_0))} \leqslant \frac{C}{\sqrt{r}} \|u\|_{L^2(\Delta_{I,2r}(x_0))}, \tag{4.76}$$

where $C > 0$ is a constant depending on the a priori data only. Thus by the doubling inequality (4.67) we have that

$$\|u\|_{L^4(\Delta_{I,r}(x_0))} \leqslant \frac{C}{\sqrt{r}} \|u\|_{L^2(\Delta_{I,r}(x_0))}. \tag{4.77}$$

Hence, we infer that for every $r \in (0, \frac{\bar{r}}{4})$ and every $x_0 \in \Gamma_I^{r_0}$, the following holds:

$$\left(\frac{1}{r^2} \int_{\Delta_{I,r}} |u|^4\right)^{\frac{1}{4}} \leqslant \left(\frac{C}{r^2} \int_{\Delta_{I,r}} |u|^2\right)^{\frac{1}{2}},$$

obtaining a reverse Hölder inequality.

The result in [9] assures the existence of some $p > 1$ and $A > 0$ depending on the a priori data only such that (4.73) holds. □

Proof of Theorem 2.1. Let x_0 be a point in $\Gamma_I^{r_0}$. Let us pick $r = \frac{\bar r}{8}$, where $\bar r$ is the radius introduced in (4.62). Thus by (4.71) with $u = u_2$ it follows that

$$\int_{\Delta_{I,\frac{\bar r}{8}}(x_0)} |u_2|^2 \mathrm{d}\sigma \geqslant C \int_{\Gamma_{I,\frac{\bar r}{16}}(x_0)} |u_2|^2 \mathrm{d}x, \tag{4.78}$$

where $C > 0$ is a constant depending on the a priori data only.

Let P_0 and $\rho_4 > 0$ be, respectively a point and a radius, such that $B_{\rho_4}(P_0) \subset \Gamma_{I,\frac{\bar r}{16}(x_0)}$. By rephrasing the argument leading to (4.66) we deduce by (4.78) that

$$\int_{\Delta_{I,\frac{\bar r}{8}}(x_0)} |u_2|^2 \mathrm{d}\sigma \geqslant C, \tag{4.79}$$

where $C > 0$ is a constant depending on the a priori data only.

Combining (4.73) and (4.79), we have that for every $x_0 \in \Gamma_I^{r_0}$ the following holds:

$$\left(\int_{\Delta_{I,\frac{\bar r}{8}}(x_0)} |u_2|^{-\frac{2}{p-1}} \mathrm{d}\sigma \right)^{p-1} \leqslant C, \tag{4.80}$$

where $C > 0$ is a constant depending on the a priori data only.

Let us now consider $x \in \Delta_{I,\frac{\bar r}{8}}(x_0)$; then by Theorem 4.2 and (1.4) we have that if $0 < \varepsilon < \varepsilon_0$, the following holds:

$$|\lambda_1(x) - \lambda_2(x)| \leqslant (\Lambda + 1)\eta(\varepsilon) \frac{1}{|u_2(x)|}. \tag{4.81}$$

Hence denoting $\delta = \frac{2}{p-1}$ and combining (4.81) and (4.80), we deduce that

$$\left(\int_{\Delta_{I,\frac{\bar r}{8}(x_0)}} |\lambda_1(x) - \lambda_2(x)|^\delta \right)^{\frac{1}{\delta}} \leqslant \eta(\varepsilon). \tag{4.82}$$

By the a priori bound (1.4), we can infer that

$$\|\lambda_1(x) - \lambda_2(x)\|_{L^2(\Delta_{I,\frac{\bar r}{8}(x_0)})} \leqslant (2\Lambda)^{1-\frac{\delta}{2}} \left(\int_{\Delta_{I,\frac{\bar r}{8}}(x_0)} |\lambda_1(x) - \lambda_2(x)|^\delta \right)^{\frac{1}{2}}. \tag{4.83}$$

Hence, by a possible further replacement of the constants C, θ in (2.12), we can infer that the last inequality and (4.82) yield

$$\|\lambda_1(x) - \lambda_2(x)\|_{L^2(\Delta_{I,\frac{\bar r}{8}}(x_0))} \leqslant \eta(\varepsilon). \tag{4.84}$$

By an interpolation inequality (see, for instance, [4, p. 777]), we have that

$$\|\lambda_1 - \lambda_2\|_{L^\infty(\Delta_{I,\frac{\bar r}{8}(x_0)})} \leqslant C \|\lambda_1 - \lambda_2\|^{\frac{1}{2}}_{L^2(\Delta_{I,\frac{\bar r}{8}}(x_0))} \|\lambda_1 - \lambda_2\|^{\frac{1}{2}}_{C^{0,1}(\Delta_{I,\frac{\bar r}{8}}(x_0))}, \tag{4.85}$$

where $C > 0$ is a constant depending on the a priori data only. Hence by (1.4) and combining (4.84) with (4.85), we obtain, by a possible further replacement of the constants C, θ in (2.12), that

$$\|\lambda_1 - \lambda_2\|_{L^\infty(\Delta_{I,\frac{\bar r}{8}}(x_0))} \leqslant \eta(\varepsilon). \tag{4.86}$$

Let us cover $\Gamma_I^{r_0}$ with the sets $\Delta_{I,\frac{\bar{r}}{8}}(x_j),\ j=1,\dots,J$, with $x_j \in \Gamma_I^{r_0}$.
Let i be an index such that

$$\|\lambda_1-\lambda_2\|_{L^\infty(\Delta_{I,\frac{\bar{r}}{8}}(x_i))} = \|\lambda_1-\lambda_2\|_{L^\infty(\Gamma_I^{r_0})}. \tag{4.87}$$

Thus, by a further possible replacement of the constant C,θ in (2.12), we deduce (2.14) from (4.87) and (4.86) with $x_0 = x_i$. □

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, San Francisco, London, 1975.

[2] V. Adolfsson and L. Escauriaza, *$C^{1,\alpha}$ domains and unique continuation at the boundary*, Comm. Pure Appl. Math, 50 (1997), pp. 935–969.

[3] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solution of elliptic partial differential equations satisfying general boundary conditions.* I., Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[4] G. Alessandrini, E. Beretta, E. Rosset, and S. Vessella, *Optimal stability for inverse elliptic boundary value problems with unknown boundaries*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 29 (2000), pp. 755–806.

[5] G. Alessandrini and A. Morassi, *Strong unique continuation for the Lamé system of elasticity*, Comm. Partial Differential Equations, 26 (2001), pp. 1787–1810.

[6] I. Bushuyev, *Stability of recovering the near-field wave from the scattering amplitude*, Inverse Problems, 12 (1996), pp. 859–867.

[7] F. Cakoni and D. Colton, *The determination of the surface impedance of a partially coated obstacle from far field data*, SIAM J. Appl. Math., 64 (2004), pp. 709–723.

[8] F. Cakoni, D. Colton, and P. Monk, *The direct and inverse scattering problems for partially coated obstacles*, Inverse Problems, 17 (2001), pp. 1997–2015.

[9] R. R. Coifman and C. L. Fefferman, *Weighted norm inequalities for maximal function and singular integrals*, Studia Math., 51 (1974), pp. 241–250.

[10] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Pure Appl. Math. (N.Y.), A Wiley-Interscience Publication, John Wiley & Sons, New York, 1983.

[11] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Appl. Math. Sci. 93, Springer-Verlag, Heidelberg, Germany, 1992.

[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Heidelberg, Germany, 1977.

[13] V. Isakov, *Stability estimates for obstacles in inverse scattering*, J. Comput. Appl. Math., 42 (1991), pp. 79–89.

[14] V. Isakov, *New stability results for soft obstacles in inverse scattering*, Inverse Problems, 9 (1993), pp. 535–543.

[15] C. Labreuche, *Stability of the recovery of surface impedances in inverse scattering*, J. Math. Anal. Appl., 231 (1999), pp. 161–176.

[16] G. M. Lieberman, *Regularized distance and its applications*, Pacific J. Math., 117 (1985), pp. 329–353.

[17] E. Sincich, *Stability and Reconstruction for the Determination of Boundary Terms by a Single Measurement*, Ph.D. thesis, S.I.S.S.A.-I.S.A.S., Trieste, Italy, 2005; available online at http://www.sissa.it/library/.

[18] G. N. Trytten, *Pointwise bound for solution of the Cauchy problem for elliptic equations*, Arch. Rational Mech. Anal., 13 (1963), pp. 222–224.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 452–467

A UNIQUENESS CRITERION FOR THE SIGNORINI PROBLEM WITH COULOMB FRICTION*

YVES RENARD†

Abstract. The purpose of this paper is to study the solutions to the Signorini problem with Coulomb friction (the so-called Coulomb problem). Some optimal a priori estimates are given, and a uniqueness criterion is exhibited. Recently, nonuniqueness examples have been presented in the continuous framework. It is proved, here, that if a solution satisfies a certain hypothesis on the tangential displacement and if the friction coefficient is small enough, it is the unique solution to the problem. In particular, this result can be useful for the search of multisolutions to the Coulomb problem because it eliminates a lot of uniqueness situations.

Key words. unilateral contact, Coulomb friction, uniqueness of solution

AMS subject classifications. 35J85, 74M10

DOI. 10.1137/050635936

Introduction. The so-called Signorini problem with Coulomb friction (or simply the Coulomb problem) has been introduced by Duvaut and Lions [4]. It does not exactly represent the equilibrium of a solid which encounters an obstacle because when the equilibrium is reached (or any steady state solution) the friction condition is no longer an irregular law. The aim of this problem is in fact to be very close to a time semidiscretization of an evolutionary problem by an implicit scheme. The fact that several solutions could coexist in an implicit scheme (independently of the size of the time step) may be an indication that the evolutionary problem has a dynamical bifurcation.

The first existence results for this problem were obtained by Nečas, Jarušek, and Haslinger in [15] for a two-dimensional elastic strip, assuming that the coefficient of friction is small enough and using a shifting technique previously introduced by Fichera and later applied to more general domains by Jarušek [11]. Eck and Jarušek [5] give a different proof using a penalization method. We emphasize that most results on existence for frictional problems involve a condition of smallness for the friction coefficient (and a compact support on Γ_C).

Recently, examples of nonunique solutions have been given by Hild in [7] and [8] for a large friction coefficient. As far as we know, for a fixed geometry, it is still an open question whether or not there is uniqueness of the solution for a sufficiently small friction coefficient. In the finite element approximation framework, the presence of bifurcation has been studied in [9].

The present paper gives the first (partial) result of uniqueness of a solution to the Coulomb problem. The summary is the following. Section 1 introduces strong and weak formulations of the Coulomb problem. Section 2 gives optimal estimates on the solutions. In particular, a comparison is made with the solution to the frictionless contact problem. Section 3 gives an additional estimate for the Tresca problem, i.e., the problem with a given friction threshold. And finally, section 4 gives the partial

*Received by the editors July 13, 2005; accepted for publication (in revised form) January 11, 2006; published electronically May 12, 2006.
http://www.siam.org/journals/sima/38-2/63593.html
†MIP, CNRS UMR 5640, INSAT, Complexe scientifique de Rangueil, 31077 Toulouse, France (Yves.Renard@insa-toulouse.fr).

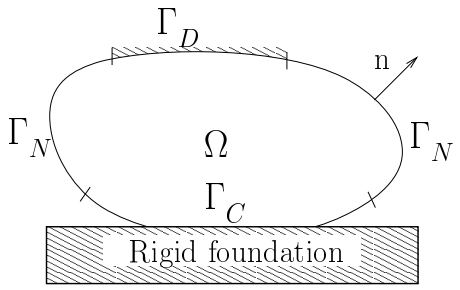


FIG. 1. *Elastic body Ω in frictional contact.*

uniqueness result. It is proved in Proposition 5 for bidimensional problems and a friction coefficient less than one that there is no multisolution with one of the solutions having a tangential displacement with a constant sign. The major result is given by Proposition 6 using the notion of a multiplier in a pair of Sobolev spaces.

1. The Signorini problem with Coulomb friction. Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) be a bounded Lipschitz domain representing the reference configuration of a linearly elastic body.

It is assumed that this body is submitted to a Neumann condition on a part of its boundary $\Gamma_{_N}$, to a Dirichlet condition on another part $\Gamma_{_D}$, and to a unilateral contact with static Coulomb friction condition on the rest of the boundary $\Gamma_{_C}$ between the body and a flat rigid foundation (see Figure 1). This latter part $\Gamma_{_C}$ is supposed to be of nonzero interior in the boundary $\partial\Omega$ of Ω. The problem consists in finding the displacement field $u(t,x)$ satisfying

$$-\operatorname{div}\sigma(u) = f \quad \text{in} \quad \Omega, \tag{1}$$

$$\sigma(u) = \mathcal{A}\varepsilon(u) \quad \text{in } \Omega, \tag{2}$$

$$\sigma(u)\mathrm{n} = F \quad \text{on } \Gamma_{_N}, \tag{3}$$

$$u = 0 \quad \text{on } \Gamma_{_D}, \tag{4}$$

where $\sigma(u)$ is the stress tensor, $\varepsilon(u)$ is the linearized strain tensor, n is the outward unit normal to Ω on $\partial\Omega$, F and f are the given external loads, and $\mathcal{A}$ is the elastic coefficient tensor which satisfies classical conditions of symmetry and ellipticity.

On $\Gamma_{_C}$, it is usual to decompose the displacement and the stress vector in normal and tangential components as follows:

$$u_{_N} = u\cdot\mathrm{n}, \quad u_{_T} = u - u_{_N}\mathrm{n},$$

$$\sigma_{_N}(u) = (\sigma(u)\mathrm{n})\cdot\mathrm{n}, \quad \sigma_{_T}(u) = \sigma(u)\mathrm{n} - \sigma_{_N}(u)\mathrm{n}.$$

To give a clear sense to this decomposition, we assume $\Gamma_{_C}$ to have the $\mathcal{C}^1$ regularity. The unilateral contact condition is expressed by the following complementary condition:

$$u_{_N} \le g, \quad \sigma_{_N}(u) \le 0, \quad (u_{_N} - g)\sigma_{_N}(u) = 0, \tag{5}$$

where g is the normal gap between the elastic solid and the rigid foundation in reference configuration (see Figure 2).

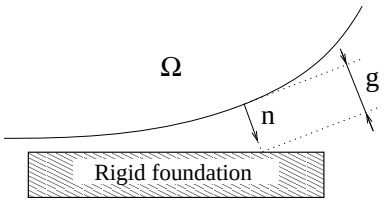


FIG. 2. *Normal gap between the elastic solid* Ω *and the rigid foundation.*

Denoting by $\mathcal{F} \geq 0$ the friction coefficient, the *static Coulomb friction* condition reads as follows:

$$\text{if } u_T = 0, \text{ then } |\sigma_T(u)| \leq -\mathcal{F}\sigma_N(u), \tag{6}$$

$$\text{if } u_T \neq 0, \text{ then } \sigma_T(u) = \mathcal{F}\sigma_N(u)\frac{u_T}{|u_T|}. \tag{7}$$

The friction force satisfies the so-called maximum dissipation principle

$$-\sigma_T(u)\cdot u_T = \sup_{\substack{\mu_T \in \mathbb{R}^{d-1} \\ |\mu_T| \leq -\mathcal{F}\sigma_N(u)}} (-\mu_T \cdot u_T). \tag{8}$$

1.1. Classical weak formulation. We present here the classical weak formulation proposed by Duvaut [3] and Duvaut and Lions [4]. Let us introduce the Hilbert spaces

$$V = \{v \in H^1(\Omega;\mathbb{R}^d), v = 0 \text{ on } \Gamma_D\},$$

$$X = \{v|_{\Gamma_C} : v \in V\} \subset H^{1/2}(\Gamma_C;\mathbb{R}^d),$$

$$X_N = \{v_N|_{\Gamma_C} : v \in V\}, \quad X_T = \{v_T|_{\Gamma_C} : v \in V\},$$

and their topological dual spaces V', X', X'_N, and X'_T. It is assumed that Γ_C is sufficiently smooth such that $X_N \subset H^{1/2}(\Gamma_C)$, $X_T \subset H^{1/2}(\Gamma_C;\mathbb{R}^{d-1})$, $X'_N \subset H^{-1/2}(\Gamma_C)$, and $X'_T \subset H^{-1/2}(\Gamma_C;\mathbb{R}^{d-1})$.

Classically, $H^{1/2}(\Gamma_C)$ is the space of the restriction on Γ_C of traces on $\partial\Omega$ of functions of $H^1(\Omega)$, and $H^{-1/2}(\Gamma_C)$ is the dual space of $H^{1/2}_{00}(\Gamma_C)$, which is the space of the restrictions on Γ_C of functions of $H^{1/2}(\partial\Omega)$ vanishing outside Γ_C. We refer to [1] and [12] for a detailed presentation of trace operators.

Now, the set of admissible displacements is defined as

$$K = \{v \in V, v_N \leq g \quad \text{a.e. on } \Gamma_C\}. \tag{9}$$

The maps

$$a(u,v) = \int_\Omega \mathcal{A}\varepsilon(u) : \varepsilon(v)dx,$$

$$l(v) = \int_\Omega f\cdot v dx + \int_{\Gamma_N} F\cdot v d\Gamma,$$

$$j(\mathcal{F}\lambda_N, v_T) = -\langle \mathcal{F}\lambda_N, |v_T| \rangle_{X'_N, X_N}$$

represent the virtual work of elastic forces, the external load, and the "virtual work" of friction forces, respectively. Standard hypotheses are as follows:

(10) $a(\cdot,\cdot)$ is a bilinear symmetric V-elliptic and continuous form on $V \times V$:
$\exists\ \alpha > 0,\ \exists\ M > 0,\ a(u,u) \geq \alpha \|u\|_V^2,\ a(u,v) \leq M\|u\|_V \|v\|_V\ \forall u, v \in V,$

(11) $l(\cdot)$ is a linear continuous form on V; i.e., $\exists\ L > 0, |l(v)| \leq L\|v\|_V\ \forall v \in V,$

(12) $g \in X_N$,

(13) $\mathcal{F} \in MX_N$ is a nonnegative multiplier in X_N.

The latter condition ensures that $j(\mathcal{F}\lambda_N, v_T)$ is linear continuous on λ_N and convex lower semicontinuous on v_T when λ_N is a nonpositive element of X_N' (see, for instance, [2]). To satisfy condition (10), it is necessary that Γ_D is of nonzero interior in the boundary of Ω and that the elastic coefficient tensor is uniformly elliptic (see [4]).

We refer to Maz'ya and Shaposhnikova [14] for the theory of multipliers. The set MX_N denotes the space of multipliers from X_N into X_N, i.e., the space of function $f : \Gamma_C \longrightarrow \mathbb{R}$ of finite norm

$$\|f\|_{MX_N} = \sup_{\substack{v_N \in X_N \\ v_N \neq 0}} \frac{\|fv_N\|_{X_N}}{\|v_N\|_{X_N}}.$$

This is the norm of the linear mapping $X_N \ni v \longmapsto (fv) \in X_N$. Of course, if $\mathcal{F}$ is a constant on Γ_C, one has $\|\mathcal{F}\|_{MX_N} = \mathcal{F}$. From the fact that Ω is supposed to be a bounded Lipschitz domain and Γ_C is supposed to have the $\mathcal{C}^1$ regularity, it is possible to deduce that for $d = 2$ the space $H^{1/2+\varepsilon}(\Gamma_C)$ is continuously included in MX_N for any $\varepsilon > 0$, and for $d = 3$ the space $H^1(\Gamma_C) \cap L^\infty(\Gamma_C)$ is included in MX_N, continuously for the norm $\|f\|_{H^1(\Gamma_C)} + \|f\|_{L^\infty(\Gamma_C)}$ (see [14]). In particular, the space of Lipschitz continuous functions is continuously included in MX_N.

Condition (10) implies in particular that $a(\cdot,\cdot)$ is a scalar product on V and that the associated norm

$$\|v\|_a = (a(v,v))^{1/2}$$

is equivalent to the usual norm of V:

$$\sqrt{\alpha}\|v\|_V \leq \|v\|_a \leq \sqrt{M}\|v\|_V \quad \forall v \in V.$$

The continuity constant of $l(\cdot)$ can also be given with respect to: $\|\cdot\|_a$:

$$\exists\ L_a > 0, |l(v)| \leq L_a\|v\|_a\ \forall v \in V.$$

Constants L and L_a can be chosen such that

$$\sqrt{\alpha}L_a \leq L \leq \sqrt{M}L_a.$$

The classical weak formulation of problem (1)–(7) is given by

(14)
$$\begin{cases} \text{Find } u \in K \text{ satisfying} \\ \\ a(u, v-u) + j(\mathcal{F}\sigma_N(u), v_T) - j(\mathcal{F}\sigma_N(u), u_T) \geq l(v-u)\ \ \forall\ v \in K. \end{cases}$$

The major difficulty about (14) is due to the coupling between the friction threshold and the contact pressure $\sigma_N(u)$. The consequence is that this problem does not represent a variational inequality, in the sense that it cannot be derived from an optimization problem.

1.2. Neumann to Dirichlet operator. In this section, the Neumann to Dirichlet operator on Γ_C is introduced together with its basic properties. This will allow to restrict the contact and friction problem to Γ_C and obtain useful estimates.

Let $\lambda = (\lambda_N, \lambda_T) \in X'$; then, under hypotheses (10) and (11), the solution u to

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ satisfying} \\ \\ a(u,v) = l(v) + \langle \lambda, v \rangle_{X',X} \quad \forall\ v \in V \end{array} \right. \tag{15}$$

is unique (see [4]). So it is possible to define the operator

$$\begin{array}{rcl} \mathbb{E} : X' & \longrightarrow & X \\ \lambda & \longmapsto & u_{|\Gamma_C}. \end{array}$$

This operator is affine and continuous. Moreover, it is invertible and its inverse is continuous. It is possible to express $\mathbb{E}^{-1}$ as follows: For $w \in X$, let u be the solution to the Dirichlet problem

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ satisfying } u_{|\Gamma_C} = w \text{ and} \\ \\ a(u,v) = l(v) \quad \forall\ v \in V,\ v_{|\Gamma_C} = 0; \end{array} \right. \tag{16}$$

then $\mathbb{E}^{-1}(w)$ is equal to $\lambda \in X'$ defined by

$$\langle \lambda, v \rangle_{X',X} = a(u,v) - l(v) \quad \forall\ v \in V.$$

In a weak sense, one has the relation $\mathbb{E}^{-1}(u) = \sigma(u)\mathrm{n}$ on Γ_C. Now, under hypotheses (10) and (11) one has

$$\|\mathbb{E}(\lambda^1) - \mathbb{E}(\lambda^2)\|_X \leq \frac{C_1^2}{\alpha} \|\lambda^1 - \lambda^2\|_{X'}, \tag{17}$$

where C_1 is the continuity constant of the trace operator on Γ_C and α is the coercivity constant of the bilinear form $a(\cdot,\cdot)$. One can verify it as follows. Let λ^1 and λ^2 be given in X'_T and let u^1, u^2 be the corresponding solutions to (15); then

$$\begin{array}{rcl} \alpha \|u^1 - u^2\|_V^2 & \leq & a(u^1 - u^2, u^1 - u^2) \ = \ \langle \lambda^1 - \lambda^2, u^1 - u^2 \rangle_{X',X} \\ & \leq & C_1 \|\lambda^1 - \lambda^2\|_{X'} \|u^1 - u^2\|_V \end{array} \tag{18}$$

and, consequently,

$$\|u^1 - u^2\|_V \leq \frac{C_1}{\alpha} \|\lambda^1 - \lambda^2\|_{X'}. \tag{19}$$

Conversely, one has

$$\|\mathbb{E}^{-1}(u^1) - \mathbb{E}^{-1}(u^2)\|_{X'} \leq M C_2^2 \|u^1 - u^2\|_X, \tag{20}$$

where M is the continuity constant of $a(\cdot,\cdot)$ and $C_2 > 0$ is the continuity constant of the homogeneous Dirichlet problem corresponding to (16) (i.e., with $l(v) \equiv 0$ and

$C_2 = \sup_{\substack{v\in X\\ v\neq 0}} \frac{\|w\|_V}{\|v\|_X}$, where $w_{|\Gamma_C} = v$ and $a(w,z) = 0 \ \ \forall z \in V$). This latter estimate can be performed as follows:

$$\begin{aligned}\|\mathbb{E}^{-1}(u^1) - \mathbb{E}^{-1}(u^2)\|_{X'} &= \sup_{\substack{v\in X\\ v\neq 0}} \frac{\langle \mathbb{E}^{-1}(u^1) - \mathbb{E}^{-1}(u^2), v\rangle_{X',X}}{\|v\|_X} \\ &= \sup_{\substack{v\in X\\ v\neq 0}} \left(\inf_{\{w\in V : w_{|\Gamma_C} = v\}} \frac{a(u^1-u^2, w)}{\|v\|_X} \right) \\ &\leq M\gamma \|u^1 - u^2\|_V, \end{aligned} \tag{21}$$

where $\gamma = \sup_{\substack{v\in X\\ v\neq 0}} \ \inf_{\{w\in V : w_{|\Gamma_C} = v\}} \frac{\|w\|_V}{\|v\|_X}$ is the continuity constant of the homogeneous Poisson problem with respect to a Dirichlet condition on Γ_C. Using $\gamma \leq C_2$, this gives (20).

It is also possible to define the following norms on Γ_C relative to $a(\cdot,\cdot)$:

$$\|v\|_{a,\Gamma_C} = \inf_{w\in V,\ w_{|\Gamma_C} = v} \|w\|_a,$$

$$\|\lambda\|_{-a,\Gamma_C} = \sup_{\substack{v\in X\\ v\neq 0}} \frac{\langle \lambda, v\rangle_{X',X}}{\|v\|_{a,\Gamma_C}} = \sup_{\substack{v\in V\\ v\neq 0}} \frac{\langle \lambda, v\rangle_{X',X}}{\|v\|_a}.$$

These are equivalent, respectively, to the norms in X and X':

$$\frac{\sqrt{\alpha}}{C_1}\|v\|_X \leq \|v\|_{a,\Gamma_C} \leq \sqrt{M}\gamma\|v\|_X,$$

$$\frac{1}{\sqrt{M}\gamma}\|\lambda\|_{X'} \leq \|\lambda\|_{-a,\Gamma_C} \leq \frac{C_1}{\sqrt{\alpha}}\|\lambda\|_{X'}.$$

With these norms, the estimates are straightforward since the following lemma holds.

LEMMA 1. *Let λ^1 and λ^2 be two elements of X' and let $u^1 = \mathbb{E}(\lambda^1)$, $u^2 = \mathbb{E}(\lambda^2)$; then under hypotheses* (10) *and* (11) *one has*

$$\|u^1 - u^2\|_{a,\Gamma_C} = \|u^1 - u^2\|_a = \|\lambda^1 - \lambda^2\|_{-a,\Gamma_C}.$$

Proof. On the one hand, one has

$$\|u^1 - u^2\|^2_{a,\Gamma_C} = \inf_{\substack{w\in V\\ w_{|\Gamma_c} = u^1 - u^2}} \|w\|_a^2 = \|u^1 - u^2\|_a^2,$$

because $u^1 - u^2$ is the minimum of $\frac{1}{2}\|w\|_a^2$ under the constraint $w_{|\Gamma_C} = u^1 - u^2$. This implies

$$\begin{aligned}\|u^1 - u^2\|^2_{a,\Gamma_C} &= a(u^1 - u^2, u^1 - u^2) = \langle \lambda^1 - \lambda^2, u^1 - u^2\rangle_{X',X} \\ &\leq \|\lambda^1 - \lambda^2\|_{-a,\Gamma_C}\|u^1 - u^2\|_{a,\Gamma_C},\end{aligned}$$

and finally

$$\|u^1 - u^2\|_{a,\Gamma_C} \leq \|\lambda^1 - \lambda^2\|_{-a,\Gamma_C}.$$

On the other hand, one has

$$\|\lambda^1-\lambda^2\|_{-a,\Gamma_C} = \sup_{\substack{v\in X\\ v\neq 0}} \frac{\langle \lambda^1-\lambda^2, v\rangle_{X',X}}{\|v\|_{a,\Gamma_C}} = \sup_{\substack{v\in X\\ v\neq 0}} \inf_{\substack{w\in V\\ w\big|_{\Gamma_C}=v}} \frac{a(u^1-u^2,w)}{\|v\|_{a,\Gamma_C}},$$

$$\leq \sup_{\substack{v\in X\\ v\neq 0}} \inf_{\substack{w\in V\\ w\big|_{\Gamma_C}=v}} \frac{\|u^1-u^2\|_a\|w\|_a}{\|v\|_{a,\Gamma_C}} = \|u^1-u^2\|_a = \|u^1-u^2\|_{a,\Gamma_C},$$

which ends the proof of the lemma. □

1.3. Direct weak inclusion formulation. Let

$$K_N = \{v_N \in X_N : v_N \leq 0 \text{ a.e. on } \Gamma_C\}$$

be the (translated) set of admissible normal displacements on Γ_C. The normal cone in X'_N to K_N at $v_N \in X_N$ is defined as

$$N_{K_N}(v_N) = \{\mu_N \in X'_N : \langle \mu_N, w_N - v_N\rangle_{X'_N,X_N} \leq 0 \ \forall w_N \in K_N\}.$$

In particular, $N_{K_N}(v_N) = \emptyset$ if $v_N \notin K_N$. The subgradient of $j(\mathcal{F}\lambda_N, u_T)$ with respect to the second variable is given by

$$\partial_2 j(\mathcal{F}\lambda_N, u_T) = \{\mu_T \in X'_T : j(\mathcal{F}\lambda_N, v_T) \geq j(\mathcal{F}\lambda_N, u_T) + \langle \mu_T, v_T - u_T\rangle_{X'_T,X_T} \ \forall v_T \in X_T\}.$$

With this notation, problem (14) is equivalent to the following problem:

$$\text{(22)} \qquad \begin{cases} \text{Find } u \in V,\ \lambda_N \in X'_N, \text{ and } \lambda_T \in X'_T \text{ satisfying} \\ (u_N, u_T) = \mathbb{E}(\lambda_N, \lambda_T), \\ -\lambda_N \in N_{K_N}(u_N - g) \quad \text{in } X'_N, \\ -\lambda_T \in \partial_2 j(\mathcal{F}\lambda_N, u_T) \quad \text{in } X'_T. \end{cases}$$

More details on this equivalence can be found in [13].

Remark 1. Inclusion $-\lambda_N \in N_{K_N}(u_N - g)$ is equivalent to the complementarity relations

$$u_N \leq g, \quad \langle \lambda_N, v_N\rangle_{X'_N,X_N} \geq 0 \ \ \forall v_N \in K_N, \quad \langle \lambda_N, u_N - g\rangle_{X'_N,X_N} = 0,$$

which is the weak formulation of the strong complementarity relations (5) for the contact conditions. Similarly, the second inclusion $-\lambda_T \in \partial_2 j(\mathcal{F}\lambda_N, u_T)$ represents the friction condition.

1.4. Hybrid weak inclusion formulation. We will now consider the sets of admissible stresses. The set of admissible normal stresses on Γ_C can be defined as

$$\Lambda_N = \{\lambda_N \in X'_N : \langle \lambda_N, v_N\rangle_{X'_N,X_N} \geq 0 \ \ \forall v_N \in K_N\}.$$

This is the opposite of K_N^*, the polar cone to K_N. The set of admissible tangential stresses on Γ_C can be defined as

$$\Lambda_T(\mathcal{F}\lambda_N) = \{\lambda_T \in X_T' : -\langle \lambda_T, w_T\rangle_{X_T', X_T} + \langle \mathcal{F}\lambda_N, |w_T|\rangle_{X_N', X_N} \leq 0 \ \ \forall w_T \in X_T\}.$$

With this, problem (14) is equivalent to the following problem:

$$\text{(23)} \qquad \begin{cases} \text{Find } u \in V,\ \lambda_N \in X_N',\ \text{and } \lambda_T \in X_T' \text{ satisfying} \\ (u_N, u_T) = \mathbb{E}(\lambda_N, \lambda_T), \\ -(u_N - g) \in N_{\Lambda_N}(\lambda_N) \quad \text{in } X_N, \\ -u_T \in N_{\Lambda_T(\mathcal{F}\lambda_N)}(\lambda_T) \quad \text{in } X_T, \end{cases}$$

where the two inclusions can be replaced by inequalities as follows:

$$\text{(24)} \qquad \begin{cases} \text{Find } u \in V,\ \lambda_N \in X_N',\ \text{and } \lambda_T \in X_T' \text{ satisfying} \\ (u_N, u_T) = \mathbb{E}(\lambda_N, \lambda_T), \\ \lambda_N \in \Lambda_N, \ \ \langle \mu_N - \lambda_N, u_N - g\rangle_{X_N', X_N} \geq 0 \ \ \forall \mu_N \in \Lambda_N, \\ \lambda_T \in \Lambda_T(\mathcal{F}\lambda_N), \ \ \langle \mu_T - \lambda_T, u_T\rangle_{X_T', X_T} \geq 0 \ \ \forall \mu_T \in \Lambda_T(\mathcal{F}\lambda_N). \end{cases}$$

Remark 2. The inclusion $-u_T \in N_{\Lambda_T(\mathcal{F}\lambda_N)}(\lambda_T)$ implies the complementarity relation

$$\langle \lambda_T, u_T\rangle_{X_T', X_T} = \langle \mathcal{F}\lambda_N, |u_T|\rangle_{X_N', X_N}$$

and the weak maximum dissipation principle

$$-\langle \lambda_T, u_T\rangle_{X_T', X_T} = \sup_{\mu_T \in \Lambda_T(\mathcal{F}\lambda_N)} \langle -\mu_T, u_T\rangle_{X_T', X_T},$$

which is the weak formulation of (8).

2. Optimal a priori estimates on the solutions to the Coulomb problem. For the sake of simplicity, a vanishing contact gap ($g \equiv 0$) will be considered in the following.

Remark 3. In the case of a nonvanishing gap, it is possible to find $u_g \in V$ such that $u_{g|_{\Gamma_C}} = g\mathrm{n}$, and then $w = u - u_g$ is solution to the problem

$$\text{(25)} \qquad \begin{cases} \text{Find } w \in V,\ \lambda_N \in X_N',\ \text{and } \lambda_T \in X_T' \text{ satisfying} \\ a(w, v) = l(v) - a(u_g, v) + \langle \lambda_N, w_N\rangle_{X_N', X_N} + \langle \lambda_T, w_T\rangle_{X_T', X_T}, \\ -w_N \in N_{\Lambda_N}(\lambda_N) \quad \text{in } X_N, \\ -w_T \in N_{\Lambda_T(\mathcal{F}\lambda_N)}(\lambda_T) \quad \text{in } X_T, \end{cases}$$

i.e., a contact problem without gap but with a modified source term.

Following Remarks 1 and 2, a solution (u, λ) to problem (22) (i.e., a solution u to problem (14)) satisfies the complementarity relations

$$\langle \lambda_N, u_N \rangle_{X'_N, X_N} = 0,$$

$$\langle \lambda_T, u_T \rangle_{X'_T, X_T} = \langle \mathcal{F}\lambda_N, |u_T| \rangle_{X'_N, X_N}.$$

This implies

$$\langle \lambda, u \rangle_{X', X} \leq 0,$$

which expresses the dissipativity of contact and friction conditions. The first consequence of this is that solutions to problem (14) can be bounded independently of the friction coefficient.

PROPOSITION 1. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied and* $g \equiv 0$, *let* (u, λ) *be a solution to problem* (22), *which means that* u *is a solution to problem* (14); *then*

$$\|u\|_a \leq L_a, \quad \|\lambda\|_{-a,\Gamma_C} \leq L_a,$$

$$\|u\|_V \leq \frac{L}{\alpha}, \quad \|\lambda\|_{X'} \leq L\gamma\sqrt{\frac{M}{\alpha}}.$$

Proof. One has

$$\|u\|_a^2 = a(u,u) = l(u) + \langle \lambda, u \rangle_{X',X} \leq L_a \|u\|_a,$$

which states the first estimates. The estimate on $\|\lambda\|_{-a,\Gamma_C}$ can be performed using the intermediary solution $u^{\mathcal{N}}$ to the following problem with a homogeneous Neumann condition on Γ_C:

$$a(u^{\mathcal{N}}, v) = l(v) \quad \forall v \in V. \tag{26}$$

Since $\|u^{\mathcal{N}}\|_a \leq L_a$ for the same reason as for u, and using Lemma 1, one has

$$\|\lambda - 0\|_{-a,\Gamma_C}^2 = a(u - u^{\mathcal{N}}, u - u^{\mathcal{N}}) = \langle \lambda, u - u^{\mathcal{N}} \rangle_{X',X} \leq -\langle \lambda, u^{\mathcal{N}} \rangle_{X',X} \leq L_a \|\lambda\|_{-a,\Gamma_C}.$$

The two last estimates can be stated thanks to equivalence of norms introduced in section 1.1. ∎

It is possible to compare $\|u\|_a$ to the corresponding norm of the solution u^c to the Signorini problem without friction defined as follows:

$$\begin{cases} \text{Find } u^c \in K \text{ satisfying} \\ \\ a(u^c, v - u^c) \geq l(v - u^c) \quad \forall\ v \in K. \end{cases} \tag{27}$$

It is well known that under hypotheses (10) and (11), this problem has a unique solution (see [12]).

PROPOSITION 2. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied and* $g \equiv 0$, *let* u *be a solution to problem* (14), *let* u^c *be the unique solution to problem* (27), *and let* $u^{\mathcal{N}}$ *be the solution to problem* (26); *then*

$$\|u\|_a \leq \|u^c\|_a \leq \|u^{\mathcal{N}}\|_a.$$

Proof. One has

$$a(u^{\mathcal{N}}, u^{\mathcal{N}}) = l(u^{\mathcal{N}}), \quad a(u^c, u^c) = l(u^c), \quad a(u,u) = l(u) + \langle \lambda_{_T}, u_{_T} \rangle_{X'_{_T}, X_{_T}}.$$

Since u^c is the solution to the Signorini problem without friction, it minimizes over K the energy functional $\frac{1}{2}a(v,v) - l(v)$. The solution $u^{\mathcal{N}}$ minimizes this energy functional over V. Thus, since $u \in K$, one has

$$\frac{1}{2}a(u^{\mathcal{N}}, u^{\mathcal{N}}) - l(u^{\mathcal{N}}) \leq \frac{1}{2}a(u^c, u^c) - l(u^c) \leq \frac{1}{2}a(u,u) - l(u),$$

and the following relations allow one to conclude that

$$a(u^c, u^c) - a(u^{\mathcal{N}}, u^{\mathcal{N}}) = l(u^c - u^{\mathcal{N}}),$$

$$a(u,u) - a(u^c, u^c) = l(u - u^c) + \langle \lambda_{_T}, u_{_T} \rangle_{X'_{_T}, X_{_T}}$$

because then

$$\frac{1}{2}a(u^c, u^c) - \frac{1}{2}a(u^{\mathcal{N}}, u^{\mathcal{N}}) \leq 0,$$

$$\frac{1}{2}a(u,u) \leq \frac{1}{2}a(u^c, u^c) + \langle \mathcal{F}\lambda_{_N}, |u_{_T}| \rangle_{X'_{_N}, X_{_N}} \leq \frac{1}{2}a(u^c, u^c). \qquad \square$$

It is also possible to estimate how far from u^c is a solution u to problem (14). Let us introduce the following norms on $\Gamma_{_C}$. For $v \in X$ let us define

$$\|v_{_T}\|_{a,\Gamma_{_C}} = \inf_{\substack{w \in V \\ w_{_T} = v_{_T}}} \|w\|_a = \inf_{\substack{z \in X \\ z_{_T} = v_{_T}}} \|z\|_{a,\Gamma_{_C}},$$

$$\|v_{_N}\|_{a,\Gamma_{_C}} = \inf_{\substack{w \in V \\ w_{_N} = v_{_N}}} \|w\|_a = \inf_{\substack{z \in X \\ z_{_N} = v_{_N}}} \|v\|_{a,\Gamma_{_C}}.$$

One has

$$\|v_{_T}\|_{a,\Gamma_{_C}} \leq \|v\|_{a,\Gamma_{_C}}, \quad \|v_{_N}\|_{a,\Gamma_{_C}} \leq \|v\|_{a,\Gamma_{_C}}.$$

Now, for $\lambda \in X'$, let us define

$$\|\lambda_{_T}\|_{-a,\Gamma_{_C}} = \sup_{\substack{v_{_T} \in X_{_T} \\ v_{_T} \neq 0}} \frac{\langle \lambda_{_T}, v_{_T} \rangle_{X'_{_T}, X_{_T}}}{\|v_{_T}\|_{a,\Gamma_{_C}}} = \sup_{\substack{v \in X \\ v \neq 0}} \frac{\langle \lambda_{_T}, v_{_T} \rangle_{X'_{_T}, X_{_T}}}{\|v\|_{a,\Gamma_{_C}}},$$

$$\|\lambda_{_N}\|_{-a,\Gamma_{_C}} = \sup_{\substack{v_{_N} \in X_{_N} \\ v_{_N} \neq 0}} \frac{\langle \lambda_{_N}, v_{_N} \rangle_{X'_{_N}, X_{_N}}}{\|v_{_N}\|_{a,\Gamma_{_C}}} = \sup_{\substack{v \in X \\ v \neq 0}} \frac{\langle \lambda_{_N}, v_{_N} \rangle_{X'_{_N}, X_{_N}}}{\|v\|_{a,\Gamma_{_C}}}.$$

Then, the following equivalences of norms are immediate:

$$\frac{\sqrt{\alpha}}{C_1}\|v_{_N}\|_{_{X_N}} \leq \|v_{_N}\|_{a,\Gamma_{_C}} \leq \gamma\sqrt{M}\|v_{_N}\|_{_{X_N}},$$

$$\frac{\sqrt{\alpha}}{C_1}\|v_{_T}\|_{_{X_T}} \leq \|v_{_T}\|_{a,\Gamma_{_C}} \leq \gamma\sqrt{M}\|v_{_T}\|_{_{X_T}},$$

$$\frac{1}{\gamma\sqrt{M}}\|\lambda_N\|_{X'_N} \leq \|\lambda_N\|_{-a,\Gamma_C} \leq \frac{C_1}{\sqrt{\alpha}}\|\lambda_N\|_{X'_N},$$

$$\frac{1}{\gamma\sqrt{M}}\|\lambda_T\|_{X'_T} \leq \|\lambda_T\|_{-a,\Gamma_C} \leq \frac{C_1}{\sqrt{\alpha}}\|\lambda_T\|_{X'_T}.$$

And the following result can be easily deduced.

LEMMA 2. *There exists $C_3 > 0$ such that for all $\lambda \in X'$*

$$\|\lambda_T\|_{-a,\Gamma_C} \leq C_3\|\lambda\|_{-a,\Gamma_C}, \quad \|\lambda_N\|_{-a,\Gamma_C} \leq C_3\|\lambda\|_{-a,\Gamma_C}.$$

This also allows us to define an equivalent norm on MX_N given for $\mathcal{F} \in MX_N$ by

$$\|\mathcal{F}\|_a = \sup_{\substack{v_N \in X_N \\ v_N \neq 0}} \frac{\|\mathcal{F}v_N\|_{a,\Gamma_C}}{\|v_N\|_{a,\Gamma_C}},$$

which satisfies

$$\frac{\sqrt{\alpha}}{C_1\gamma\sqrt{M}}\|\mathcal{F}\|_a \leq \|\mathcal{F}\|_{MX_N} \leq \frac{C_1\gamma\sqrt{M}}{\sqrt{\alpha}}\|\mathcal{F}\|_a.$$

With these definitions, the following result holds.

LEMMA 3. *There exists $C_4 > 0$ such that*

$$\|\mathcal{F}|v_T|\,\|_{a,\Gamma_C} \leq C_4\|\mathcal{F}\|_a\|v_T\|_{a,\Gamma_C} \quad \forall v_T \in X_T.$$

Proof. One has

$$\|\mathcal{F}|v_T|\,\|_{a,\Gamma_C} \leq \|\mathcal{F}\|_a\|\,|v_T|\,\|_{a,\Gamma_C}.$$

Moreover, it is known (see [1]) that the norm $\|\cdot\|_{X_N}$ is equivalent to the norm

$$\|v_N\|^2_{1/2,\Gamma_C} = \|v_N\|^2_{L^2(\Gamma_C)} + \int_{\Gamma_C}\int_{\Gamma_C} \frac{|v_N(x) - v_N(y)|^2}{|x-y|^d}dxdy,$$

and it is easy to verify that $\|\,|v_T|\,\|_{1/2,\Gamma_C} \leq \|v_T\|_{1/2,\Gamma_C}$ for any $v_T \in X_T$. Thus, the result can be deduced from the previously presented equivalences of norms. ∎

Of course the tangential stress on Γ_C corresponding to u^c is vanishing. The tangential stress corresponding to u can be estimated as follows. As $\lambda_T \in \Lambda_T(\mathcal{F}\lambda_N)$, one has

$$\|\lambda_T\|_{-a,\Gamma_C} = \sup_{\substack{v_T \in X_T \\ v_T \neq 0}} \frac{\langle \lambda_T, v_T\rangle_{X'_T,X_T}}{\|v_T\|_{a,\Gamma_C}} \leq \sup_{\substack{v_T \in X_T \\ v_T \neq 0}} \frac{-\langle \mathcal{F}\lambda_N, |v_T|\rangle_{X'_N,X_N}}{\|v_T\|_{a,\Gamma_C}}$$
$$\leq C_4\|\mathcal{F}\|_a\|\lambda_N\|_{-a,\Gamma_C}.$$

Now, with the result of Proposition 1 this means that

$$\|\lambda_T\|_{-a,\Gamma_C} \leq L_a C_3 C_4\|\mathcal{F}\|_a, \tag{28}$$

and the following result holds.

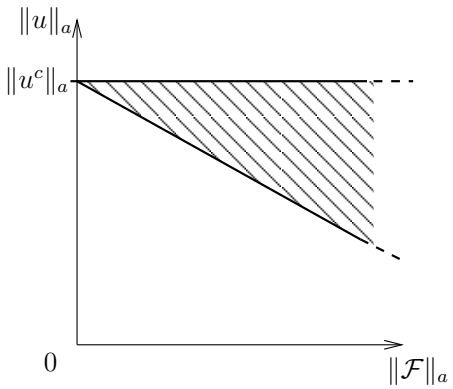


FIG. 3. *Admissibility zone for* $\|u\|_a$.

PROPOSITION 3. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied and* $g \equiv 0$, *let* u *be a solution to problem* (14) *and let* u^c *be the solution to problem* (27); *then*

$$\|u^c - u\|_a \leq L_a C_3 C_4 \|\mathcal{F}\|_a,$$
$$\|u^c - u\|_V \leq \frac{L C_3 C_4}{\alpha} \|\mathcal{F}\|_a.$$

Proof. With $\lambda \in X'$ and $\lambda^c \in X'$ the corresponding stresses on Γ_C, because $-\lambda_N^c \in N_{K_N}(u_N^c)$ and $-\lambda_N \in N_{K_N}(u_N - g)$ and because of the fact that N_{K_N} is a monotone set-valued map, one has

$$\langle \lambda_N^c - \lambda_N, u_N^c - u_N \rangle_{X'_N, X_N} \leq 0.$$

Now, $\|u^c - u\|_a$ can be estimated as follows:

$$\|u^c - u\|_a^2 = a(u^c - u, u^c - u) = \langle \lambda^c - \lambda, u^c - u \rangle_{X',X} \leq \|\lambda_T\|_{-a,\Gamma_C} \|u^c - u\|_a,$$

which gives the result taking into account (28). □

The latter result implies that if problem (14) has several solutions, then they are in a ball of radius $L_a C_3 C_4 \|\mathcal{F}\|_a$ centered around u^c. In particular, if u^1 and u^2 are two solutions to problem (14), one has $\|u^1 - u^2\|_a \leq 2 L_a C_3 C_4 \|\mathcal{F}\|_a$. This is illustrated by Figure 3.

Remark 4. For a friction coefficient $\mathcal{F}$ constant on Γ_C, the graph in Figure 3 can be more precise for $\mathcal{F} = \|\mathcal{F}\|_a$ small, since, from the proof of Proposition 2 and the continuity result given by the latter proposition, one can deduce $\|u\|_a^2 \leq \|u^c\|_a^2 + \mathcal{F}\langle \lambda_N^c, |u_T^c| \rangle_{X'_N, X_N}$ at least if $\langle \lambda_N^c, |u_T^c| \rangle_{X'_N, X_N} < 0$. Of course, if $\langle \lambda_N^c, |u_T^c| \rangle_{X'_N, X_N} = 0$, the solution u^c to the Signorini problem without friction is also a solution to the Coulomb problem for any friction coefficient.

3. Elementary estimates on the Tresca problem. What is usually called the Tresca problem is the friction problem with a given friction threshold. Let $\theta \in X'_N$ be given. Then it can be formulated as follows:

$$\text{(29)} \qquad \begin{cases} \text{Find } u \in K \text{ satisfying} \\ \\ a(u, v-u) + j(\theta, v_T) - j(\theta, u_T) \geq l(v-u) \quad \forall\, v \in K. \end{cases}$$

It is well known that under standard hypotheses (10), (11), and (13), this problem has a unique solution (see [12]) which minimizes the functional $\frac{1}{2}a(u,u)+j(\theta,u)-l(u)$.

In fact, it is not difficult to verify that all the estimates given in the previous section for the solutions to the Coulomb problem are still valid for the solution to the Tresca problem. Moreover, the solution to the Tresca problem continuously depends on the friction threshold θ. This result is stated in the following lemma.

LEMMA 4. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied, if* u^1, u^2 *are the solutions to problem* (29) *for a friction threshold* $\theta^1 \in \Lambda_N$ *and* $\theta^2 \in \Lambda_N$, *respectively, then there exists a constant* $C_5 > 0$ *independent of* θ^1 *and* θ^2 *such that the following estimate holds:*

$$\|u^1 - u^2\|_a^2 \leq C_5 \|\theta^1 - \theta^2\|_{-a,\Gamma_C}.$$

Proof. One has

$$a(u^1, u^2 - u^1) - l(u^2 - u^1) + j(\theta^1, u^2) - j(\theta^1, u^1) \geq 0,$$

$$a(u^2, u^1 - u^2) - l(u^1 - u^2) + j(\theta^2, u^1) - j(\theta^2, u^2) \geq 0,$$

which implies

$$\|u^1 - u^2\|_a^2 \quad \leq \quad \langle \theta^1 - \theta^2, |u_T^1| - |u_T^2| \rangle_{X'_N, X_N},$$

which gives the estimate using Proposition 1 (in fact, $C_5 \leq 2C_4 L_a$). ∎

Remark 5. It does not seem possible to establish a Lipschitz continuity with respect to the friction threshold θ. Such a result would automatically imply the uniqueness of the solution to the Coulomb problem for a sufficiently small friction coefficient.

4. A uniqueness criterion. Hild in [7, 8] exhibits some multisolutions for the Coulomb problem on triangular domains. These solutions have been obtained for a large friction coefficient ($\mathcal{F} > 1$) and for a tangential displacement having a constant sign. For the moment, it seems that no multisolution has been exhibited for an arbitrary small friction coefficient in the continuous case, although such a result exists for finite element approximation in [6], albeit for a variable geometry. As far as we know, no uniqueness result has been proved even for a sufficiently small friction coefficient. The result presented here is a partial uniqueness result, which determines some cases where it is possible to say that a particular solution of the Coulomb problem is in fact the unique solution. A contrario, this result can be used to search multisolutions for an arbitrary small friction coefficient, by the fact that it eliminates a lot of situations. The partial uniqueness results we present in this section are deduced from the estimate given by the following lemma.

LEMMA 5. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied and* $g \equiv 0$, *if* u^1 *and* u^2 *are two solutions to problem* (14) *and* λ^1 *and* λ^2 *are the corresponding contact stresses on* Γ_C, *then one has the following estimate:*

$$\|u^1 - u^2\|_a^2 = \|\lambda^1 - \lambda^2\|_{-a,\Gamma_C}^2 \leq \quad \langle \zeta - \lambda_T^2, u_T^1 - u_T^2 \rangle_{X'_T, X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\lambda_N^1, u_T^2).$$

Proof. One has

$$\|u^1 - u^2\|_a^2 = \|\lambda^1 - \lambda^2\|_{-a,\Gamma_C}^2 = \langle \lambda_N^1 - \lambda_N^2, u_N^1 - u_N^2 \rangle_{X'_N, X_N} + \langle \lambda_T^1 - \lambda_T^2, u_T^1 - u_T^2 \rangle_{X'_T, X_T}.$$

Because N_{K_N} is a monotone set-valued map, one has $\langle \lambda_N^1 - \lambda_N^2, u_N^1 - u_N^2 \rangle_{X'_N, X_N} \leq 0$. Thus

$$\|u^1 - u^2\|_a^2 \leq \langle (\lambda_T^1 - \zeta) + (\zeta - \lambda_T^2), u_T^1 - u_T^2 \rangle_{X'_T, X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\lambda_N^1, u_T^2).$$

But $\partial_2 j(\mathcal{F}\lambda_N, u_T)$ is also a monotone set-valued map with respect to its second variable, which implies the result (and also the fact that $\|u^1 - u^2\|_a^2 \leq \inf_{\zeta \in -\partial_2 j(\mathcal{F}\lambda_N^1, u_T^2)} \|\zeta - \lambda_T^2\|_{-a,\Gamma_C}$). ∎

An immediate consequence of this lemma is the following result for a vanishing tangential displacement.

PROPOSITION 4. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied and* $g \equiv 0$, *if* u *is a solution to problem* (14) *such that* $u_T = 0$ *a.e. on* Γ_C *and if* $C_3 C_4 \|\mathcal{F}\|_a < 1$, *then* u *is the unique solution to problem* (14).

Proof. Let us assume that $\overline{u}$ is another solution to problem (14). Then from Lemma 5 one has

$$\|u - \overline{u}\|_a^2 \leq \langle \zeta - \overline{\lambda}_T, u_T - \overline{u}_T \rangle_{X'_T, X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\lambda_N, \overline{u}_T),$$

but, because $u_T = 0$ and due to the complementarity relations $\langle \overline{\lambda}_T, \overline{u}_T \rangle_{X'_T, X_T} = \langle \mathcal{F}\overline{\lambda}_N, |\overline{u}_T| \rangle_{X'_N, X_N}$ and $\langle \zeta, \overline{u}_T \rangle_{X'_T, X_T} = \langle \mathcal{F}\lambda_N, |\overline{u}_T| \rangle_{X'_N, X_N}$, it implies using Lemma 3 that

$$\begin{aligned} \|u - \overline{u}\|_a^2 &\leq \langle \mathcal{F}(\overline{\lambda}_N - \lambda_N), |\overline{u}_T| \rangle_{X'_N, X_N} \leq C_3 C_4 \|\mathcal{F}\|_a \|\lambda - \overline{\lambda}\|_{-a,\Gamma_C} \|u_T - \overline{u}_T\|_{a,\Gamma_C}^2 \\ &\leq C_3 C_4 \|\mathcal{F}\|_a \|u - \overline{u}\|_a^2, \end{aligned}$$

which concludes the proof. ∎

In the case $d = 2$, it is possible to give a result to a solution having a tangential displacement with a constant sign on Γ_C. We will say that a tangential displacement $u_T \in X_T$ is strictly positive if $\langle \mu_T, u_T \rangle_{X'_T, X_T} > 0$ for all $\mu_T \in X'_T$ such that $\mu_T \geq 0$ (i.e., $\langle \mu_T, v_T \rangle_{X'_T, X_T} \geq 0$ for all $v_T \in X_T$, $v_T \geq 0$, a.e. on Γ_C) and $\mu_T \neq 0$.

PROPOSITION 5. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied,* $g \equiv 0$, *and* $d = 2$, *if* u *is a solution to problem* (14) *such that* $u_T > 0$ *and* $C_3 \|\mathcal{F}\|_a < 1$, *then* u *is the unique solution to problem* (14) *(when* $\mathcal{F}$ *is constant over* Γ_C, *the condition reduces to* $C_3 \mathcal{F} < 1$*).*

Proof. Let us assume that $\overline{u}$ is another solution to problem (14), with $\overline{\lambda}_N$ and $\overline{\lambda}_T$ the corresponding contact stresses on Γ_C. Then from Lemma 5 one has

$$\|\overline{u} - u\|_a^2 \leq \langle \zeta - \lambda_T, \overline{u}_T - u_T \rangle_{X'_T, X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\overline{\lambda}_N, u_T).$$

Because $u_T > 0$, one has $\lambda_T = \mathcal{F}\lambda_N$ and $-\partial_2 j(\mathcal{F}\overline{\lambda}_N, u_T)$ contains $\mathcal{F}\overline{\lambda}_N$. Thus, taking $\zeta = \mathcal{F}\overline{\lambda}_N$, one obtains

$$\|\overline{u} - u\|_a^2 \leq \langle \mathcal{F}(\overline{\lambda}_N - \lambda_N), \overline{u}_T - u_T \rangle_{X'_T, X_T} \leq \|\overline{\lambda} - \lambda\|_{-a,\Gamma_C} \|\mathcal{F}(\overline{u} - u)\|_a \leq \|\mathcal{F}\|_a \|\overline{u} - u\|_a^2,$$

which implies $\overline{u} = u$ when $\|\mathcal{F}\|_a < 1$. ∎

Of course, the same reasoning is valid for $u_T < 0$.

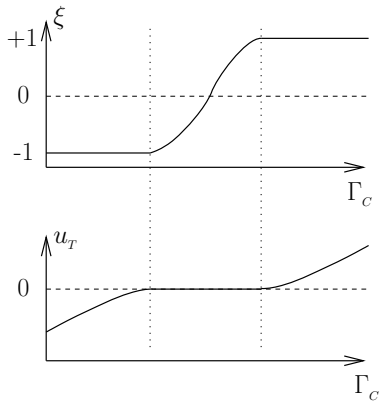


FIG. 4. *Example of tangential displacement u_T and a possible corresponding multiplier ξ for $d = 2$.*

Let us now define the space of multipliers $M(X_T \to X_N)$ of the functions $\xi : \Gamma_C \to \mathbb{R}^d$ such that $\xi \cdot \mathrm{n} = 0$ a.e. on Γ_C and such that the following two equivalent norms are finite:

$$\|\xi\|_{M(X_T \to X_N)} = \sup_{\substack{v_T \in X_T \\ v_T \neq 0}} \frac{\|\xi \cdot v_T\|_{X_N}}{\|v_T\|_{X_T}} \quad \text{and} \quad \|\xi\|_a = \sup_{\substack{v_T \in X_T \\ v_T \neq 0}} \frac{\|\xi \cdot v_T\|_{a,\Gamma_C}}{\|v_T\|_{a,\Gamma_C}}.$$

Because Γ_C is assumed to have the $\mathcal{C}^1$ regularity, $M(X_T \to X_N)$ is isomorphic to $(MX_N)^{d-1}$.

It is possible to give a more general result assuming that $\lambda_T = \mathcal{F}\lambda_N\xi$, with $\xi \in M(X_T \to X_N)$. It is easy to see that this implies that $|\xi| \leq 1$ a.e. on the support of λ_N and, more precisely, that $\xi \in \mathrm{Dir}_T(u_T)$ a.e. on the support of λ_N, where $\mathrm{Dir}_T(.)$ is the subderivative of the convex map $\mathbb{R}^d \ni x \longmapsto |x_T|$. This means that it is reasonable to assume that $\xi \in \mathrm{Dir}_T(u_T)$ a.e. on Γ_C.

PROPOSITION 6. *Assuming hypotheses* (10), (11), *and* (13) *are satisfied and $g \equiv 0$, if u is a solution to problem* (14) *such that $\lambda_T = \mathcal{F}\lambda_N\xi$, with $\xi \in M(X_T \to X_N)$, $\xi \in \mathrm{Dir}_T(u_T)$ a.e. on Γ_C, and $C_3\|\mathcal{F}\|_a\|\xi\|_a < 1$, then u is the unique solution to problem* (14).

Proof. Let us assume that $\overline{u}$ is another solution to problem (14), with $\overline{\lambda}_N$ and $\overline{\lambda}_T$ the corresponding contact stresses on Γ_C. Then from Lemma 5 one has

$$\|\overline{u} - u\|_a^2 \leq \langle \zeta - \lambda_T, \overline{u}_T - u_T \rangle_{X_T', X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\overline{\lambda}_N, u_T).$$

Then, a possible choice is $\zeta = \mathcal{F}\overline{\lambda}_N\xi$, which, together with the fact that $\|\mathcal{F}\xi\|_a \leq \|\mathcal{F}\|_a\|\xi\|_a$, gives

$$\begin{aligned}\|\overline{u} - u\|_a^2 \leq \langle \mathcal{F}\xi(\overline{\lambda}_N - \lambda_N), \overline{u}_T - u_T \rangle_{X_T', X_T} &\leq C_3\|\mathcal{F}\|_a\|\xi\|_a\|\overline{\lambda} - \lambda\|_{-a,\Gamma_C}\|\overline{u} - u\|_a \\ &\leq C_3\|\mathcal{F}\|_a\|\xi\|_a\|\overline{u} - u\|_a^2,\end{aligned}$$

which implies $\overline{u} = u$ when $C_3\|\mathcal{F}\|_a\|\xi\|_a < 1$. ∎

Remark 6. Using equivalences of norms, one can deduce that a more restrictive condition than $C_3\|\mathcal{F}\|_a\|\xi\|_a < 1$ is the condition $\|\mathcal{F}\|_{MX_N}\|\xi\|_{M(X_T \to X_N)} < \frac{\sqrt{\alpha}}{C_1C_3\gamma\sqrt{M}}$.

As illustrated in Figure 4, for $d = 2$, the multiplier ξ has to vary from -1 to $+1$ each time the sign of the tangential displacement changes from negative to positive.

The set $M(X_T \to X_N)$ does not contain any multiplier having a discontinuity of the first kind. This implies that in order to satisfy the assumptions of Proposition 6 the tangential displacement of the solution u cannot pass from a negative value to a positive value, being zero on only a single point of Γ_C.

Perspectives. As far as we know, the result given by Propositions 4, 5, and 6 are the first results dealing with the uniqueness of the solution to the Coulomb problem without considering a regularization of the contact or the friction law. In the future, it may be interesting to investigate the following open problems.

Is it possible to prove that, for a sufficiently regular domain and a sufficiently regular loading, a solution of the Coulomb problem is necessarily such that $\lambda_T = \mathcal{F}\lambda_N\xi$ with $\xi \in M(X_T \to X_N)$? This could be a way to prove a uniqueness result for a sufficiently small friction coefficient and regular loadings.

The more the tangential displacement u_T oscillates around 0 (i.e., the more u_T changes its sign for $d = 2$), the more the multiplier ξ varies and thus the greater $\|\xi\|_{M(X_T \to X_N)}$ is. Does it mean that a multisolution for an arbitrary small friction coefficient and a fixed geometry has to be searched with very oscillating tangential displacement (necessarily for all the solutions)?

Finally, the convergence of finite element methods in the uniqueness framework given by Proposition 6 will be presented in [10].

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, London, 1975.

[2] L.-E. Andersson, *Existence results for quasistatic contact problems with Coulomb friction*, Appl. Math. Optim., 42 (2000), pp. 169–202.

[3] G. Duvaut, *Problèmes unilatéraux en mécanique des milieux continus*, in Actes du Congrès International des Mathématiciens (Nice 1970), Tome 3, Gauthier-Villars, Paris, 1971, pp. 71–77.

[4] G. Duvaut and J. L. Lions, *Les Inéquations en Mécanique et en Physique*, Dunod, Paris, 1972.

[5] C. Eck and J. Jarušek, *Existence results for the static contact problem with Coulomb friction*, Math. Models Methods Appl. Sci., 8 (1998), pp. 445–468.

[6] R. Hassani, P. Hild, I. Ionescu, and N.-D. Sakki, *A mixed finite element method and solution multiplicity for Coulomb frictional contact*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 4517–4531.

[7] P. Hild, *An example of nonuniqueness for the continuous static unilateral contact model with Coulomb friction*, C. R. Acad. Sci. Paris, 337 (2003), pp. 685–688.

[8] P. Hild, *Non-unique slipping in the Coulomb friction model in two-dimensional linear elasticity*, Quart. J. Mech. Appl. Math., 57 (2004), pp. 225–235.

[9] P. Hild and Y. Renard, *Local uniqueness and continuation of solutions for the discrete Coulomb friction problem in elastostatics*, Quart. Appl. Math., 63 (2005), pp. 553–573.

[10] P. Hild and Y. Renard, *An error estimate for the Signorini problem with Coulomb friction approximated by finite elements*, submitted.

[11] J. Jarušek, *Contact problems with bounded friction. Coercive case*, Czechoslovak Math. J., 33 (1983), pp. 237–261.

[12] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalties and Finite Element Methods*, SIAM Stud. Appl. Math. 8, SIAM, Philadelphia, 1988.

[13] P. Laborde and Y. Renard, *Fixed point strategies for elastostatic frictional contact problems*, submitted.

[14] V. G. Maz'ya and T. O. Shaposhnikova, *Theory of Multipliers in Spaces of Differentiable Functions*, Pitman, Boston, 1985.

[15] J. Nečas, J. Jarušek, and J. Haslinger, *On the solution of variational inequality to the Signorini problem with small friction*, Boll. Un. Mat. Ital. B(5), 17 (1980), pp. 796–811.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 468–477

SCALING OF THE ENERGY FOR THIN MARTENSITIC FILMS*

NIRMALENDU CHAUDHURI† AND STEFAN MÜLLER‡

Abstract. We study the scaling behavior of thin martensitic films. Specifically we consider an elastic energy with two $SO(3)$ invariant wells which are strongly incompatible in the sense of Matos and Šverák, but whose two-dimensional projections may be compatible. We show that in a thin film of thickness h the energy per unit height scales like h. This scaling lies in between the classical membrane theory (where the energy per unit height is of order 1) and the Kirchhoff bending theory, which corresponds to a scaling of h^2.

Key words. thin films, martensitic phase transitions, variational methods for solids

AMS subject classifications. 74K15, 74N10, 49J45

DOI. 10.1137/04061581X

1. Introduction.

1.1. Main result. We study the scaling of the elastic energy for a thin film made of a multiphase material. Specifically we consider the cylindrical domain

$$\Omega_h := S \times \left(-\frac{h}{2}\,,\,\frac{h}{2}\right) \subset \mathbb{R}^3, \tag{1}$$

an elastic deformation

$$v : \Omega_h \to \mathbb{R}^3, \tag{2}$$

and its energy (per unit height)

$$E^h(v) := \frac{1}{h}\int_{\Omega_h} W(\nabla v(x))dx. \tag{3}$$

We suppose that the stored energy density W, which is defined on the space $\mathbb{M}^{3\times 3}$ of 3×3 matrices, is nonnegative and vanishes exactly on the set

$$K := SO(3) \cup SO(3)\,H, \quad \det H > 0, \tag{4}$$

which consists of two copies of the group $SO(3) = \{F \in \mathbb{M}^{3\times 3} : F^TF = Id_3, \det F = 1\}$ of rotations, corresponding to two preferred crystalline configurations or phases (see (7)–(9) and (10) below for the full list of assumptions on W). We are interested in low-energy deformations, and these are characterized by the fact that ∇v is close to K, except possibly on a set of small measure.

Bhattacharya and James [5] made the crucial observation that for a number of interesting materials the low-energy states are very different in three-dimensional

*Received by the editors September 27, 2004; accepted for publication (in revised form) April 26, 2005; published electronically May 26, 2006.
http://www.siam.org/journals/sima/38-2/61581.html
†Centre for Mathematics and Its Applications, Australian National University, Canberra ACT 0200, Australia (chaudhur@maths.anu.edu.au).
‡Max Planck Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (sm@mis.mpg.de). This research was partially supported by the TMR network on phase transitions in crystalline materials (FMRX-CT98-0229) and the DFG priority program on analysis, modeling and simulation of multiscale problems (SPP 1095).

(bulk) samples and in the thin film limit. If Id_3 represents the austenite (high-temperature) phase and H represents one of the martensitic phases, then these are usually incompatible in bulk; in particular, there are no nontrivial zero energy states. By contrast, the limiting thin film membrane energy

$$I_{\text{membrane}} := \int_S W_{\text{membrane}}(\nabla' v)\, dx', \quad \text{where } x' = (x_1, x_2), \quad \nabla' := (\partial_1, \partial_2), \tag{5}$$

which, roughly speaking, is the Γ-limit of E^h (see section 1.2 for a more detailed discussion), admits many nontrivial zero energy states, including lamellar arrangements of the two phases, as well as more complicated, e.g., tent-like, structures; see [5]. This drastic difference in the behavior stems from the fact that the three-dimensional compatibility requires the existence of an invariant plane (i.e., Id_3 and RH have to agree on a plane, for some $R \in SO(3)$), while two-dimensional compatibility requires only an invariant line (in the film plane), i.e., one needs a vector v in the film plane such that $v = RHv$. Suppose such a vector v exists. Then in a film of small, but finite, thickness h the juxtaposition of the deformation gradients Id_3 and RH (along a line in direction v) leads to a mismatch of the deformations of order h. Separating the two regions by a strip of width proportional to h, one sees easily that there exist three-dimensional deformations $v^{(h)}$ which have a nontrivial thin film limit and whose energy is bounded from above by Ch.

Our main result that this scaling is optimal. To state the result precisely it is convenient to introduce the rescaling $y(x) := v(x_1, x_2, hx_3)$ and the notation $\Omega := \Omega_1 = S \times (-1/2, 1/2)$. Then $y : \Omega \to \mathbb{R}^3$. As above, we write $\nabla' y = (y_{,1}, y_{,2}) = y_{,1} \otimes e_1 + y_{,2} \otimes e_2$ for the gradient in the plane, where $\otimes$ denotes the tensor product, and $\nabla_h y = (y_{,1}, y_{,2}, \frac{1}{h} y_{,3})$. Thus the elastic energy per unit height is given by

$$I^h(y) := E^h(v) = \int_\Omega W(\nabla_h y) dx\,. \tag{6}$$

We assume that W is Borel measurable and satisfies

$$W \text{ is } C^2 \text{ in a neighborhood of } K = SO(3) \cup SO(3)\, H, \tag{7}$$

$$W \text{ is frame indifferent: } W(F) = W(RF) \text{ for all } R \in K\,, \tag{8}$$

$$W(F) \geq C\text{dist}^2(F\,, K),\ C > 0, \text{ and } W(F) = 0 \text{ if } F \in K\,. \tag{9}$$

We suppose that the two wells are strongly incompatible in the sense of Matos [21] and Šverák [30]. By polar decomposition we may suppose that H is symmetric. Hence H is diagonal in an orthonormal basis (v_1, v_2, v_3), with eigenvalues $\lambda_1, \lambda_2, \lambda_3$. From now on we suppose that the film has been cut so that v_3 is the film normal. Thus, after a suitable rotation, we may suppose that H is diagonal in the standard basis (e_1, e_2, e_3). The strong incompatibility condition reads

$$\sum_{i=1}^{3} (1 - \lambda_i) \left(1 - \frac{\det H}{\lambda_i}\right) > 0. \tag{10}$$

The above Matos–Šverák [30] condition is stronger than the usual rank-one incompatibility condition. By $O(2,3)$ we denote the set of linear isometries from $\mathbb{R}^2$ to $\mathbb{R}^3$,

i.e., the set of all 3×2 matrices F with $F^T F = Id_2$. It is easy to see that the convex hull conv $O(2,3)$ of $O(2,3)$ is given by matrices with $F^T F \leq Id_2$ (in the sense of symmetric matrices), i.e., by all linear maps with Lipschitz constant less than or equal to 1. Let $\tilde{H} := \mathrm{diag}(\lambda_1, \lambda_2)$ be the two-dimensional projection of H. Then $\tilde{H}$ is compatible with the identity if one eigenvalue is bigger than (or equal to) 1 and the other eigenvalue is less than (or equal to) 1. An example for two-dimensional compatibility and three-dimensional strong incompatibility is given by the choice $H = (\lambda, \mu, \mu)$ with $\lambda < 1 < \mu$ and $(\lambda - 1)(3\mu + 1) + 2(\mu - 1) > 0$.

THEOREM 1. *Suppose that W satisfies the conditions* (7)–(9) *and* (10). *Consider a sequence $y^{(h)}$ which satisfies*

$$\frac{1}{h} I^h(y^{(h)}) \leq C \quad \text{for all } 0 < h < h_0. \tag{11}$$

Then, as $h \to 0$, there exists a subsequence (not relabeled) such that $y^{(h)} \rightharpoonup y$ $W^{1,2}(\Omega, \mathbb{R}^3)$ (weakly) and y is independent of x_3.

Moreover, $\nabla' y \in$ conv $O(2,3)$ or $\nabla' y \, \tilde{H}^{-1} \in$ conv $O(2,3)$ a.e. in S. In other words, $\mathcal{L}^2(S \setminus E_1 \cup E_2) = 0$, where

$$E_1 := \{x' \in S : (\nabla' y(x'))^T (\nabla' y(x')) \leq Id_2\},$$
$$E_2 := \{x' \in S : \tilde{H}^{-T} (\nabla' y(x'))^T (\nabla' y(x')) \tilde{H}^{-1} \leq Id_2\}.$$

In addition we have

$$\liminf_{h \to 0} \frac{1}{h} I^h(y^{(h)}) \geq C \inf\{\mathcal{H}^1(S \cap \partial^* E) : E \text{ has finite perimeter } E \subset E_1, S \setminus E \subset E_2\}. \tag{12}$$

Remarks 1. Note that the sets E_1 and E_2 need not be disjoint.

2. The situation is complicated by the fact that in the thin film limit microstructure (i.e., fine scale oscillation of the deformation gradient) can arise from two different sources: phase mixtures and loss of compactness in thin films due to the crumbling under compression, which already occurs in single phase materials (see (17) below). It is due to the crumbling that one can only assure that the limiting deformation gradient is in the *convex hull* of $O(2,3)$ (or $O(2,3)H$) rather than $O(2,3)$ itself. Even for single phase materials crumbling can only be excluded if one has the much stronger energy bound $I^h(y^{(h)}) \leq Ch^2$ (see [12, section 5]).

3. Estimate (12) assures that the scaling proportional to h is optimal if the limit involves a nontrivial *phase* mixture. Indeed if $(1/h) I^h(y^{(h)}) \to 0$, then one can choose either $E_1 = S$ or $E_2 = S$ (up to a nullset); i.e., the corresponding limiting y can already be reached by a single phase material (see next subsection).

4. If the two-dimensional projections of the wells are incompatible, then either $\lambda_1 < 1$ and $\lambda_2 < 1$ or $\lambda_1 > 1$ and $\lambda_2 > 1$. It suffices to consider the former case. Then $E_1 \supset E_2$. Thus we have $\nabla' y \in \mathrm{conv}\, O(2,3)$ a.e., whenever (11) holds. Conversely, if $\nabla' y \in \mathrm{conv}\, O(2,3)$, then by a recent result of Conti and Maggi [9] there exist maps $y^{(h)} \rightharpoonup y$ in $W^{1,2}$ such that $h^{-1} I^h(y^{(h)}) \to 0$; see the next subsection. Thus for two-dimensional incompatible wells the admissible limiting deformations can always be reached by using a single well and crumbling.

5. The scaling $I^h \sim h$ is an unconventional one in terms of classical membrane and plate theories; it lies in between the scaling for membranes ($I^h \sim 1$) and Kirchhoff plates ($I^h \sim h^2$); see the next subsection for details.

6. The Γ-limit of the scaled functionals $(1/h)I^h$ is not known. One difficulty is that, in contrast to many other situations, one cannot expect the Γ-limit to be independent of the boundary conditions. Thus many of the usual cut-and-paste arguments do not apply; see the next subsection for further comments. For rods the corresponding Γ-limit is known; see [22].

1.2. Mathematical context. To put the result above in context, we very briefly review the theory of thin film limits for a single phase material, i.e., for energy functions W which satisfy the coercivity condition

$$W(F) \geq C \operatorname{dist}^2(F, SO(3)), \quad C > 0, \tag{13}$$

instead of (9) (for a more extended review with further references, see [14]).

The derivation of lower dimensional theories of elasticity from the three-dimensional theory has a very long and distinguished history which dates back to the beginning of elasticity theory. Rigorous results, starting from *nonlinear* elasticity, however, have only been obtained since the early 90s, beginning with the work of LeDret and Raoult [18, 19]. They showed, under an additional growth hypothesis on W from above, that the Γ-limit (with respect to strong L^2 convergence in $W^{1,2}(\Omega, \mathbb{R}^3)$) of the functionals I^h exists and is given by

$$I_{\text{membrane}} := \begin{cases} \int_S W_{\text{membrane}}(\nabla' v)\, dx' & \text{if } v_{,3} = 0, \\ \infty & \text{else.} \end{cases} \tag{14}$$

The membrane energy can be computed in two steps. First one minimizes out the derivatives in the third component (corresponding to the third column of F) and defines the energy W_2 on 3×2 matrices by

$$W_2(G) = \min\{W(G + b \otimes e_3) : b \in \mathbb{R}^3\}. \tag{15}$$

Then W_{membrane} is given as the quasi-convex hull of W_2, i.e., by minimizing out over all possible fine-scale oscillations:

$$W_{\text{membrane}}(G) := \inf \left\{ \int_{(0,1)^2} W_2(G + \nabla' \varphi)\, dx' : \varphi \in C_0^\infty((0,1)^2; \mathbb{R}^3) \right\}. \tag{16}$$

For a single well material (i.e., if W vanishes on $SO(3)$ and satisfies (13)) the reduced energy W_2 vanishes on $O(2,3)$ and

$$W_{\text{membrane}}(G) = 0 \quad \Leftrightarrow \quad G \in \operatorname{conv} O(2,3) \tag{17}$$

(the convex, quasi-convex, and rank-one convex hull of $O(2,3)$ agree). Thus the membrane energy is fully degenerate for compressions, which agrees with the physical intuition that a membrane can withstand only tension but not compression. Based on this intuition a so-called tension field theory for membranes has been used in the engineering literature for a long time [32, 28]. Pipkin [25, 26] has shown that tension-field theory arises naturally as a consequence of relaxation and quasi-convexification.

We leave for a moment the case of single phase materials to mention that the limit considered by Bhattacharya and James [5] is slightly different from that studied by LeDret and Raoult. Bhattacharya and James added a regularizing higher gradient perturbation $\kappa^2 |\nabla^2 v|^2$ to the integrand in (3) and (after the usual rescaling

$y(x_1, x_2, x_3) = v(x_1, x_2, hx_3))$ pass to the limit $h \to 0$ for fixed $\kappa > 0$. They thus obtain a limiting two-dimensional energy which involves $W_2(\nabla' y)$ (plus a higher gradient contribution involving κ) rather than $W_{\text{membrane}}(\nabla' y)$. If one takes the limit $\kappa \to 0$ in the Bhattacharya–James limiting energy, then one recovers (14). Shu [29] has shown that one also obtains (14) if one considers a $\kappa(h)$ with $\lim_{h\to 0} \kappa(h) = 0$; in fact, he gives a detailed analysis of a variety of multiparameter limits related to the scale of material heterogeneities in the tangential and normal directions (corresponding, e.g., to polycrystals or multilayers).

Let us now return to single phase materials, i.e., those for which W vanishes on $SO(3)$ and satisfies (13). For those materials it has recently become possible to study also the Γ-limit of the rescaled functionals $h^{-\beta} I^h$ and to derive a full hierarchy of plate theories. For $\beta = 2$ one obtains Kirchhoff's geometrically nonlinear bending theory [11, 12, 23, 24]. This theory imposes the isometry constraint $\nabla' y \in O(2,3)$, and the energy is given by a quadratic expression in the curvature, more precisely in the second fundamental form $A = -(\nabla')^2 y \cdot \nu$, where $\nu := y_{,1} \wedge y_{,2}$ is the normal to the deformed surface. For $\beta = 4$ one obtains the von Kármán plate theory [13]; in fact, the full range of exponents $\beta \geq 2$ is now understood [14].

In contrast, relatively little is known in the range $0 < \beta < 2$. Conti [8] has recently shown that for $0 < \beta < 1$ the Γ-limit of

$$\frac{1}{h^\beta}\left[I^h(y) - \int_\Omega h^\beta f(x') \cdot y(x)\, dx\right] \tag{18}$$

is given by

$$J(y) = \begin{cases} -\int_S f \cdot y\, dx' & \text{if } y_{,3} = 0, \quad \nabla' y \in \operatorname{conv} O(2,3), \\ \infty & \text{else.} \end{cases} \tag{19}$$

The range $1 \leq \beta < 2$ is largely unexplored in terms of rigorous analysis. In the context of delamination and blistering of thin films [15] one is lead to the study of compressive *Dirichlet* boundary conditions such as $y^{(h)}(x', x_3) = (\lambda x', h x_3)$ on $\partial S \times I$, with $0 \leq \lambda < 1$. If this boundary condition is imposed, one can show that $ch \leq \inf I^h(y^{(h)}) \leq Ch$, with $c > 0$; see [4] (as well as [17, 3] for related work). For the extension to anisotropic boundary compression, see [8]. The Γ-limit of $h^{-1} I^h$ is not known.

If instead of Dirichlet boundary conditions we assume only that $y^{(h)} \rightharpoonup (\lambda x', 0)$ in $W^{1,2}$, then much less is known. Venkataramani has constructed maps with periodic boundary conditions whose energy scales like $h^{5/3}$ [31]. His construction shows that for $\lambda = 0$ one can achieve an energy bound $Ch^{5/3}$. Conti and Maggi [9] have generalized this construction to a much larger class of limit maps. They have also shown that every short map (i.e., every map satisfying $(\nabla' y)^T \nabla' y \leq Id_2$) can be approximated in L^∞ (and weakly in $W^{1,2}$) by maps y_h with energy bounded by $Ch^{5/3-\varepsilon}$. The scaling exponent $h^{5/3}$ has been suggested in the physics literature on crumbling as a natural exponent based on a formal scaling argument and an assumed equipartition of bending and stretching energy [20, 10] (see also [2, 6]; for complex folding patterns at free boundaries and their potential relevance for certain growth models in biology, see as well [27, 1]). For a single ridge with well-defined boundary conditions Venkataramini recently showed that the energy scales indeed like $h^{5/3}$ [31].

2. Proof of the lower bound. The key ingredient in the proof is the following rigidity result.

THEOREM 2 (see [7, Theorem 2]). *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain, $n \geq 2$ and $K := SO(n) \cup SO(n)\, H$, where $H = \mathrm{diag}(\lambda_1, \cdots \lambda_n)$, $\lambda_i > 0$, such that $\sum_{i=1}^{n} (1 - \lambda_i)\,(1 - \det H/\lambda_i) > 0$. There exists a positive constant $C(\Omega, H)$ with the following property. For each $u \in W^{1,2}(\Omega, \mathbb{R}^n)$ there is an associated $R := R(u, \Omega) \in K$ such that*

$$\|\nabla u - R\|_{L^2(\Omega)} \leq C(\Omega, H)\, \|\mathrm{dist}(\nabla u\,, K\,)\|_{L^2(\Omega)}\,. \tag{20}$$

The above rigidity estimate is proved in [7], but for the convenience of the readers we give a brief outline of the proof. Suppose $\|\mathrm{dist}(\nabla u\,, K\,)\|_{L^2(\Omega)}$ is small. Consider a smooth uniformly convex function F on $\mathbb{M}^{3\times 3}$ with quadratic growth whose gradient ∇F is the cofactor map when restricted to the set K (such a function exists; see [21]). We then decompose a given Sobolev function u into a part w which satisfies the nonlinear elliptic system of PDEs $\mathrm{div}\, \nabla F(\nabla w) = 0$ with $w = u$ on the boundary $\partial\Omega$. Then w satisfies the standard $W^{2,2}_{\mathrm{loc}}$ estimate. Exploiting that $\nabla(F - \det) = 0$ on K and the uniform monotonicity of ∇F, one can easily prove that the $W^{1,2}$ norm of the remainder $u - w$ is small. By a separation lemma (Lemma 2.4, [7]) and a careful covering argument we prove that ∇w cannot oscillate too much between the wells $SO(n)$ and $SO(n)H$. This shows that ∇w is L^2-close to a single well. Finally the rigidity estimate for a single well energy [12] yields (20).

We note that the inequality (20) is invariant under uniform scaling and translation of the domain; i.e., the same value of C serves for $\lambda\Omega + c$, and the rescaled function $\lambda v((x - c)/\lambda)$ may be associated with the same choice of $R \in K$.

Proof of Theorem 1. Suppose first that $S = (0, 1)^2$ and $h = 1/N$, $N \in \mathbb{N}$. Divide S into squares $S_{a,h}$ of side h with center at $a.$, i.e.,

$$S_{a,h} := a + \left(-\frac{h}{2}, \frac{h}{2}\right)^2, \quad a \in \mathbb{Z}^2_h, \qquad \mathbb{Z}_h := \left\{\frac{h}{2}, \frac{3h}{2}, \ldots, 1 - \frac{h}{2}\right\}. \tag{21}$$

Then $S = \cup S_{a,h}$, up to a set of measure zero. Let us define, $v^{(h)} : \Omega_h \to \mathbb{R}^3$ by $v^{(h)}(x_1, x_2, hx_3) := y^{(h)}(x_1, x_2, x_3)$ for all $x = (x_1, x_2, x_3) \in \Omega := \Omega_1$. Now we apply the rigidity Theorem 3 for $v^{(h)}$ for the domain $S_{a,h} \times \left(-\frac{h}{2}, \frac{h}{2}\right)$ to obtain a constant $C > 0$ (independent of a and h), and $R^h_a \in K = SO(3) \cup SO(3)\, H$, such that

$$\int_{S_{a,h}\times\left(-\frac{h}{2}, \frac{h}{2}\right)} |\nabla v^{(h)}(z) - R^h_a|^2 dz \leq C \int_{S_{a,h}\times\left(-\frac{h}{2}, \frac{h}{2}\right)} \mathrm{dist}^2(\nabla v^{(h)}(z), K) dz. \tag{22}$$

This yields

$$\int_{S_{a,h}\times\left(-\frac{1}{2}, \frac{1}{2}\right)} |\nabla_h y^{(h)}(x) - R^h_a|^2 dx \leq C \int_{S_{a,h}\times\left(-\frac{1}{2}, \frac{1}{2}\right)} \mathrm{dist}^2(\nabla_h y^{(h)}, K) dx, \tag{23}$$

where, as before, $\nabla_h y^{(h)} := \left(\nabla' y^{(h)}, \frac{1}{h} y_{,3}\right)$.

Define the piecewise constant map $R^h : S \to K$ by $R^h := R^h_a$ in $S_{a,h}$. Summing (23) over all $S_{a,h}$ and using (9) and (11), we obtain

$$\begin{aligned} \int_{S\times\left(-\frac{1}{2}, \frac{1}{2}\right)} |\nabla_h y^{(h)}(x) - R^h(x')|^2 dx &\leq C \int_{S\times\left(-\frac{1}{2}, \frac{1}{2}\right)} \mathrm{dist}^2(\nabla_h y^{(h)}(x), K) dx \\ &\leq C\, h\,. \end{aligned} \tag{24}$$

Thus

$$\text{(25)}\qquad \begin{cases} \nabla_h y^{(h)} - R^h \to 0 \quad \text{strongly in } L^2(\Omega, \mathbb{M}^{3\times 3}), \\ \nabla_h y^{(h)} \rightharpoonup (\nabla' y, b) \quad \text{weakly in } L^2(\Omega, \mathbb{M}^{3\times 3}), \\ R^h \rightharpoonup \overline{R} \quad \text{weakly in } L^2(S, \mathbb{M}^{3\times 3}). \end{cases}$$

From (25) we have $\overline{R} = (\nabla' y, b)$, and hence y is independent of x_3.

Let $\epsilon > 0$ be sufficiently small. We divide the family of squares $S_{a,h}$ into three different groups $\mathcal{A}_i$, $i = 0,\ 1,\ 2$, in the following manner:

$$\text{(26)}\qquad a \in \mathcal{A}_0 \quad \text{if and only if } \int_{S_{a,h}\times I} W(\nabla_h y^{(h)}(x))dx \geq \epsilon\, h^2,$$

where $I := \left(-\frac{1}{2}, \frac{1}{2}\right)$. If $a \notin \mathcal{A}_0$, Theorem 2 yields

$$\text{(27)}\qquad \frac{1}{h^2}\int_{S_{a,h}\times I} |\nabla_h y^{(h)}(x) - R^h_a|^2 dx \leq \frac{C}{h^2}\int_{S_{a,h}\times I} W(\nabla_h y^{(h)}(x))dx \leq C\epsilon\,.$$

Now define

$$\text{(28)}\qquad a \in \mathcal{A}_1 \quad \text{if and only if (27) holds for } R^h_a \in SO(3)\,,$$

$$\text{(29)}\qquad a \in \mathcal{A}_2 \quad \text{if and only if (27) holds for } R^h_a \in SO(3)\,H\,.$$

Clearly $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$. Thus the sets

$$\text{(30)}\qquad \Omega^h_i := \mathrm{int} ft(\cup_{a\in\mathcal{A}_i}\ \bar{S}_{a,h}),$$

$i = 0,\ 1,\ 2$, are disjoint and cover S.

Note also that the area of the set Ω^h_0 is bounded by Ch in view of (11). We now would like to estimate the length of the boundary $\partial\Omega^h_1$. Clearly this boundary consists of a union of vertical and horizontal segments of lengths h. The main observation is that each such boundary segment must also be in the boundary of one of the squares in Ω^h_0 (see Lemma 3 below). Then a simple counting argument yields that the length of $\partial\Omega^h_1$ is bounded by a constant independent of h.

To state the argument precisely we introduce the following notation. Let $e_\pm := (0, \pm 1)$ and $\tilde{e}_\pm = (\pm 1, 0)$. Then the boundary $\partial S_{a,h}$ of the square $S_{a,h}$ consists of four segments, namely the top and bottom boundaries $\partial^{e_\pm} S_{a,h} := a + \left(-\frac{h}{2}, \frac{h}{2}\right) \times \left\{\pm\frac{h}{2}\right\}$ and the left and right boundaries $\partial^{\tilde{e}_\pm} S_{a,h} := a + \left\{\pm\frac{h}{2}\right\} \times \left(-\frac{h}{2}, \frac{h}{2}\right)$. Let us denote the line segment of the boundary of $\partial S_{a,h}$ in the direction of the vector $e \in \{(\pm 1, 0),\ (0, \pm 1)\}$ by $\partial^e S_{a,h}$. Thus the boundary of Ω^h_1 is the union of the line segments $\partial^e S_{a,h} \subset \partial S_{a,h}$ for $S_{a,h} \subset \Omega^h_1$ such that $\partial^e S_{a,h} \cap \Omega^h_1 = \emptyset$. In other words,

$$\text{(31)}\qquad \partial\Omega^h_1 = \bigcup_{\substack{a\in\mathcal{A}_1 \\ \partial^e S_{a,h}\cap\Omega^h_1 = \emptyset}} \partial^e S_{a,h}\,.$$

LEMMA 3. *Let $a \in \mathcal{A}_1$ and $\partial^e S_{a,h} \subset \partial\Omega^h_1 \setminus \partial S$ for some $e \in \{(\pm 1, 0),\ (0, \pm 1)\}$. Then $S_{b,h} \subset \Omega^h_0$ for $b := a + h\,e$.*

Proof of Lemma 3. Since $\partial^e S_{a,h} \subset \partial\Omega^h_1 \setminus \partial S$, the square $S_{b,h}$ belongs to S. Suppose $b \notin \mathcal{A}_0$ and apply Theorem 2 to the domain $(S_{a,h} \cup S_{b,h}) \times \left(-\frac{h}{2}, \frac{h}{2}\right)$ to

obtain a constant $C > 0$ (independent of h, a, and b) and a matrix R^h_{ab} such that

$$\frac{1}{h^2}\int_{(S_{a,h}\cup S_{b,h})\times\left(-\frac{1}{2},\frac{1}{2}\right)} |\nabla_h y^{(h)} - R^h_{ab}|^2 dx \le \frac{C}{h^2}\int_{(S_{a,h}\cup S_{b,h})\times\left(-\frac{1}{2},\frac{1}{2}\right)} W(\nabla_h y^{(h)})\,dx \le 2C\,\epsilon\,. \tag{32}$$

Since $S_{a,h} \subset \Omega^h_1$, there exists $R^h_a \in SO(3)$ such that

$$\frac{1}{h^2}\int_{S_{a,h}\times\left(-\frac{1}{2},\frac{1}{2}\right)} |\nabla_h y^{(h)} - R^h_a|^2 dx \le C\,\epsilon\,. \tag{33}$$

Therefore (32) and (33) yield $|R^h_{ab} - R^h_a| \le 4C\epsilon$. Similarly there exist $R^h_b \in SO(3)H$ such that $|R^h_{ab} - R^h_b| \le 4C\epsilon$. We thus obtain a contradiction if ϵ is chosen sufficiently small, and the proof of the lemma is finished. □

Proof of Theorem 1 *(continued).* Let $a \in \mathcal{A}_1$, and $\partial^e S_{a,h} \subset \partial\Omega^h_1 \setminus \partial S$, for some e. Then by Lemma 3, the square adjacent to the side $\partial^e S_{a,h}$ is in Ω^h_0. There can be at most four edges $\partial^e S_{a,h} \subset \partial\Omega^h_1 \setminus \partial S$, touching a single square in Ω^h_0. Thus from (31), (26), and (11) we obtain

$$\begin{aligned}\mathcal{H}^1(\partial\Omega^h_1 \setminus \partial S) &= \sum_{\substack{a\in\mathcal{A}_1\\ \partial^e S_{a,h}\subset\partial\Omega^h_1\setminus\partial S}} \mathcal{H}^1(\partial^e S_{a,h})\\ &= h\,\mathrm{card}\left\{a \in \mathcal{A}_1 \,:\, \partial^e S_{a,h} \subset \partial\Omega^h_1 \setminus \partial S\right\}\\ &\le 4h\,\mathrm{card}\mathcal{A}_0\\ &\le 4h\,\frac{1}{h^2\epsilon}\,I^h(y^{(h)})\\ &\le C\,, \end{aligned} \tag{34}$$

where "card" stands for the cardinality of a set. Hence $\chi_{\Omega^h_1}$ is bounded in $BV(S)$, functions of bounded variation on S (see Example 1.4 in [16]), and passing to a subsequence, we get $\chi_{\Omega^h_1} \overset{*}{\rightharpoonup} \chi_E$ in $BV(S)$. Therefore by the lower semicontinuity and compactness theorems for BV functions (Theorems 1.9 and 1.19, respectively, in [16]) we obtain

$$\mathrm{Per}(E) = \int_S |\nabla\chi_E| \le \liminf_{h\to 0}\int_S |\nabla\chi_{\Omega^h_1}| = \liminf_{h\to 0}\mathcal{H}^1(\partial\Omega^h_1\setminus\partial S) \le C\,\liminf_{h\to 0}\frac{1}{h}I^h(y^{(h)})\,,$$

and $\chi_{\Omega^h_1} \to \chi_E$ strongly in $L^1(S)$.

Furthermore, it follows from (9) that $\mathrm{dist}^2(F, SO(3)) \le C(W(F)+1)$. Using (11), we deduce that

$$\int_{N_h} \mathrm{dist}^2(\nabla_h y^{(h)}, SO(3))\,dx \to 0 \quad \text{whenever } \mathcal{L}^3(N_h) \to 0. \tag{35}$$

Since the map $X \mapsto \mathrm{dist}(X, M)$ is convex, whenever M is a convex set, standard

convexity and lower semicontinuity arguments yield (with $I := (-1/2, 1/2)$)

$$\begin{aligned}
\int_\Omega \chi_{E\times I}\,\mathrm{dist}^2(\nabla' y, \mathrm{conv}\,O(2,3))dx &\le \liminf_{h\to 0}\int_\Omega \chi_{E\times I}\,\mathrm{dist}^2(\nabla' y^{(h)}, \mathrm{conv}\,O(2,3))dx\\
&\le \liminf_{h\to 0}\int_\Omega \chi_{E\times I}\,\mathrm{dist}^2(\nabla_h y^{(h)}, SO(3))dx\\
&\le \liminf_{h\to 0}\int_\Omega \chi_{\Omega_1^h\times I}\,\mathrm{dist}^2(\nabla_h y^{(h)}, SO(3))dx\\
&\le \liminf_{h\to 0}\sum_{a\in\mathcal{A}_1}\int_{S_{a,h}\times I}|\nabla_h y^{(h)} - R_a^h|^2 dx\\
&\le C\,\liminf_{h\to 0}\sum_{a\in\mathcal{A}_1}\int_{S_{a,h}\times I} W(\nabla_h y^{(h)})dx\\
&\le C\,\liminf_{h\to 0} I^h(y^{(h)})\\
&= 0, \qquad (36)
\end{aligned}$$

where we used (35) to obtain the third inequality. Hence $\nabla' y \in \mathrm{conv}\,O(2,3)$ a.e. in E. Since $\mathcal{L}^2(\Omega_0^h) = h^2\,\mathrm{card}\mathcal{A}_0 \le \frac{1}{\epsilon}\,I^h(y^{(h)}) \to 0$ as $h \to 0$ we have

$$\chi_{\Omega_2^h} = 1 - \chi_{\Omega_0^h} - \chi_{\Omega_1^h} \to (1 - \chi_E) \text{ strongly in } L^1(S)\,.$$

Applying the above arguments with $\nabla' y$ replaced by $(\nabla' y)\tilde{H}^{-1}$ and $\chi_{E\times I}$ replaced by $1 - \chi_{E\times I}$, we conclude similarly that $(\nabla' y)\tilde{H}^{-1} \in \mathrm{conv}\,O(2,3)$ a.e. in $S \setminus E$. Thus $E \subset E_1$, $S \setminus E \subset E_2$, and

$$\mathrm{Per}(E) \le C \liminf \frac{1}{h} I^h(y^{(h)})\,.$$

This finishes the proof for S being the unit square and $1/h \in \mathbb{N}$. For a general bounded open set S the assertion follows similarly by first considering the subset S^h which consists of the union of all squares $S_{a,h}$ that are contained in S. ∎

Acknowledgment. This work was initiated when N.C. held a position at the Max-Planck Institute for Mathematics in the Sciences, Germany.

REFERENCES

[1] B. Audoly and A. Boudaoud, *Self-similar structures near boundaries in strained systems*, Phys. Rev. Lett., 91 (2003), paper 086105.

[2] M. Ben Amar and Y. Pomeau, *Crumpled paper*, Proc. Roy. Soc. London A, 453 (1997), pp. 729–755.

[3] H. Ben Belgacem, S. Conti, A. DeSimone, and S. Müller, *Rigorous bounds for the Föppl–von Kármán theory of isotropically compressed plates*, J. Nonlinear Sci., 10 (2000), pp. 661–685.

[4] H. Ben Belgacem, S. Conti, A. DeSimone, and S. Müller, *Energy scaling of compressed elastic films—Three-dimensional elasticity and reduced theories*, Arch. Ration. Mech. Anal., 164 (2002), pp. 1–37.

[5] K. Bhattacharya and R. D. James, *A theory of thin films of martensitic materials with applications to microactuators*, J. Mech. Phys. Solids, 47 (1999), pp. 531–576.

[6] E. Cerda, S. Chaieb, F. Melo, and L. Mahadevan, *Conical dislocations in crumpling*, Nature, 401 (1999), pp. 46–49.

[7] N. Chaudhuri and S. Müller, *Rigidity estimate for two incompatible wells*, Calc. Var. Partial Differential Equations, 19 (2004), pp. 379–390.

[8] S. Conti, *Low Energy Deformations of Thin Elastic Plates: Isometric Embeddings and Branching Patterns*, Habilitation thesis, University of Leipzig, Leipzig, Germany, 2003.
[9] S. Conti and F. Maggi, in preparation.
[10] B. A. DiDonna, S. C. Venkataramani, T. A. Witten, and E. M. Kramer, *Singularities, structures and scaling in deformed elastic m-sheets*, Phys. Rev. E., 65 (2002), paper 016603.
[11] G. Friesecke, R. D. James, and S. Müller, *Rigorous derivation of nonlinear plate theory and geometric rigidity*, C. R. Acad. Sci. Paris Sér. I, 334 (2002), pp. 173–178.
[12] G. Friesecke, R. D. James, and S. Müller, *A theorem on geometric rigidity and the derivation of nonlinear plate theory from three dimensional elasticity*, Comm. Pure Appl. Math., 55 (2002), pp. 1461–1506.
[13] G. Friesecke, R. D. James, and S. Müller, *The Föppl–von Kármán plate theory as a low energy Γ-limit of nonlinear elasticity*, C. R. Math. Acad. Sci. Paris Sér. I, 335 (2002), pp. 201–206.
[14] G. Friesecke, R. D. James, and S. Müller, *A hierarchy of plate models derived from nonlinear elasticity by Gamma-convergence*, Arch. Ration. Mech. Anal., to appear.
[15] G. Gioia and M. Ortiz, *Delamination of compressed thin films*, Adv. Appl. Mech., 33 (1997), pp. 119–192.
[16] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, Basel, Stuttgart, 1984.
[17] W. M. Jin and P. Sternberg, *Energy estimates for the von Kármán model of thin film blistering*, J. Math. Phys., 42 (2001), pp. 192–199.
[18] H. LeDret and A. Raoult, *Le modèle de membrane non linéaire comme limite variationelle de l'élasticité non linéaire tridimensionelle*, C. R. Acad. Sci. Paris Sér. I, 317 (1993), pp. 221–226.
[19] H. LeDret and A. Raoult, *The nonlinear membrane model as a variational limit of nonlinear three-dimensional elasticity*, J. Math. Pures Appl., 73 (1995), pp. 549–578.
[20] A. E. Lobkovsky, *Boundary layer analysis of the ridge singularity in a thin plate*, Phys. Rev. E., 53 (1996), pp. 3750–3759.
[21] J. P. Matos, *Young measures and the absence of fine microstructures in a class of phase transitions*, European J. Appl. Math., 6 (1992), pp. 31–54.
[22] M. G. Mora and S. Müller, *Derivation of a rod theory for multiphase materials*, online at http://www.mis.mpg.de/preprints/2005/prepr2005_40.html.
[23] O. Pantz, *Une justification partielle du modèle de plaque en flexion par Γ-convergence*, C. R. Acad. Sci. Paris Sér. I, 332 (2001), pp. 587–592.
[24] O. Pantz, *On the justification of the nonlinear inextensional plate model*, Arch. Ration. Mech. Anal., 167 (2003), pp. 179–209.
[25] A. C. Pipkin, *The relaxed energy density for isotropic elastic membranes*, IMA J. Appl. Math., 36 (1986), pp. 85–99.
[26] A. C. Pipkin, *Continuously distributed wrinkles in fabrics*, Arch. Ration. Mech. Anal., 95 (1986), pp. 93–115.
[27] E. Sharon, B. Roman, M. Marder, G. S. Shin, and H. L. Swinney, *Mechanics: Buckling cascades in free sheets—Wavy leaves may not depend only on their genes to make their edges crinkle*, Nature, 419 (2002), p. 579.
[28] E. Reissner, *On tension field theory*, 1938 paper reprinted in *Selected Works in Applied Mechanics and Mathematics*, Jones and Bartlett, London, 1996, pp. 134–146.
[29] Y. C. Shu, *Heterogeneous thin films of martensitic materials*, Arch. Ration. Mech. Anal., 153 (2000), pp. 39–90.
[30] V. Šverák, *On the problem of two wells. Microstructure and phase transition*, in IMA Vol. Math. Appl. 54, Springer, New York, 1993, pp. 183–189.
[31] S. Venkataramani, *Lower bounds for the energy in a crumpled sheet—A minimal ridge*, Nonlinearity, 17 (2004), pp. 301–312.
[32] H. Wagner, *Ebene Blechwandträger mit sehr dünnem Steigblech*, Z. Flugtechnik u. Motorluftschiffahrt, 20 (1929), pp. 200–207, 227–233, 256–262, 279–284, 306–314.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 478–502

A GENERALIZED DYNAMICAL APPROACH TO THE LARGE TIME BEHAVIOR OF SOLUTIONS OF HAMILTON–JACOBI EQUATIONS*

ANDREA DAVINI† AND ANTONIO SICONOLFI‡

Abstract. We consider the Hamilton–Jacobi equation

$$\partial_t u + H(x, Du) = 0 \qquad \text{in } (0, +\infty) \times \mathbb{T}^N,$$

where $\mathbb{T}^N$ is the flat N-dimensional torus, and the Hamiltonian $H(x, p)$ is assumed continuous in x and strictly convex and coercive in p. We study the large time behavior of solutions, and we identify the limit through a Lax-type formula. Some convergence results are also given for H solely convex. Our qualitative method is based on the analysis of the dynamical properties of the Aubry set, performed in the spirit of [A. Fathi and A. Siconolfi, *Calc. Var. Partial Differential Equations*, 22 (2005), pp. 185–228]. This can be viewed as a generalization of the techniques used in [A. Fathi, *C. R. Acad. Sci. Paris Ser. I Math.*, 327 (1998), pp. 267–270] and [J. M. Roquejoffre, *J. Math. Pures Appl.* (9), 80 (2001), pp. 85–104]. Analogous results have been obtained in [G. Barles and P. E. Souganidis, *SIAM J. Math. Anal.*, 31 (2000), pp. 925–939] using PDE methods.

Key words. Hamilton-Jacobi equations, viscosity solutions, Aubry set

AMS subject classifications. 49L25, 35B40, 58J37, 37J50

DOI. 10.1137/050621955

1. Introduction.

This paper is about the large time behavior of the equation

$$\partial_t u + H(x, Du) = 0$$

in the flat torus $\mathbb{T}^N$. Here and in what follows, (sub-, super-) solutions are meant in the viscosity sense (see [2, 3, 14]).

The subject has been extensively investigated, first in [16], and subsequently in [11], [4], [18]. It is therefore well understood that, under suitable assumptions on H,

$$u(t, x) + ct$$

converges uniformly, for t diverging positively, to a solution v of the stationary equation

$$H(x, D\phi) = c \qquad \text{in } \mathbb{T}^N,$$

where c is the so-called critical value of the Hamiltonian, i.e., given by

$$c = \min\{a \,:\, H(x, D\phi) = a \text{ has a subsolution}\}.$$

This is also the unique value of a for which $H(x, D\phi) = a$ admits a solution on the whole torus; see [15], [13]. Any (sub) solution of the previous equation with $a = c$ will be called critical in what follows.

*Received by the editors January 4, 2005; accepted for publication (in revised form) December 5, 2005; published electronically May 26, 2006. The first author has been partially supported by MIUR through the Cofin project 2002 "Viscosity, metric and control theoretic methods in nonlinear PDEs."

http://www.siam.org/journals/sima/38-2/62195.html

†Dipartimento di Matematica Pura ed Applicata, Università degli Studi di Padova, via Belzoni, 7, 35100 Padova, Italy (davini@math.unipd.it).

‡Dipartimento di Matematica, Università di Roma "La Sapienza," P.le Aldo Moro 2, 00185 Roma, Italy (siconolf@mat.uniroma1.it).

This problem has been attacked, in the quoted literature, either by means of dynamical techniques or by using viscosity solutions methods.

The dynamical approach, which can be found in [11], [18] (see also [8], [12] for a general introduction on the subject), requires strong regularity assumptions on the Hamiltonian (C^2-regularity, strict convexity, and superlinearity at infinity in the second variable), since it is based on the analysis of the associated Hamiltonian flow. The latter is related to the solution u through the Lax–Oleinik formula. As first pointed out in [11], a crucial role is played by the Aubry set, which consists of accumulation points of the flow and is invariant.

The weakening of the conditions on H has a theoretical relevance and is also an important issue for the applications, since, for instance, the Hamiltonians appearing in control theory are not smooth.

A significant step in this direction has been performed in [4] by means of pure PDE methods. The authors were able to prove the convergence, assuming that H is only continuous and satisfies a coercivity condition. Moreover, they require the Hamiltonian to fulfill a convexity-type inequality, which includes also some nonconvex functions, but not all strictly convex Hamiltonians.

The main contribution of the present paper is to employ generalized dynamical methods to achieve the above convergence result in the presence of weak regularity of H, which is taken continuous, strictly convex, and coercive.

Our procedure yields, even in the continuous case, deeper insight into the convergence phenomenon as well as remarkably simple proofs which avoid any technicality.

The core of our argument is the discovery of some distinguished curves on the torus along which the difference between u and any critical subsolution ϕ enjoys a monotonicity property. This is a generalization of something already proved in [18] for curves of the Hamiltonian flow lying on the Aubry set. The crux is that, of course, no Hamiltonian flow can be generally defined in our setting.

We overcome this difficulty following the ideas of [13], where some aspects of the Aubry–Mather theory are extended to continuous quasi-convex Hamiltonians. Using a nonsymmetric semidistance, denoted by S, suitably related to the c-sublevel set of the Hamiltonian, it is possible to define a generalized (projected) Aubry set, say $\mathcal{A}$, and some relevant properties, holding for the classical Aubry set when H is C^2, are recovered.

Under the additional assumption that H is Lipschitz continuous with respect to x, it is proved in [13], for instance, that a multivalued dynamics can be defined on $\mathcal{A}$. We take a further step here by showing that, even for continuous H, some dynamical properties are encoded in the structure of the Aubry set. We prove indeed that through any point of $\mathcal{A}$ there passes a curve η defined on $\mathbb{R}$ and satisfying

$$S(\eta(t_1),\eta(t_2)) = \int_{t_1}^{t_2} (L(\eta,\dot\eta)+c)\ \mathrm{d}s = -S(\eta(t_2),\eta(t_1)) \quad \text{for any } t_1,t_2\in\mathbb{R},$$

where L is the Lagrangian function related to H. These are precisely the curves, called critical, satisfying the monotonicity property previously mentioned.

Beside this, we get the convergence result by exploiting, as in [18], the relaxed semilimits theory and a generalization of the fact, proved in [13], that all critical subsolutions are differentiable on $\mathcal{A}$ and have the same gradient.

An advantage of our method is that we can single out the point where the strict convexity condition on H—or, to be more precise, the C^1-regularity of L, which is an equivalent condition—is employed (see Lemma 5.2). More precisely, such a property allows us to estimate how the line integral of the Lagrangian on a critical curve varies under suitable perturbation of the parametrization; see also Remark 5.8. This is a crucial step since it is well known, as shown in an example in [4], that the simple convexity of H does not ensure, in general, the convergence phenomenon.

However, we can prove that the mere convexity is actually enough when the equilibrium points form a uniqueness set for the critical equation. This accounts for the fact that a small perturbation, in a convex Hamiltonian, can produce a passage from a convergence to a nonconvergence situation; see Example 5.10. This generalizes the results of [16], where the Hamiltonian is taken only convex, as well.

We are furthermore able to identify the limit function v through a representation formula, which involves $u(0,\cdot)$, the Aubry set, and the semidistance S. It is the critical solution coinciding on $\mathcal{A}$ with the maximal critical subsolution not exceeding $u(0,\cdot)$. This should be compared to the formula given in [11] for Hamiltonians of class C^2 using the Peierls barrier. Our formula has been exploited in [19] to perform a numerical approximation of the Aubry set.

The paper is organized as follows. In section 2 some preliminary material is collected, including definitions of the semidistance S and the generalized Aubry set, as well as some properties of the critical solutions and of the Lax–Oleinik semigroup. In section 3 we introduce, through a representation formula, a distinguished critical solution, which will be proved to be the limit of $u(t,x)+ct$ for $t \to +\infty$. Section 4 is devoted to studying the dynamical properties of the Aubry set and singling out a class of special curves covering $\mathcal{A}$. The main results are finally proved in section 5. In the appendix we show that the usual integral representation formula for the Lax–Oleinik semigroup holds also in the case when H is coercive, but not necessarily superlinear at infinity, and so the Lagrangian L is possibly infinite valued at some points of $\mathbb{T}^N \times \mathbb{R}^N$.

2. Assumptions and preliminary results. We write below a list of symbols used throughout this paper.

N	an integer number.
$B_R(x_0)$	the closed ball in $\mathbb{R}^N$ centered at x_0 of radius R.
B_R	the closed ball in $\mathbb{R}^N$ centered at 0 of radius R.
$\langle \cdot\,, \cdot \rangle$	the scalar product in $\mathbb{R}^N$.
$\lvert \cdot \rvert$	the Euclidean norm in $\mathbb{R}^N$.
$\mathbb{R}_+$	the set of nonnegative real numbers.
$\mathbb{T}^N$	the N-dimensional flat torus.
$\mathrm{C}(\mathbb{T}^N)$	the space of real-valued continuous functions on $\mathbb{T}^N$.
$\mathrm{Lip}(\mathbb{T}^N)$	the space of real-valued Lipschitz-continuous functions on $\mathbb{T}^N$.

A subset of $\mathbb{R}^k$ is called *negligible* if its k-dimensional Lebesgue measure is equal to zero. We say that a property holds *almost everywhere* (*a.e.*) on $\mathbb{R}^k$ if it holds up to a negligible subset of $\mathbb{R}^k$. Given a measurable function $\varphi : \mathbb{T}^N \to \mathbb{R}$, its L^∞-norm on $\mathbb{T}^N$ will be denoted by $\|\varphi\|_\infty$. We will write $\varphi_n \rightrightarrows \varphi$ on $\mathbb{T}^N$ to mean that the sequence of functions $(\varphi_n)_n$ uniformly converges to φ on $\mathbb{T}^N$.

By *modulus* we mean a nondecreasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$, vanishing and continuous at 0. Given a closed convex subset Z of $\mathbb{R}^k$, and $p_0 \in Z$, we define the normal cone of Z at p_0, in symbols $N_Z(p_0)$, as the set $\{q \in \mathbb{R}^N : \langle q, p_0\rangle = \max_{p\in Z}\langle q,p\rangle\}$.

We endow the flat torus $\mathbb{T}^N$ with the Riemannian metric induced by the Euclidean metric on $\mathbb{R}^N$. We recall that $\mathbb{T}^N$ can be viewed as the quotient space $\mathbb{R}^N/\mathbb{Z}^N$, obtained by identifying all points $x, y \in \mathbb{R}^N$ such that $x - y \in \mathbb{Z}^N$.

With the term *curve*, without any further specification, we refer to a Lipschitz-continuous function from some given interval $[a, b]$ to $\mathbb{T}^N$. The space of all such curves is denoted by $\mathrm{Lip}([a, b], \mathbb{T}^N)$, while $\mathrm{Lip}_{x,y}([a, b], \mathbb{T}^N)$ stands for the family of curves γ joining x to y, i.e., such that $\gamma(a) = x$ and $\gamma(b) = y$, for any fixed x, y in $\mathbb{T}^N$. We denote by $W^{1,1}([a, b], \mathbb{T}^N)$ the space of absolutely continuous curves defined in $[a, b]$. Given a curve γ defined on some interval $[a, b]$, a curve γ' defined on $[a', b']$ will be called a *reparametrization of* γ if there exists an order preserving Lipschitz-continuous map $f : [a', b'] \to [a, b]$ surjective and such that $\gamma' = \gamma \circ f$. The Euclidean length of a curve γ is denoted by $\ell(\gamma)$.

Unless otherwise specified, the term (sub-, super-) solution of some PDE is understood in the viscosity sense. Given a continuous function g defined in $\mathbb{R}^k$ and $x_0 \in \mathbb{R}^k$, we denote by $D^+g(x_0)$ (resp., $D^-g(x_0)$) the superdifferential (resp., the subdifferential) of g at x_0, i.e., the (possibly empty) set made up of the differentials of the viscosity test functions from above (resp., from below) to g at x_0. Note that, in the case when g is convex, D^-g coincides with the usual subdifferential of convex analysis. When g is defined on $\mathbb{R}^m \times \mathbb{R}^k$ and $(x_0, p_0) \in \mathbb{R}^m \times \mathbb{R}^k$, we will denote by $D_p^- g(x_0, p_0)$ the subdifferential of the function $g(x_0, \cdot)$ at p_0. For a function $g : \mathbb{R}^k \to (-\infty, +\infty]$, we denote by $\mathrm{dom}(g)$ its effective domain, i.e., the subset of $\mathbb{R}^k$ where g is finite valued.

We deal with a Hamiltonian H, defined on the cotangent bundle $T^*\mathbb{T}^N$, identified with $\mathbb{T}^N \times \mathbb{R}^N$, and satisfying the following set of assumptions:

(H1) $H : \mathbb{T}^N \times \mathbb{R}^N \to \mathbb{R}$ is continuous;

(H2) $p \mapsto H(x, p)$ is convex on $\mathbb{R}^N$ for any $x \in \mathbb{T}^N$;

(H3) $\lim_{|p|\to+\infty}(\inf_{x\in\mathbb{T}^N} H(x, p)) = +\infty$;

(H4) the set of minimizers of $p \mapsto H(x, p)$ has empty interior for any $x \in \mathbb{T}^N$.

To obtain our general convergence result (see in particular Proposition 5.3, which will constitute a crucial step for that), we will, moreover, assume the following:

(H2)′ $p \mapsto H(x, p)$ is strictly convex on $\mathbb{R}^N$ for any $x \in \mathbb{T}^N$.

Notice that condition (H4) is certainly satisfied when (H2)′ holds true, since, in this case, the set of minimizers of $H(x, \cdot)$ reduces to a point for any $x \in \mathbb{T}^N$.

Remark 2.1. Exploiting the subdifferentiability properties of the function $p \mapsto H(x, p)$, for any fixed x, we see that the Lipschitz constant of such a function in B_R, for any $R > 0$, can be estimated, uniformly with respect to x, in terms of R and of $\max\{H(x, p) : (x, p) \in \mathbb{T}^N \times B_{2R}\}$; see, e.g., [17, Proposition 2.2.6].

Remark 2.2. The problem we are dealing with can be equivalently formulated in $\mathbb{R}^N$, instead of $\mathbb{T}^N$, with $\mathbb{Z}^N$-periodicity conditions.

We consider the family of Hamilton–Jacobi equations

$$H(x, D\phi) = a \qquad \text{on } \mathbb{T}^N, \tag{1}$$

with a real parameter, and set

$$c := \inf\{a \in \mathbb{R} : (1) \text{ has a subsolution}\}.$$

This is called the *critical value* of the Hamiltonian H and is characterized by the property of being the unique value for a such that (1) admits (at least) one solution (see, e.g., [15], [13]). A solution (resp., supersolution, subsolution) of

$$H(x, D\phi) = c \qquad \text{in } \mathbb{T}^N \tag{2}$$

will be called critical in what follows. Thanks to hypothesis (H3), all subsolutions of (1) are Lipschitz continuous. Moreover, by the convexity assumption, there is complete equivalence between the notions of (viscosity) subsolution and a.e. subsolution (see [2]).

Following [13], we carry out the study of properties of subsolutions to (1) by means of the semidistances S_a defined on $\mathbb{T}^N \times \mathbb{T}^N$, for $a \geq c$, as follows:

$$S_a(x,y) = \inf \left\{ \int_0^1 \sigma_a(\gamma(s), \dot\gamma(s))\, \mathrm{d}s \,:\, \gamma \in \mathrm{Lip}_{x,y}([0,1], \mathbb{T}^N) \right\}, \tag{3}$$

where $\sigma_a(x,q)$ is the support function of the a-sublevel of H, namely

$$\sigma_a(x,q) := \sup \{ \langle q, p \rangle \,:\, H(x,p) \leq a \}. \tag{4}$$

The function $\sigma_a(x,q)$ is convex in q and upper semicontinuous in x (and even continuous in all points x, where the set $\{p \in \mathbb{R}^N \,:\, H(x,p) \leq a\}$ has a nonempty interior or reduces to a point), while S_a satisfies the following properties:

$$\begin{aligned} &S_a(x,y) \leq S_a(x,z) + S_a(z,y), \\ &S_a(x,y) \leq b_a |x-y| \end{aligned}$$

for all $x, y, z \in \mathbb{T}^N$ and for some positive constant b_a. The following properties hold (see [13]).

PROPOSITION 2.3. *Given $a \geq c$, we have that*

(i) *for any $y \in \mathbb{T}^N$, the functions $S_a(y,\cdot)$ and $-S_a(\cdot,y)$ are both subsolutions of* (1);

(ii) *a function ϕ is a subsolution of* (1) *if and only if*

$$\phi(x) - \phi(y) \leq S_a(y,x) \qquad \textit{for all } x, y \in \mathbb{T}^N.$$

To ease notation, in what follows we will write S, σ in place of S_c, σ_c, respectively.

In the analysis of the behavior of critical subsolutions, a special role is played by a set $\mathcal{A}$, which has been called in [13] the *(projected) Aubry set*, defined as the collection of points $y \in \mathbb{T}^N$ such that

$$\inf \left\{ \int_0^1 \sigma(\gamma, \dot\gamma)\, \mathrm{d}s \,:\, \gamma \in \mathrm{Lip}_{y,y}([0,1], \mathbb{T}^N),\, \ell(\gamma) \geq \delta \right\} = 0 \quad \text{for some } \delta > 0,$$

or, equivalently (cf. [13, Lemma 5.1]),

$$\inf \left\{ \int_0^1 \sigma(\gamma, \dot\gamma)\, \mathrm{d}s \,:\, \gamma \in \mathrm{Lip}_{y,y}([0,1], \mathbb{T}^N),\, \ell(\gamma) \geq \delta \right\} = 0 \quad \text{for any } \delta > 0.$$

The set $\mathcal{A}$ is closed and nonempty (cf. [13, Corollaries 5.7 and 5.9]). In the next theorem we outline the main properties linking $\mathcal{A}$ to (2) (see [13]).

THEOREM 2.4.

(i) *If ϕ and w are a subsolution and a supersolution of* (2)*, respectively, and $\phi \leq w$ on $\mathcal{A}$, then $\phi \leq w$ on $\mathbb{T}^N$. In particular, if two solutions of* (2) *coincide on $\mathcal{A}$, then they coincide on $\mathbb{T}^N$.*

(ii) *If w_0 is a function defined on $C \subset \mathcal{A}$ such that*

$$w_0(x) - w_0(y) \leq S(y,x) \quad \textit{for every } x, y \in C,$$

then the function

$$w(x) := \min_{y\in C} \big(w_0(y) + S(y,x)\big) \tag{5}$$

is the maximal critical subsolution of (2) *equaling* w_0 *on* C*, and a critical solution as well.*

(iii) *If we furthermore set* $C = \mathcal{A}$ *in* (5)*, then* w *is the unique critical solution equaling* w_0 *on* $\mathcal{A}$*.*

We call $y \in \mathbb{T}^N$ an *equilibrium point* if $\min_p H(y,p) = c$. The collection of all such points will be denoted by $\mathcal{E}$. The set $\mathcal{E}$ is a (possibly empty) closed subset of $\mathcal{A}$ (cf. [13, Lemma 5.2]). This property depends on the fact that the c-sublevel $\{p : H(y,p) \le c\}$ is nonvoid and has an empty interior when $y \in \mathcal{E}$ (the latter is a consequence of (H4), and this is actually the unique point where such a condition is used). It is apparent that $c \ge \max_{x\in\mathbb{T}^N} \min_{p\in\mathbb{R}^N} H(x,p)$; we point out that $\mathcal{E}$ is nonempty if and only if the previous formula holds with an equality. In this case, $\mathcal{E}$ is made up of the points x where the maximum is attained.

Let us now focus our attention on the Cauchy problem

$$\begin{cases} \partial_t u + H(x,Du) = 0 & \text{in } (0,+\infty)\times\mathbb{T}^N, \\ u(0,x) = u_0(x) & \text{on } \mathbb{T}^N, \end{cases} \tag{6}$$

where u_0 is a continuous initial datum. The following result holds (see, e.g., [6]).

THEOREM 2.5. *Assume* H *satisfies assumptions* (H1), (H2), (H3), (H4). *Then the Cauchy problem* (6) *admits a unique uniformly continuous solution* $u(t,x)$ *on* $\mathbb{R}_+ \times \mathbb{T}^N$*, for any* $u_0 \in \mathrm{C}(\mathbb{T}^N)$*. If, moreover, the initial datum* $u_0 \in \mathrm{Lip}(\mathbb{T}^N)$*, then* $u(t,x)$ *is Lipschitz continuous on* $\mathbb{R}_+ \times \mathbb{T}^N$ *and satisfies*

$$\|Du\|_\infty \le M, \quad \|\partial_t u\|_\infty \le \operatorname{ess\,sup}\{|H| : |p| \le M\}$$

for any positive constant M *such that*

$$M > \|Du_0\|_\infty, \quad \inf\{H : |p| > M\} > \sup\{|H| : |p| \le \|Du_0\|_\infty\}. \tag{7}$$

In view of the previous theorem, we can define, for any $t > 0$, a nonlinear operator $\mathcal{S}(t)$ on $\mathrm{C}(\mathbb{T}^N)$ by setting $\mathcal{S}(t)\phi := u(t,\cdot)$ for every $\phi \in \mathrm{C}(\mathbb{T}^N)$, where $u(t,x)$ denotes the unique solution of the Cauchy problem (6) with $u_0 = \phi$. The family of operators $\big(\mathcal{S}(t)\big)_{t>0}$ forms a semigroup, whose main properties are summarized below.

PROPOSITION 2.6.

(i) (Semigroup property.) *For any* $t, s > 0$ *we have* $\mathcal{S}(t+s) = \mathcal{S}(t)\circ\mathcal{S}(s)$.

(ii) (Monotonicity property.) *For every* $\phi, \psi \in \mathrm{C}(\mathbb{T}^N)$ *and each* $t > 0$ *we have*

$$\phi \le \psi \;\Rightarrow\; \mathcal{S}(t)\phi \le \mathcal{S}(t)\psi.$$

(iii) *For any* $a \in \mathbb{R}$ *and* $\phi \in \mathrm{C}(\mathbb{T}^N)$*, we have* $\mathcal{S}(t)(\phi + a) = \mathcal{S}(t)\phi + a$.

(iv) (Nonexpansiveness.) *For each* $t > 0$*, the map* $\mathcal{S}(t)$ *is nonexpansive, i.e.,*

$$\|\mathcal{S}(t)\phi - \mathcal{S}(t)\psi\|_\infty \le \|\phi - \psi\|_\infty \quad \textit{for every } \phi, \psi \in \mathrm{C}(\mathbb{T}^N).$$

(v) *For every* $\phi \in \mathrm{C}(\mathbb{T}^N)$*, we have* $\lim_{t\to 0} \mathcal{S}(t)\phi = \phi$.

We define the *Fenchel transform* $L : \mathbb{T}^N \times \mathbb{R}^N \to (-\infty, +\infty]$ of H via

$$L(x,q) := \sup_{p\in\mathbb{R}^N} \{\langle p, q\rangle - H(x,p)\}. \tag{8}$$

The function L is called the Lagrangian related to the Hamiltonian H. We record for later use the following properties.

PROPOSITION 2.7. *Let H satisfy assumptions* (H1), (H2), (H3). *Then the following properties hold for the Lagrangian L:*

(i) *$L(x,q)$ is lower semicontinuous on $\mathbb{T}^N \times \mathbb{R}^N$, and convex in q for any fixed $x \in \mathbb{T}^N$.*

(ii) *L is continuous on $\mathrm{int}(\mathrm{dom}L) =: \Omega$.*

If, in addition, H satisfies (H2)$'$ *then*

(iii) *for every $(x,q) \in \Omega$, L is differentiable with respect to q, and $(x,q) \mapsto D_q L(x,q)$ is continuous on Ω.*

(iv) *if (x,q) is such that the supremum in the definition of $L(x,q)$ is a maximum, then (x,q) belongs to Ω.*

We refer to the appendix for the proof.

Each operator $\mathcal{S}(t)$ can be represented through the integral formula

$$\big(\mathcal{S}(t)\phi\big)(x) = \inf\left\{\phi(\gamma(0)) + \int_0^t L(\gamma(s),\dot\gamma(s))\,\mathrm{d}s \ : \ \gamma \in W^{1,1}([0,t],\mathbb{T}^N),\ \gamma(t) = x\right\} \tag{9}$$

for any $\phi \in \mathrm{C}(\mathbb{T}^N)$. The family of operators $\big(\mathcal{S}(t)\big)_{t>0}$ is called the Lax–Oleinik semigroup.

Remark 2.8. When $\phi \in \mathrm{Lip}(\mathbb{T}^N)$ and L is finite valued, the validity of (9) can be seen, for instance, by combining [10, Theorem 1.1] with Theorem 2.5. This is the case when H is uniformly superlinear in p. The infimum in (9) is then a minimum by classical results of the calculus of variations (see, e.g., [5]), and all minimizers are Lipschitz continuous (cf. [1] for some results on this topic).

We present in the appendix a proof of (9) for $\phi \in \mathrm{C}(\mathbb{T}^N)$ and general L, possibly infinite valued in some subset of $\mathbb{T}^N \times \mathbb{R}^N$, and we show the existence of minimizers in this case too.

We will use the following Tonelli-type semicontinuity theorem (see, e.g., [5, Theorem 3.6]) in the proof of Propositions 4.12 and A.6.

THEOREM 2.9. *Let J be a bounded interval of $\mathbb{R}$, and let $F : \mathbb{R}^N \times \mathbb{R}^N \to (-\infty, +\infty]$ be a function satisfying the following conditions:*

(i) *F is lower semicontinuous;*

(ii) *$F(x,\cdot)$ is convex on $\mathbb{R}^N$ for every $x \in \mathbb{R}^N$;*

(iii) *F is bounded from below by a constant.*

Then the functional

$$\mathcal{F}(\gamma) := \int_J F(\gamma(s),\dot\gamma(s))\,\mathrm{d}s$$

is sequentially weakly lower semicontinuous in $W^{1,1}(J,\mathbb{R}^N)$; i.e., if $(\gamma_k)_k$ converges weakly in $W^{1,1}(J,\mathbb{R}^N)$ to γ, then

$$\mathcal{F}(\gamma) \le \liminf_{k\to+\infty} \mathcal{F}(\gamma_k). \tag{10}$$

Equivalently, we can say that (10) *holds if $(\gamma_k)_k$ converges uniformly to γ and the measures $\nu_k(E) := \int_E |\dot\gamma_k|\,\mathrm{d}s$ are equiabsolutely continuous on J with respect to the Lebesgue measure.*

3. A distinguished critical solution. Before attacking the convergence problem, we try to guess what the asymptotic limit of $\mathcal{S}(t)u_0 + ct$ should be. We start by providing a Lax-type formula which involves the initial datum u_0, the Aubry set, and the semidistance S, and we show that this defines a critical solution, more precisely, the solution whose trace on $\mathcal{A}$ coincides with that of the maximal critical subsolution not exceeding u_0. This formula furthermore generalizes the formula given in (5).

THEOREM 3.1. *Let $w_0 : \mathbb{T}^N \to \mathbb{R}$ be any function bounded from below. Set*

$$v(x) := \inf_{y\in\mathcal{A}} \left(S(y,x) + \inf_{z\in\mathbb{T}^N} (w_0(z) + S(z,y)) \right) \quad \text{for every } x \in \mathbb{T}^N. \tag{11}$$

Then

(i) $\inf_{y\in\mathbb{T}^N} (S(y,\cdot) + w_0(y)) =: v_0$ *is the maximal critical subsolution not exceeding w_0 on $\mathbb{T}^N$.*

(ii) *the function v is the critical solution equaling v_0 on $\mathcal{A}$.*

(iii) *if the inequality $w_0(y) - w_0(x) \le S(x,y)$ holds for all x, $y \in \mathbb{T}^N$, then $v = \min_{y\in\mathcal{A}}(w_0(y) + S(y,\cdot))$ on $\mathbb{T}^N$, and $v_0 = w_0$ on $\mathcal{A}$.*

We show separately, in the next lemma, the relevant fact on which the proof of Theorem 3.1 relies.

LEMMA 3.2. *Let C be a subset of $\mathbb{T}^N$ and $w_0 : C \to \mathbb{R}$ be any function bounded from below. Then*

$$w(x) := \inf_{z\in C} \big(w_0(z) + S(z,x)\big)$$

is the maximal subsolution of (2) not exceeding w_0 on C. The function w is, moreover, a critical solution in $\mathbb{T}^N \setminus \overline{C}$ and in the whole $\mathbb{T}^N$ whenever $C \subset \mathcal{A}$.

Proof. It is easy to check, exploiting the very definition of w, that $w \le w_0$ on C and $w(x) - w(y) \le S(y,x)$ for every $x, y \in \mathbb{T}^N$. The latter inequality implies that w is a critical subsolution by Proposition 2.3. If ϕ is any critical subsolution with $\phi \le w_0$ on C then, taking into account that $\phi(x) - \phi(y) \le S(y,x)$ for every $x, y \in \mathbb{T}^N$, we get

$$\phi(x) \le \min_{z\in C} \big(\phi(z) + S(z,x)\big) \le w(x) \qquad \text{for every } x \in \mathbb{T}^N,$$

which gives the maximality of w. Such a property also implies that w is a supersolution of (2) in $\mathbb{T}^N \setminus \overline{C}$ through a standard argument (see, e.g., the proof of Proposition 3.2 in [13]). If furthermore $C \subset \mathcal{A}$, then by Theorem 2.4(ii) $w = \min_{z\in C} \big(w(z) + S(z,x)\big)$, and so it is a critical solution in $\mathbb{T}^N$. □

Proof of Theorem 3.1. Item (i) comes directly from Lemma 3.2 with $C = \mathbb{T}^N$; (ii) is therefore a consequence of Theorem 2.4(iii). Item (iii) can be finally deduced from Theorem 2.4(ii). □

The proof that the function given in formula (11), with $w_0 = u_0$, actually coincides with the asymptotic limit of $\mathcal{S}(t)u_0 + ct$ is, of course, the main goal of our analysis, and will be attained in the subsequent sections. The remainder of the present one is devoted, instead, to some preliminary remarks which give support to our guess and cast some light on our further analysis.

We start by noticing that, given a solution w of (2) and a general initial datum $u_0 \in \mathrm{C}(\mathbb{T}^N)$, there exist, since $\mathbb{T}^N$ is compact, some constants α, β such that

$$w + \alpha \le u_0 \le w + \beta \quad \text{on } \mathbb{T}^N.$$

This implies, in view of the relation $\mathcal{S}(t)w = w - ct$, which holds for every $t > 0$, and the monotonicity property of the semigroup $\big(\mathcal{S}(t)\big)_{t>0}$,

$$w + \alpha - ct \le \mathcal{S}(t)u_0 \le w + \beta - ct \quad \text{on } T^N$$

for any $t > 0$, or, in other words,

$$w + \alpha \leq \mathcal{S}(t)u_0 + ct \leq w + \beta \quad \text{on } \mathbb{T}^N. \tag{12}$$

Since the family of functions $\left(\mathcal{S}(t)u_0 + ct\right)_{t>0}$ is equicontinuous (in view of Theorem 2.5), and equibounded thanks to (12), we can define the relaxed semilimits,

$$\underline{u}(x) = \limsup_{t\to+\infty}{}^{*} \left(\mathcal{S}(t)u_0\right)(x) + ct := \sup\left\{\limsup_{n\to+\infty} \left(\mathcal{S}(t_n)u_0\right)(x_n) + ct_n\right\}, \tag{13}$$

$$\overline{u}(x) = \liminf_{t\to+\infty}{}_{*} \left(\mathcal{S}(t)u_0\right)(x) + ct := \inf\left\{\liminf_{n\to+\infty} \left(\mathcal{S}(t_n)u_0\right)(x_n) + ct_n\right\}, \tag{14}$$

where the supremum and the infimum in (13) and (14), respectively, are taken for all sequences $(x_n)_n$ converging to x and all diverging sequences $(t_n)_n$. Moreover, thanks to the uniform continuity of the function $\left(\mathcal{S}(t)u_0\right)(x)$ on $\mathbb{R}_+ \times \mathbb{T}^N$ (cf. Theorem 2.5), the sequences $(x_n)_n$ may be chosen identically equal to x, so that the following identities hold true:

$$\underline{u}(x) = \sup\left\{\psi(x) \,:\, \psi \in \omega_{\mathcal{S}}(u_0)\right\},$$
$$\overline{u}(x) = \inf\left\{\psi(x) \,:\, \psi \in \omega_{\mathcal{S}}(u_0)\right\},$$

where

$$\omega_{\mathcal{S}}(u_0) := \left\{\psi \in \mathrm{C}(\mathbb{T}^N) \,:\, \psi = \lim_{n\to+\infty} \mathcal{S}(t_n)u_0 + ct_n \text{ for some diverging sequence } (t_n)_n\right\}.$$

We have the following theorem (cf. proof of Theorem 1 in [16]).

THEOREM 3.3. *The functions $\underline{u}$ and $\overline{u}$ defined by* (13) *and* (14) *are a subsolution and a supersolution of* (2)*, respectively.*

We proceed to establish the asymptotic convergence of $\mathcal{S}(t)u_0 + ct$ to the function v given in (11) with $w_0 = u_0$, provided u_0 is a critical sub- or supersolution.

THEOREM 3.4. *Let $u_0 \in \mathrm{C}(\mathbb{T}^N)$ be either a subsolution or a supersolution of* (2)*. Then $\mathcal{S}(t)u_0 + ct$ uniformly converges, as t goes to $+\infty$, to the critical solution v defined by* (11) *with $w_0 = u_0$.*

Proof. Let us first assume u_0 to be a subsolution of (2). By Theorem 3.1(iii), v is the maximal critical subsolution satisfying $v = u_0$ on $\mathcal{A}$, and hence $v \geq u_0$ on $\mathbb{T}^N$. As $u_0 - ct$ and $v - ct$ are a subsolution and a supersolution of (2), respectively, the comparison principle yields

$$u_0 - ct \leq \mathcal{S}(t)u_0 \leq v - ct \qquad \text{on } \mathbb{T}^N,$$

and consequently, since $v = u_0$ on $\mathcal{A}$, we get

$$u_0 = S(t)u_0 + ct = v \qquad \text{on } \mathcal{A}$$

for every $t > 0$. It follows that $v = \underline{u} = \overline{u}$ on $\mathcal{A}$, and we finally deduce from Theorem 2.4(i) that $v = \underline{u} = \overline{u}$ on $\mathbb{T}^N$. This proves the assertion when u_0 is a critical subsolution.

Let us now assume u_0 to be a supersolution of (2). Let v_0 be the maximal critical subsolution not exceeding u_0 on $\mathbb{T}^N$, i.e.,

$$v_0 = \min_{y\in\mathbb{T}^N} \left(S(y,\cdot) + u_0(y)\right).$$

The maximality of v_0, combined with the fact that u_0 is a critical supersolution, implies that v_0 is a critical solution as well, so that the identity $v = v_0$ on $\mathbb{T}^N$ holds true. Arguing as in the first part of the proof, we therefore obtain

$$v \le \mathcal{S}(t)u_0 + ct \le u_0 \qquad \text{on } \mathbb{T}^N$$

for every $t > 0$. This entails $v \le \overline{u} \le \underline{u} \le u_0$ on $\mathbb{T}^N$. From the fact that $\underline{u}$ is a critical subsolution, and from the maximality property of $v_0 = v$, we get $\underline{u} \le v$, and so $v = \overline{u} = \underline{u}$ on $\mathbb{T}^N$. ∎

We deduce the following from Theorem 3.4.

PROPOSITION 3.5. *Assume $u_0 \in \mathrm{C}(\mathbb{T}^N)$, and let v be the function defined by* (11) *with $w_0 = u_0$. Then the relaxed semilimits $\underline{u}$ and $\overline{u}$, defined by* (13) *and* (14)*, respectively, satisfy*

$$v(x) \le \overline{u}(x) \le \underline{u}(x) \qquad \textit{for every } x \in \mathbb{T}^N. \tag{15}$$

Proof. Set $v_0 = \min_{y\in\mathbb{T}^N}(S(y,\cdot) + u_0(y))$. It is apparent that $v_0 \le u_0$ on $\mathbb{T}^N$, and hence, by the monotonicity property of the semigroup $(\mathcal{S}(t))_{t>0}$, we obtain $\mathcal{S}(t)v_0 + ct \le \mathcal{S}(t)u_0 + ct$ on $\mathbb{T}^N$, and (15) follows in view of Theorems 3.4 and 3.1. ∎

Proposition 3.5, Theorem 2.4, and the facts that v is a critical solution and $\underline{u}$ a critical subsolution imply that the convergence result we aim at is proved as soon as the equality $v = \underline{u}$ is obtained on $\mathcal{A}$. This suggests, in the end, that what really matters in our analysis is the asymptotic behavior of $\mathcal{S}(t)u_0 + ct$ on $\mathcal{A}$.

4. Dynamical properties of the projected Aubry set. Here we define a family of curves, called critical, fully covering the Aubry set, which will play an important role in the convergence result of the next section. We will furthermore investigate the behavior of critical subsolutions on such curves. Throughout the section, conditions (H1), (H2), (H3), (H4) are assumed.

DEFINITION 4.1. *A curve γ defined on an interval J is called* critical *if*

$$S(\gamma(t_1),\gamma(t_2)) = \int_{t_1}^{t_2} (L(\gamma,\dot\gamma) + c)\,\mathrm{d}s = -S(\gamma(t_2),\gamma(t_1))$$

for every t_1, t_2 in J with $t_2 \ge t_1$.

LEMMA 4.2. *Any critical curve is contained in the Aubry set.*

Proof. Let γ be a critical curve, which we first assume to be nonconstant, defined in some interval J. Given t_1, t_2 in J with $t_2 \ge t_1$ and $\gamma(t_1) \ne \gamma(t_2)$, we can find two sequences of curves $\gamma_n^1 \in \mathrm{Lip}_{\gamma(t_1),\gamma(t_2)}([0,1],\mathbb{T}^N)$ and $\gamma_n^2 \in \mathrm{Lip}_{\gamma(t_2),\gamma(t_1)}([0,1],\mathbb{T}^N)$, which approximate the semidistance S of their end points up to $1/n$, for any n. The cycles γ_n, obtained by juxtaposition of γ_n^1 and γ_n^2 and change of parametrization to $[0,1]$, are of length $\ell(\gamma_n) \ge 2|\gamma(t_2) - \gamma(t_1)|$ and satisfy, by Definition 4.1,

$$\lim_n \int_0^1 \sigma(\gamma_n(s),\dot\gamma_n(s))\,\mathrm{d}s = 0,$$

which shows that $\gamma(t_1)$, $\gamma(t_2)$ are in $\mathcal{A}$. If, on the contrary, the support of γ is reduced to a point, say x_0, we find

$$\int_J (L(x_0,0) + c)\,\mathrm{d}s = 0,$$

which implies $x_0 \in \mathcal{E} \subset \mathcal{A}$. ∎

A further step in the analysis is carried out by picking up a special parametrization for curves on the torus. To do this, we use the Lagrangian function related to H.

DEFINITION 4.3. *A curve γ defined on an interval J is said to have a* Lagrangian parametrization *if*

$$L(\gamma(t),\dot\gamma(t))+c=\sigma(\gamma(t),\dot\gamma(t)) \qquad \text{for a.e. } t\in J. \tag{16}$$

The definition of the semidistance S and the inequality $L(x,q)\geq\sigma(x,q)-c$, which holds for every x and q, imply the following.

PROPOSITION 4.4. *Any critical curve has a Lagrangian parametrization.*

More generally, the following reparametrization proposition holds.

PROPOSITION 4.5. *Any curve of finite length and with closure of the support disjoint from $\mathcal{E}$ admits a Lagrangian reparametrization defined on some compact interval $[0,T]$.*

Proof. The first step is to show the existence of an upper semicontinuous (resp., lower semicontinuous) function $\overline\lambda(x,q)$ (resp., $\underline\lambda(x,q)$) defined in $(\mathbb{T}^N\setminus\mathcal{E})\times(\mathbb{R}^N\setminus\{0\})$ such that the equality

$$L(x,\overline\lambda(x,q)q)=\overline\lambda(x,q)\sigma(x,q)-c \tag{17}$$

and the similar one obtained by replacing $\overline\lambda(\cdot,\cdot)$ by $\underline\lambda(\cdot,\cdot)$ hold true.

Given $(x,q)\in(\mathbb{T}^N\setminus\mathcal{E})\times(\mathbb{R}^N\setminus\{0\})$, and denoting $\{p\,:\,H(x,p)\leq c\}$ by Z, we have that $q\in N_Z(p_0)$ for some p_0 with $H(x,p_0)=c$, and therefore

$$\lambda q\in D_p^-H(x,p_0)\quad\text{for some }\lambda>0 \tag{18}$$

in view of Theorem 23.7 of [17]. Consequently the set of nonnegative λ satisfying (17) in place of $\overline\lambda(x,q)$, denoted by $F(x,q)$, is nonempty; see Theorem A.2. It is, moreover, a compact subset of $(0,+\infty)$. We see, in fact, that, for λ large, relation (18) is impossible, when $H(x,p_0)=c$, since Z is compact and $H(x,\cdot)$ is locally Lipschitz continuous. This shows that $F(x,q)$ is bounded from above. It is also closed thanks to the continuity of $\sigma(x,\cdot)$ and $L(x,\cdot)$, and to the inequality

$$L(x,\lambda q)\geq\sigma(x,\lambda q)-c\quad\text{for every }\lambda\geq0.$$

Moreover, $0\notin F(x,q)$ because $x\notin\mathcal{E}$, and consequently $L(x,0)=-\min_pH(x,p)<-c$. We then define

$$\overline\lambda(x,q)=\max_{F(x,q)}\lambda,\quad \underline\lambda(x,q)=\min_{F(x,q)}\lambda,$$

and we see that these functions, for (x,q) varying in $(\mathbb{T}^N\setminus\mathcal{E})\times(\mathbb{R}^N\setminus\{0\})$, are upper semicontinuous and lower semicontinuous, respectively. In particular,

$$0<\inf_{(x,q)\in C}\underline\lambda(x,q)\leq\sup_{(x,q)\in C}\overline\lambda(x,q)<+\infty \tag{19}$$

for every compact set C contained in $(\mathbb{T}^N\setminus\mathcal{E})\times(\mathbb{R}^N\setminus\{0\})$.

The assertion is now obtained, arguing as in [13, Proposition 7.4]. We give the proof for the reader's convenience. Let γ be a curve as in the statement. We may

assume that it is parametrized by arc-length on $[0, \ell(\gamma)]$, with $\ell(\gamma) > 0$. Set

$$\lambda(\varsigma) := \underline{\lambda}(\gamma(\varsigma), \dot\gamma(\varsigma)) \qquad \text{for a.e. } \varsigma \in (0, \ell(\gamma))$$

and

$$T := \int_0^{\ell(\gamma)} \frac{1}{\lambda(\varsigma)}\, \mathrm{d}\varsigma.$$

Such a quantity is finite and positive, thanks to the measurable character of $\lambda(\cdot)$ and to (19) with $C := \gamma\big([0, \ell(\gamma)]\big) \times \{p : |p| = 1\}$. We will now prove the existence of a curve $\overline\gamma$, in $[0, T]$, which is a Lagrangian reparametrization of γ. To this aim, let us define

$$f(s) := \int_0^s \frac{1}{\lambda(\varsigma)}\, \mathrm{d}\varsigma \qquad \text{for any } s \in [0, \ell(\gamma)],$$

and set $\varphi := (f)^{-1}$ on $[0, T]$. As

$$\dot\varphi(\tau) = \lambda(\varphi(\tau)) \qquad \text{for a.e. } \tau \in [0, T],$$

we immediately derive that φ is an order preserving bi-Lipschitz diffeomorphism. Letting $\overline\gamma := \gamma \circ \varphi$, we get

$$\dot{\overline\gamma}(\cdot) := \lambda(\varphi(\cdot))\ \dot\gamma(\varphi(\cdot)) \qquad \text{a.e. on } [0, T],$$

and the conclusion follows at once by the definition of $\lambda(\cdot)$. □

Remark 4.6. A notion of Lagrangian parametrization can be given at any level $a > c$ by replacing in (16) c and σ with a and σ_a, respectively. Proposition 4.5 can be accordingly generalized, providing Lagrangian reparametrizations for any curve, without the requirement of an empty intersection with $\mathcal{E}$. Such a restriction comes, in fact, from the necessity of avoiding that a p_0 satisfying $H(x_0, p_0) = c$, for some x_0, be a minimizer of $p \mapsto H(x_0, p)$. This possibility is actually ruled out for a p_0 with $H(x_0, p_0) = a$ when $a > c$.

Exploiting the previous remark, we can provide, in a sense, a generalization of Proposition 4.5. This result will be used in the proof of Proposition 5.5.

LEMMA 4.7. *Let $\gamma \in \mathrm{Lip}([0,1], \mathbb{T}^N)$. For any $T > 0$ we set*

$$[\gamma]_T := \{\, \xi \in \mathrm{Lip}([0, T], \mathbb{T}^N) \,:\, \xi \text{ is a reparametrization of } \gamma \,\}.$$

Then

$$\int_0^1 \sigma(\gamma, \dot\gamma)\, \mathrm{d}s = \inf \left\{ \int_0^T \Big(L(\xi, \dot\xi) + c \Big)\, \mathrm{d}s \;:\; \xi \in [\gamma]_T,\ T > 0 \right\}. \tag{20}$$

Proof. It is apparent that the left-hand term of (20) is not greater than the right-hand term. To prove the converse inequality, we select a decreasing sequence $(\delta_n)_n$ with $\delta_n \downarrow 0$. Since $\sigma(x, q) = \inf_n \sigma_{c+\delta_n}(x, q)$ for every $(x, q) \in \mathbb{T}^N \times \mathbb{R}^N$, by the monotone convergence theorem we get

$$\int_0^1 \sigma(\gamma, \dot\gamma)\, \mathrm{d}s = \inf_n \int_0^1 \sigma_{c+\delta_n}(\gamma, \dot\gamma)\, \mathrm{d}s. \tag{21}$$

Taking into account Remark 4.6, we have a Lagrangian reparametrization γ_n of γ at level $a = c + \delta_n$, for any n, defined in some interval $[0, T_n]$, with $T_n > 0$, such that

$$\int_0^1 \sigma_{c+\delta_n}(\gamma, \dot\gamma)\,\mathrm{d}s = \int_0^{T_n} \sigma_{c+\delta_n}(\gamma_n, \dot\gamma_n)\,\mathrm{d}s = \int_0^{T_n} (L(\gamma_n, \dot\gamma_n) + c + \delta_n)\ \mathrm{d}s$$
$$\geq \int_0^{T_n} (L(\gamma_n, \dot\gamma_n) + c)\ \mathrm{d}s.$$

The assertion therefore follows from (21). □

The main result we aim at, in this section, is the following.

THEOREM 4.8. *Through any point of $\mathcal{A}$ there passes a critical curve defined on the whole $\mathbb{R}$.*

We start with a lemma, then prove a local version of Theorem 4.8, and thereafter get the full result by using Zorn's lemma.

LEMMA 4.9. *There exists a real number $R > 0$ such that*

$$\{q \in \mathbb{R}^N : L(x, q) + c = \sigma(x, q) \text{ for some } x \in \mathbb{T}^N\} \subseteq B_R.$$

Proof. We can take R as the Lipschitz constant of the function $p \mapsto H(x, p)$ for $x \in \mathbb{T}^N$ and p satisfying $H(x, p) = c$. To see that this quantity is actually well defined, note that the condition on (x, p) singles out a compact set in $\mathbb{T}^N \times \mathbb{R}^N$ in view of the coercivity assumption (H3), and take into account Remark 2.1.

If $q \in \mathbb{R}^N$, $x_0 \in \mathbb{T}^N$ are such that $L(x_0, q) + c = \sigma(x_0, q)$, then $q \in D_p^- H(x_0, p_0)$ for some p_0 with $H(x_0, p_0) = c$, and so $|q| \leq R$. □

LEMMA 4.10. *For any $y \in \mathcal{A}$, there exists $\delta \in (0, +\infty]$ and a critical curve η which is defined in $(-\delta, \delta)$ and satisfies $\eta(0) = y$.*

Proof. If $y \in \mathcal{E}$, we simply set $\eta(t) = y$ for every $t \in \mathbb{R}$. By the definition of equilibrium point, we have

$$L(y, 0) + c = \max_{p \in \mathbb{R}^N} -H(x, p) + c = 0 \tag{22}$$

for every $t \in \mathbb{R}$, which shows that η is indeed a critical curve. If $y \in \mathcal{A} \setminus \mathcal{E}$, we exploit Lemma 9.4 of [13] to see that there exists a curve γ contained in $\mathcal{A}$, and defined in some neighborhood J of $t = 0$, such that $\gamma(0) = y$ and

$$S(\gamma(t_1), \gamma(t_2)) = \int_{t_1}^{t_2} \sigma(\gamma, \dot\gamma)\,\mathrm{d}s = -S(\gamma(t_2), \gamma(t_1))$$

for every t_1, $t_2 \in J$ with $t_2 > t_1$. Note that this result does not require the Lipschitz continuity of H in x, which was assumed in that paper, and therefore holds also in our present setting.

Because of the local character of the construction, we can assume that γ stays away from $\mathcal{E}$. We thus consider a Lagrangian reparametrization of γ, which does exist in view of Proposition 4.5, to get the required curve. □

PROPOSITION 4.11. *Let $y \in \mathcal{A}$. Then there exists a critical curve η defined on $\mathbb{R}$ with $\eta(0) = y$.*

Proof. In view of Lemma 4.10, we may assume that $y \in \mathcal{A} \setminus \mathcal{E}$. We denote by $\mathcal{C}$ the set of pairs (T, η), where $T \in (0, +\infty]$, and η is a critical curve defined on $(-T, T)$ and equaling y at 0. We give an order relation in $\mathcal{C}$ by defining

$$(T, \eta) \preceq (T', \eta') \quad \text{if} \quad T \leq T' \ \text{ and } \ \eta'_{|(-T,T)} = \eta.$$

The set $\mathcal{C}$ is nonempty by Lemma 4.10. To prove that $\mathcal{C}$ is inductively ordered, we take a nonempty chain $\{(T_i, \eta_i)\}$, with i in some set of indices I, and observe that an upper bound $(\hat{T}, \hat{\eta}) \in \mathcal{C}$ can be defined through

$$\hat{T} := \sup_i T_i \quad \text{and} \quad \hat{\eta}(t) := \eta_i(t) \quad \text{if} \quad t \in (-T_i, T_i) \text{ for every } i \in I.$$

Zorn's lemma provides the existence of a maximal element (T_y, η_y) in $\mathcal{C}$. We claim that $T_y = +\infty$. If, in fact, this were not the case, and $T_y < +\infty$, then the curve η_y would have a limit (belonging to $\mathcal{A}$) for t going to $\pm T_y$, in view of Lemma 4.9. It would then be possible to extend η_y to some interval $(-T_y - \delta, T_y + \delta)$ for a suitable $\delta > 0$ by applying Lemma 4.10 to these limit points. This would violate the maximality of (T_y, η_y). □

We denote by $\mathcal{K}$ the family of all maximal critical curves, and by $\mathcal{K}(y)$ the subset of $\mathcal{K}$ made up of those equaling y at $t = 0$, for each $y \in \mathcal{A}$.

We proceed to prove a compactness property for $\mathcal{K}$.

PROPOSITION 4.12. *$\mathcal{K}$ is a compact metric space with respect to the local uniform convergence on $\mathbb{R}$.*

Proof. Let $(\eta_k)_k$ be a sequence in $\mathcal{K}$. The curves η_k are uniformly bounded by the compactness of $\mathbb{T}^N$, and equi-Lipschitz continuous by Lemma 4.9, and hence we can apply the Ascoli–Arzelà theorem to infer the existence of a subsequence (not relabeled) which converges locally uniformly to some curve η defined on $\mathbb{R}$. The limit curve η is contained in $\mathcal{A}$, as the Aubry set is closed, and clearly satisfies

$$S(\eta(t_1), \eta(t_2)) = -S(\eta(t_2), \eta(t_1)) \tag{23}$$

for every t_1, t_2 in $\mathbb{R}$. If, in addition, $t_2 > t_1$, we have

$$S(\eta_k(t_1), \eta_k(t_2)) = \int_{t_1}^{t_2} (L(\eta_k(s), \dot{\eta}_k(s)) + c)\, \mathrm{d}s$$

for every k, and we therefore deduce, thanks to Theorem 2.9,

$$S(\eta(t_1), \eta(t_2)) = \lim_{k \to +\infty} \int_{t_1}^{t_2} (L(\eta_k(s), \dot{\eta}_k(s)) + c)\, \mathrm{d}s \geq \int_{t_1}^{t_2} (L(\eta(s), \dot{\eta}(s)) + c)\, \mathrm{d}s.$$

Since the converse inequality is apparent, we get in the end

$$S(\eta(t_1), \eta(t_2)) = \int_{t_1}^{t_2} (L(\eta(s), \dot{\eta}(s)) + c)\, \mathrm{d}s. \tag{24}$$

Relations (23), (24) show that $\eta \in \mathcal{K}$. □

Given $\eta \in \mathcal{K}$, we denote by $\omega(\eta)$ the set of its ω-limits, i.e., of the points x_0 satisfying

$$x_0 = \lim_k \eta(s_k) \quad \text{with } s_k \to +\infty \text{ as } k \to +\infty. \tag{25}$$

We deduce from Proposition 4.12 that through any point x_0 of $\omega(\eta)$ there passes a critical curve entirely lying in $\omega(\eta)$. If, in fact, (25) holds, then $\{\eta(s_k + \cdot)\}_k$ converges locally uniformly, up to a subsequence, to a curve γ, which equals x_0 at 0, and is contained in $\omega(\eta)$.

Remark 4.13. We can describe more precisely $\omega(\eta)$ if the sequence s_k, appearing in (25), is increasing and such that $s_{k+1} - s_k$ converges to a finite limit, necessarily nonnegative, say T, and $\{\eta(s_k + \cdot)\}_k$ converges locally uniformly to a curve γ.

In this case, $\omega(\eta)$ coincides with the support of γ, which is a cycle of period T because of the relations

$$\gamma(t+T) = \lim_k \eta(s_k + T + t) = \lim_k \eta(s_{k+1} + t) = \gamma(t),$$

which hold for any t. If, in fact, $y_0 := \lim_n \eta(t_n)$ belongs to $\omega(\eta)$, with $(t_n)_n$ a diverging sequence, then we can select, for any n, an index $k_n \in \mathbb{N}$ satisfying $s_{k_n} \leq t_n < s_{k_n+1}$. The sequence $t_n - s_{k_n}$ is therefore bounded and thus convergent, up to a subsequence, to some $t_0 \in [0, T]$. It then follows that $y_0 = \gamma(t_0)$.

If in particular $T = 0$, then γ reduces to a point, which must be the support of a critical curve, and consequently belongs to $\mathcal{E}$.

We know from [13] that, if H is Lipschitz continuous in x, all critical subsolutions are strictly differentiable at any point of the Aubry set and have the same derivative. This implies that they coincide, up to an additive constant, on every rectifiable subset of $\mathcal{A}$. These results are based upon some semiconcavity estimates which, in turn, depend essentially on the Lipschitz character of the Hamiltonian in x that we do not have here. We can nevertheless find something similar in our setting by looking at the behavior of the critical subsolutions on curves of $\mathcal{K}$.

THEOREM 4.14. *Let $\eta \in \mathcal{K}$. Then all critical subsolutions coincide on $\eta(\mathbb{R})$, up to an additive constant. There exists, in addition, a negligible set $\Sigma \subset \mathbb{R}$ such that, for any critical subsolution ϕ, the map $\phi \circ \eta$ is differentiable on $\mathbb{R} \setminus \Sigma$ and satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\phi \circ \eta\right)(t_0) = \sigma(\eta(t_0), \dot\eta(t_0)) \qquad \text{whenever } t_0 \in \mathbb{R} \setminus \Sigma. \tag{26}$$

We show first an auxiliary lemma, on which the proof of Theorem 4.14 is based.

PROPOSITION 4.15. *Let $\eta \in \mathcal{K}$. Then there exists a negligible set $\Sigma \subset \mathbb{R}$ such that the functions $\eta(\cdot)$, $S(\eta(t_0), \eta(\cdot))$ and $-S(\eta(\cdot), \eta(t_0))$ are differentiable at any t_0 in $\mathbb{R} \setminus \Sigma$, and*

$$\frac{\mathrm{d}}{\mathrm{d}t} S(\eta(t_0), \eta(t))\Big|_{t=t_0} = -\frac{\mathrm{d}}{\mathrm{d}t} S(\eta(t), \eta(t_0))\Big|_{t=t_0} = \sigma(\eta(t_0), \dot\eta(t_0)). \tag{27}$$

Proof. Let Σ be a negligible subset of $\mathbb{R}$ such that every $t_0 \in \mathbb{R} \setminus \Sigma$ is a differentiability point for $\eta(\cdot)$ and a Lebesgue point for the function $\sigma(\eta(\cdot), \dot\eta(\cdot))$. The existence of such a set is guaranteed by Rademacher and Lebesgue differentiability theorems. As the curve η is critical, we have

$$\frac{S(\eta(t_0), \eta(t))}{t - t_0} = \frac{1}{t - t_0} \int_{t_0}^{t} \sigma(\eta(s), \dot\eta(s))\, \mathrm{d}s \qquad \text{for every } t > t_0.$$

Since t_0 is a Lebesgue point of $\sigma(\eta(\cdot), \dot\eta(\cdot))$, we derive

$$\lim_{t \to t_0^+} \frac{S(\eta(t_0), \eta(t))}{t - t_0} = \sigma(\eta(t_0), \dot\eta(t_0))$$

for every $t_0 \in \mathbb{R} \setminus \Sigma$. A similar limit relation for $t \to t_0^-$ can be deduced analogously. ∎

Proof of Theorem 4.14. Let Σ and ϕ be the subset of $\mathbb{R}$ given by Proposition 4.15 and a critical subsolution, respectively. By Proposition 2.3, we have

$$-S(\eta(t),\eta(t_0)) \le \phi(\eta(t)) - \phi(\eta(t_0)) \le S(\eta(t_0),\eta(t)) \qquad \text{for every } t, t_0 \in \mathbb{R},$$

and hence we get (26), for $t_0 \in \mathbb{R} \setminus \Sigma$, in view of Proposition 4.15. This fully proves the assertion. ∎

We point out two consequences of the previous theorem that we use in the next section and judge to be of independent interest as well.

PROPOSITION 4.16. *Two critical subsolutions coinciding on* $\mathcal{M} := \bigcup_{\eta\in\mathcal{K}} \omega(\eta)$ *must also coincide on* $\mathcal{A}$.

Proof. Let ϕ_1, ϕ_2 be two critical subsolutions coinciding on $\mathcal{M}$. Take y and η in $\mathcal{A}$ and in $\mathcal{K}(y)$, respectively. Let $(t_n)_n$ be a diverging sequence such that $\lim_n \eta(t_n) = x \in \mathcal{M}$. As $S(y,\cdot)$ is a critical subsolution (cf. Proposition 2.3), Theorem 4.14 yields

$$\phi_i(y) = \phi_i(\eta(0)) - S(y,\eta(0)) = \phi_i(\eta(t_n)) - S(y,\eta(t_n))$$

for every $n \in \mathbb{N}$, $i \in \{1,2\}$. Sending n to $+\infty$, we get

$$\begin{aligned}\phi_1(y) &= \lim_{n\to+\infty} \phi_1(\eta(t_n)) - S(y,\eta(t_n)) = \phi_1(x) - S(y,x) = \phi_2(x) - S(y,x)\\ &= \lim_{n\to+\infty} \phi_2(\eta(t_n)) - S(y,\eta(t_n)) = \phi_2(y),\end{aligned}$$

whence we get the assertion since y is an arbitrary point of $\mathcal{A}$. ∎

Remark 4.17. As the curve $\eta(t) := y$, for every $t \in \mathbb{R}$, is critical whenever $y \in \mathcal{E}$, it is apparent from the definitions that the set $\mathcal{E}$ is always contained in $\mathcal{M}$.

PROPOSITION 4.18. *The set* $\mathcal{M}$ *is a uniqueness set for* (2); *i.e., two solutions of* (2) *coinciding on* $\mathcal{M}$ *coincide on the whole torus too.*

Proof. The assertion comes from the previous proposition and from the property that $\mathcal{A}$ is a uniqueness set for (2), as established in Theorem 2.4. ∎

5. Convergence to steady states. We are now ready to prove our main convergence result. Throughout this section we will assume, without loss of generality, $c = 0$. We also assume that H satisfies conditions (H1), (H2)′, (H3). We recall that $u_0 \in C(\mathbb{T}^N)$ is the initial datum of the Cauchy problem (6) and that $\omega_{\mathcal{S}}(u_0)$ denotes the family of the uniform limits of $\mathcal{S}(t_n)u_0$ for some diverging sequence $(t_n)_n$. We start by establishing some monotonicity properties for the function $\mathcal{S}(t)\psi - \phi$ on the curves of $\mathcal{K}$, where ψ is any continuous function and ϕ any critical subsolution. The next result is analogous to Lemma 3.1 in [18].

PROPOSITION 5.1. *Let* $\eta \in \mathcal{K}$. *Then the map* $t \mapsto \big(\mathcal{S}(t)\psi\big)(\eta(t)) - \phi(\eta(t))$ *is nonincreasing on* $\mathbb{R}_+$ *for any* $\psi \in C(\mathbb{T}^N)$ *and any critical subsolution* ϕ.

Proof. Let t_1, t_2 in $\mathbb{R}_+$ with $t_2 \ge t_1$. Taking into account Theorem 4.14 and the integral representation formula for the Lax–Oleinik semigroup, we get

$$\big(\mathcal{S}(t_2)\psi\big)(\eta(t_2)) - \big(\mathcal{S}(t_1)\psi\big)(\eta(t_1)) \le \int_{t_1}^{t_2} L(\eta(s),\dot\eta(s))\,\mathrm{d}s = \phi(\eta(t_2)) - \phi(\eta(t_1)),$$

which proves the assertion. ∎

We proceed to prove that a strict monotonicity property actually holds on the critical curves under appropriate assumptions. This result relies on a lemma, which we demonstrate first and which estimates the modification of the line integral of the Lagrangian on a critical curve when the Lagrangian parametrization is suitably

perturbed. As already pointed out in the introduction, we emphasize that, for this, we essentially use the differentiability of L in q and the continuity of $D_qL(x,q)$ in $\mathrm{int}(\mathrm{dom}L)$, a property that is equivalent, for a continuous Hamiltonian, to the strict convexity of H in the second variable (cf. [7]). These results are key tools for the forthcoming convergence theorem.

LEMMA 5.2. *There is a modulus $\omega(\cdot)$ such that, if η is any curve in $\mathcal{K}$ and λ is suitably close to* 1, *we have*

$$\int_{t_1}^{t_2} L(\eta_\lambda, \dot\eta_\lambda)\,\mathrm{d}s \le S(\eta_\lambda(t_1), \eta_\lambda(t_2)) + |\lambda-1|\,\omega(|\lambda-1|)(t_2-t_1)$$

for every t_1, t_2 with $t_2 > t_1$, where $\eta_\lambda(t) := \eta(\lambda t)$ for all $t \in \mathbb{R}$.

Proof. We claim that $K := \{(x,q) \in \mathcal{A}\times\mathbb{R}^N : L(x,q) = \sigma(x,q)\}$ is a compact subset of $\mathrm{int}(\mathrm{dom}L)$. It is in fact closed by the lower and upper semicontinuity of L and σ, respectively, bounded by Lemma 4.9, and contained in $\mathrm{int}(\mathrm{dom}L)$ thanks to Proposition 2.7(iv). There thus exists $\delta > 0$ such that the set $K_\delta := \{(x,\lambda q) : (x,q) \in K,\ |\lambda-1| \le \delta\}$ is compactly contained in $\mathrm{int}(\mathrm{dom}L)$.

Let us now fix λ in $(1-\delta, 1+\delta)$ and denote by θ a continuity modulus for the function $(x,q) \mapsto D_qL(x,q)$ in K_δ. For a.e. $s \in \mathbb{R}$ we have

$$(\eta(\lambda s), \dot\eta(\lambda s)) \in K, \tag{28}$$

$$\langle D_q\, L(\eta(\lambda s), \dot\eta(\lambda s)), \dot\eta(\lambda s)\rangle = \sigma(\eta(\lambda s), \dot\eta(\lambda s)), \tag{29}$$

where the first relation comes from the very definition of critical curve, and the second one holds in view of Theorem A.2. Let s be such that (28) and (29) hold. The application of the mean value theorem to the function $\mu \mapsto L(\eta(\lambda s), \mu\eta(\lambda s))$ in the interval with end points 1 and λ yields

$$L(\eta(\lambda s), \lambda\,\dot\eta(\lambda s)) - L(\eta(\lambda s), \dot\eta(\lambda s)) = (\lambda-1)\langle D_q\, L(\eta(\lambda s), \mu_0\,\dot\eta(\lambda s)), \dot\eta(\lambda s)\rangle,$$

where μ_0 is a suitable constant between λ and 1. By using (28), (29), and the definition of $\theta(\cdot)$, we derive from this identity

$$L(\eta(\lambda s), \lambda\,\dot\eta(\lambda s)) \le \lambda\,\sigma(\eta(\lambda s), \dot\eta(\lambda s)) + R|\lambda-1|\,\theta(|\lambda-1|R),$$

where R is the positive constant provided by Lemma 4.9. We now exploit the previous estimate and the fact that η is a critical curve to get for any t_1, t_2 in $\mathbb{R}$ with $t_2 > t_1$

$$\begin{aligned}\int_{t_1}^{t_2} L(\eta_\lambda, \dot\eta_\lambda)\,\mathrm{d}s &= \int_{t_1}^{t_2} L(\eta(\lambda s), \lambda\dot\eta(\lambda s))\,\mathrm{d}s\\ &\le \int_{t_1}^{t_2} \lambda\,\sigma(\eta(\lambda s), \dot\eta(\lambda s))\,\mathrm{d}s + (t_2-t_1)|\lambda-1|\,R\;\theta(R\,|\lambda-1|)\\ &= S(\eta_\lambda(t_1), \eta_\lambda(t_2)) + (t_2-t_1)|\lambda-1|\,R\;\theta(R\,|\lambda-1|).\end{aligned}$$

The proof is complete. □

PROPOSITION 5.3. *Let $\eta \in \mathcal{K}$, $\psi \in C(\mathbb{T}^N)$, and ϕ be a critical subsolution. Let us assume $D^+\big((\psi-\phi)\circ\eta\big)(0)\setminus\{0\} \neq \emptyset$ (recall that D^+ indicates the superdifferential); then*

$$(\mathcal{S}(t)\psi)(\eta(t)) - \phi(\eta(t)) < \psi(\eta(0)) - \phi(\eta(0)) \qquad \textit{for every } t > 0. \tag{30}$$

Proof. We fix $t > 0$. Inequality (30) will be proved for $\phi := -S(\cdot, \eta(t))$, which is enough to get the full result, in view of Theorem 4.14. We also assume, without loss of generality in view of Proposition 2.6(iii), that $\psi(\eta(0)) - \phi(\eta(0)) = 0$. We are thus lead to show that the left-hand term of (30) is strictly negative. To this aim, we take into account the integral formula for the Lax–Oleinik semigroup, given in section 2, to get, for λ close to 1 and η_λ defined as in Lemma 5.2,

$$(\mathcal{S}(t)\psi)(\eta(t)) - \phi(\eta(t)) = (\mathcal{S}(t)\psi)(\eta(t)) \leq \int_{(1/\lambda-1)t}^{t/\lambda} L(\eta_\lambda, \dot\eta_\lambda)\,\mathrm{d}s + \psi\big(\eta((1-\lambda)t)\big),$$

whence, by Lemma 5.2,

$$(\mathcal{S}(t)\psi)(\eta(t)) - \phi(\eta(t)) \leq \psi\big(\eta((1-\lambda)t)\big) - \phi\big(\eta((1-\lambda)t)\big) + t|\lambda-1|\,\omega(|\lambda-1|).$$

If $m \neq 0$ is an element of $D^+\big((\psi-\phi)\circ\eta\big)(0)$, we therefore have

$$(\mathcal{S}(t)\psi)(\eta(t)) - \phi(\eta(t)) \leq m\big((1-\lambda)t\big) + o\big((1-\lambda)t\big) + t|\lambda-1|\,\omega(|\lambda-1|),$$

where $o(\cdot)$ satisfies $\lim_{\lambda\to 1} \frac{o((1-\lambda)t)}{1-\lambda} = 0$. A suitable choice of λ close to 1 makes the left-hand term of the previous formula strictly negative, and consequently proves the assertion, for the arbitrariness of t. □

We combine the information gathered in Propositions 5.1 and 5.3 with some properties of the Lax–Oleinik semigroup to get the following.

PROPOSITION 5.4. *Let ϕ be a critical subsolution, and let $\psi \in \omega_{\mathcal{S}}(u_0)$. For any $x_0 \in \mathcal{M}$ there exists a curve $\gamma \in \mathcal{K}(x_0)$ such that the function $t \mapsto \psi(\gamma(t)) - \phi(\gamma(t))$ is constant on $\mathbb{R}$.*

Proof. Let $(s_k)_k$ and $(t_k)_k$ be two diverging sequences, η a curve of $\mathcal{K}$ such that $x_0 = \lim_k \eta(s_k)$, and ψ the uniform limit of $\mathcal{S}(t_k)u_0$ in $\mathbb{T}^N$. We can assume that the curve γ, defined by $\gamma(t) = \lim_k \eta(t+s_k)$, for any t, is the local uniform limit of the sequence $\eta(s_k+\cdot)$ in $\mathbb{R}$, and so $\gamma \in \mathcal{K}$. We assume, in addition, that $t_k - s_k \to +\infty$, as $k \to +\infty$, and that $\mathcal{S}(t_k - s_k)u_0$ uniformly converges to some $\psi_1 \in \omega_{\mathcal{S}}(u_0)$. The nonexpansiveness of the Lax–Oleinik semigroup implies

$$\|\mathcal{S}(t_k)u_0 - \mathcal{S}(s_k)\psi_1\|_\infty = \|\mathcal{S}(s_k+t_k-s_k)u_0 - \mathcal{S}(s_k)\psi_1\|_\infty \leq \|\mathcal{S}(t_k-s_k)u_0 - \psi_1\|_\infty,$$

which entails $\mathcal{S}(s_k)\psi_1 \rightrightarrows \psi$ in $\mathbb{T}^N$. We know from Proposition 5.1 that the function

$$s \mapsto (\mathcal{S}(s)\psi_1)(\eta(s)) - \phi(\eta(s))$$

is nonincreasing in $\mathbb{R}_+$, and hence it admits a limit, denoted by l, as $s \to +\infty$. Such a limit is furthermore finite, since it is greater than or equal to $-\|\overline{u} - \phi\|_\infty$. Given $t > 0$, we have

$$l = \lim_{k\to+\infty} (\mathcal{S}(s_k+t)\psi_1)(\eta(s_k+t)) - \phi(\eta(s_k+t)) = (\mathcal{S}(t)\psi)(\gamma(t)) - \phi(\gamma(t)).$$

The function $t \mapsto (\mathcal{S}(t)\psi)(\gamma(t)) - \phi(\gamma(t))$ is therefore constant on $\mathbb{R}_+$. From this we deduce, by applying Proposition 5.3 to the curve $\gamma(s+\cdot) \in \mathcal{K}$, for any fixed s, that $D^+\big((\psi-\phi)\circ\gamma\big)(s) \setminus \{0\} = \emptyset$ for any $s \in \mathbb{R}$. This implies that $\psi - \phi$ is constant on γ. □

The previous proposition shows that any function ψ in $\omega_{\mathcal{S}}(u_0)$ coincides, on any given critical curve γ lying in $\mathcal{M}$, with some critical subsolution ϕ. Such a critical subsolution may a priori depend on the curve γ and on ψ. We proceed to show that, on the contrary, ϕ is uniquely determined and coincides with the function v defined by (11), putting u_0 in place of w_0. In force of Proposition 5.4, it will be enough to prove the following fact.

PROPOSITION 5.5. *Given $\eta \in \mathcal{K}$, $\psi \in \omega_{\mathcal{S}}(u_0)$, $\varepsilon > 0$, there exists $\tau \in \mathbb{R}$ such that*

$$|v(\eta(\tau)) - \psi(\eta(\tau))| < \varepsilon,$$

where v is the critical solution defined by (11) *with $w_0 = u_0$.*

Proof. Since the curve η is contained in $\mathcal{A}$, and in view of Theorem 3.1(ii), we have

$$v(\eta(0)) = \min_{z \in \mathbb{T}^N} \Big(u_0(z) + S(z, \eta(0)) \Big),$$

and hence $v(\eta(0)) = u_0(z_0) + S(z_0, \eta(0))$, for some $z_0 \in \mathbb{T}^N$. We choose a curve $\gamma \in \mathrm{Lip}_{z_0,\eta(0)}([0,1], \mathbb{T}^N)$ such that

$$v(\eta(0)) + \varepsilon/2 = u_0(z_0) + S(z_0, \eta(0)) + \varepsilon/2 > u_0(z_0) + \int_0^1 \sigma(\gamma, \dot\gamma)\, \mathrm{d}s.$$

We thereafter take into account Lemma 4.7 and the integral representation formula for the Lax–Oleinik semigroup to get

$$v(\eta(0)) + \varepsilon/2 > u_0(z_0) + \int_0^T L(\gamma_T, \dot\gamma_T)\, \mathrm{d}s \geq \big(\mathcal{S}(T)u_0\big)(\eta(0)),$$

where γ_T is a suitable reparametrization of γ on $[0,T]$, for some $T > 0$. Now letting $(\tau_n)_n$ be a diverging sequence with $\mathcal{S}(\tau_n)u_0 \rightrightarrows \psi$, we have

$$\|\mathcal{S}(\tau_n)u_0 - \psi\|_\infty < \varepsilon/2 \qquad \text{and} \qquad \tau_n - T > 0 \qquad \text{for } n \text{ sufficiently large.}$$

Pick such an n and set $\tau = \tau_n - T$, and then use the above inequalities and Theorem 4.14 to obtain

$$\begin{aligned} \psi(\eta(\tau)) - \varepsilon/2 &< \big(\mathcal{S}(\tau_n)u_0\big)(\eta(\tau)) = \big(\mathcal{S}(\tau)\mathcal{S}(T)u_0\big)(\eta(\tau)) \\ &\leq \big(\mathcal{S}(T)u_0\big)(\eta(0)) + \int_0^\tau L(\eta, \dot\eta)\, \mathrm{d}s \\ &< \varepsilon/2 + v(\eta(0)) + \int_0^\tau L(\eta, \dot\eta)\, \mathrm{d}s = \varepsilon/2 + v(\eta(\tau)). \end{aligned}$$

This gives the assertion since $\psi(\eta(\tau)) - v(\eta(\tau)) \geq 0$ by Proposition 3.5. ∎

We directly derive the following theorem from Propositions 5.4 and 5.5.

THEOREM 5.6. *Any function in $\omega_{\mathcal{S}}(u_0)$ coincides with v on $\mathcal{M}$, where v is the critical subsolution defined by* (11)*, with u_0 in place of w_0.*

We finally prove our main result.

THEOREM 5.7. *Let H satisfy conditions* (H1), (H2)′, (H3) *and $u_0 \in \mathrm{C}(\mathbb{T}^N)$. Then $\mathcal{S}(t)u_0$ uniformly converges to v on $\mathbb{T}^N$ as t goes to $+\infty$, where v is the critical solution given by formula* (11) *with $w_0 = u_0$.*

Proof. Theorem 5.6 implies that v and $\underline{u}$ coincide on $\mathcal{M}$; therefore they coincide on $\mathcal{A}$ thanks to Proposition 4.16. The comparison principle given in Theorem 2.4 hence tells us that $\underline{u} \leq v$ on the whole torus, since $\underline{u}$ is a critical subsolution and v is a critical solution. The assertion is at last obtained thanks to Proposition 3.5. ∎

Remark 5.8. We stress that the only place in the present section (actually, in the whole paper) where the strict convexity assumption is directly employed is Lemma 5.2. More precisely, it uses the global continuity of $D_q L(x,q)$ in $\mathrm{int}(\mathrm{dom} L)$, a property that is equivalent to the strict convexity of H in p, as previously noted. As a matter of fact, we do not exploit such a condition in its full strength. The existence of a continuity modulus for $D_q L(x,q)$ in a neighborhood of the image of the map $t \mapsto (\eta, \dot{\eta})$, for each $\eta \in \mathcal{K}$, might be sufficient.

Since the stationary curve $\gamma(\cdot) = y$ belongs to $\mathcal{K}$ whenever $y \in \mathcal{E}$, Proposition 5.5—which has been proved without exploiting the strict convexity assumption (H2)′—directly implies that any function of $\omega_{\mathcal{S}}(u_0)$ coincides with v on $\mathcal{E}$. Hence, whenever $\mathcal{E}$ is a uniqueness set for the critical equation (2), the argument of Theorem 5.7 gives the convergence result, bypassing Proposition 5.4, which instead relies on Lemma 5.2. This happens, for instance, when $\mathcal{M} = \mathcal{E}$. We can therefore state the following.

THEOREM 5.9. *Let H satisfy conditions* (H1), (H2), (H3), (H4) *and $u_0 \in \mathrm{C}(\mathbb{T}^N)$. Then any function in $\omega_{\mathcal{S}}(u_0)$ coincides with v on $\mathcal{E}$, where v is the critical subsolution defined by* (11), *with u_0 in place of w_0. In particular, $\mathcal{S}(t)u_0$ uniformly converges to v on $\mathbb{T}^N$, as t goes to $+\infty$, when $\mathcal{M} = \mathcal{E}$, and, more generally, whenever $\mathcal{E}$ is a uniqueness set for the critical equation* (2).

Note that the previous theorem includes the results of [16], where the Hamiltonian under investigation was assumed only convex and with the Aubry set consisting of equilibria.

The next one-dimensional example deals with a family of Hamiltonians, depending on a parameter $\alpha \in \mathbb{R}$, which satisfy assumptions (H1), (H2), (H3), (H4). It is shown that a suitable initial datum for the time-dependent equation can be selected in such a way that the convergence to a steady state does not take place whenever the Hamiltonian under consideration does not satisfy the assumptions of Theorem 5.9. It can be viewed as a development of the example given in [4, section 5].

Example 5.10. Consider the $\mathbb{Z}$-periodic Hamiltonian

$$H(x,p) = |p| - f(x)$$

defined in $\mathbb{R}$ (cf. Remark 2.2), where f is a continuous periodic potential with $f \not\equiv 0$, $f \geq 0$, and $\min_{\mathbb{R}} f = 0$. The effective Hamiltonian $\overline{H}(\alpha)$, i.e., the critical value of $H(x, p+\alpha)$, is given, for any $\alpha \in \mathbb{R}$, by

$$\overline{H}(\alpha) = \max\left\{0, |\alpha| - \int_0^1 f \,\mathrm{d}s\right\};$$

see [15]. It is not difficult to check that, for $\alpha \in \overline{H}^{-1}(0)$, the set of equilibria $\mathcal{E}(\alpha)$, relative to $H(x, p+\alpha)$, coincides with $f^{-1}(0)$ and is a uniqueness set for the corresponding critical equation, while $\mathcal{A}(\alpha) = \mathcal{M}(\alpha) = \mathbb{R}$ and $\mathcal{E}(\alpha) = \emptyset$ as soon as α lies outside the flat part.

Given $\alpha \notin \overline{H}^{-1}(0)$, we define in $[0, +\infty) \times \mathbb{R}$ the function

$$w(t,x) = u_0(x - \mathrm{sgn}\,\alpha\; t) + \mathrm{sgn}\,\alpha \int_0^x f \,\mathrm{d}s - \left(\mathrm{sgn}\,\alpha \int_0^1 f \,\mathrm{d}s\right) x,$$

where sgn indicates the sign function, and u_0 is a C^1 nonconstant periodic function satisfying

$$\operatorname{sgn}\left\{u_0'(x)+\operatorname{sgn}\alpha\,\left(|\alpha|-\int_0^1 f\,\mathrm{d}s\right)\right\}=\operatorname{sgn}\alpha\qquad\text{for all } x\in\mathbb{R}. \tag{31}$$

Note that relation (31) implies

$$\operatorname{sgn}\left(u_0'(x)+\operatorname{sgn}\alpha\, f(y)-\operatorname{sgn}\alpha\,\int_0^1 f\,\mathrm{d}s+\alpha\right)=\operatorname{sgn}\alpha\qquad\text{for all } x,y\in\mathbb{R}, \tag{32}$$

as f is nonnegative. The function $w(t,\cdot)$ is periodic in $\mathbb{R}$ for any t, as easily seen. By taking into account (32), a direct calculation shows that

$$\partial_t w(t,x)+|\partial_x w(t,x)+\alpha|-f(x)-|\alpha|+\int_0^1 f\,\mathrm{d}s=0$$

for every $(t,x)\in(0,+\infty)\times\mathbb{R}$. Hence, w is a periodic C^1-solution of the time-dependent equation

$$\partial_t u+H(x,\partial_x u+\alpha)-\overline{H}(\alpha)=0\qquad\text{in } (0,+\infty)\times\mathbb{R},$$

but it does not converge to any steady state for $t\to+\infty$. Note that H is not strictly convex in the second argument and that $\mathcal{E}(\alpha)=\emptyset$.

Such a construction is clearly not possible when $\alpha\in\overline{H}^{\,-1}(0)$, or, in other words, when $|\alpha|\le\int_0^1 f\,\mathrm{d}s$, because condition (31) implies, in this case, that u_0' does not change sign on $\mathbb{R}$, in contrast with u_0 being nonconstant and periodic.

Appendix. We consider a Hamiltonian $H:\mathbb{T}^N\times\mathbb{R}^N\to\mathbb{R}$ satisfying conditions (H1), (H2), (H3), and we denote by $L:\mathbb{T}^N\times\mathbb{R}^N\to(-\infty,+\infty]$ the corresponding Lagrangian defined through the Fenchel transform (8). As H is assumed coercive but not superlinear, the Lagrangian L is not finite valued in general. Our aim is to give first a proof of Proposition 2.7, and afterward to show the validity of the integral representation formula (9) for the Lax–Oleinik semigroup. We start by recalling some basic facts of convex analysis, and by giving a characterization of the interior of $\mathrm{dom}(L)$, where $\mathrm{dom}(L):=\{(x,q)\in\mathbb{T}^N\times\mathbb{R}^N:\,L(x,q)<+\infty\}$.

THEOREM A.1. *Let $f:\mathbb{R}^N\to(-\infty,+\infty]$ be a convex function with $f\not\equiv+\infty$. Then $D^-f(q)$ is a nonempty bounded set if and only if $q\in\mathrm{int}(\mathrm{dom}f)$, and it is empty for $q\notin\mathrm{dom}f$.*

We refer to [17, Theorem 23.4] for a proof. Note that $L(x,\cdot)$ turns out to be convex and lower semicontinuous on $\mathbb{R}^N$, as supremum of continuous convex functions, for any $x\in\mathbb{T}^N$. Moreover, $L(x,\cdot)\not\equiv+\infty$ (cf. [17, Theorem 12.2]). We have (cf. [17, Theorem 23.5]) the following.

THEOREM A.2. *Let $x\in\mathbb{T}^N$ and $p,q\in\mathbb{R}^N$. The following conditions are equivalent:*

(a) $H(x,p)+L(x,q)\le\langle p,q\rangle$;
(b) $H(x,p)+L(x,q)=\langle p,q\rangle$;
(c) $q\in D_p^-H(x,p)$;
(d) *the function* $\langle\,\cdot\,,q\rangle-L(x,\cdot)$ *achieves its maximum at* p;
(e) $p\in D_q^-L(x,q)$;
(f) *the function* $\langle p,\,\cdot\,\rangle-L(x,\cdot)$ *achieves its maximum at* q.

THEOREM A.3.

(i) *For any* $x \in \mathbb{T}^N$, $\mathrm{dom}(L(x,\cdot))$ *has a nonempty interior.*

(ii) $\mathrm{int}(\mathrm{dom}L) = \bigcup_{x\in\mathbb{T}^N}\{x\} \times \mathrm{int}(\mathrm{dom}L(x,\cdot))$.

Proof. According to Corollary 13.4.2 in [17], the assertion in (i) holds true if and only if there are no lines along which $H(x,\cdot)$ is (finite and) affine. Such a condition is actually a consequence of the coercivity assumption (H3).

To ease notation, let us temporarily denote by $\widetilde{\Omega}$ the set on the right-hand side of the equality in item (ii). It is apparent that $\mathrm{int}(\mathrm{dom}L) \subset \widetilde{\Omega}$.

To prove the converse inclusion, we assume by contradiction the existence of $(x_0,q_0) \in \widetilde{\Omega} \setminus \mathrm{int}(\mathrm{dom}L)$. According to Theorem A.1, this implies that $D_q^- L(x_0,q_0)$ is nonempty and bounded, and, moreover, that there is a sequence $((x_n,q_n))_n$ converging to (x_0,q_0), with $D_q^- L(x_n,q_n)$ either empty or unbounded. In any case, we may find a sequence $(p_n)_n$ such that $|p_n| \to +\infty$ and

$$\langle p_0, q_n\rangle - H(x_n,p_0) \le \langle p_n,q_n\rangle - H(x_n,p_n), \tag{33}$$

where p_0 is any fixed element of $D_q^- L(x_0,q_0)$. Since the function $p \mapsto \langle p,q_n\rangle - H(x_n,q_n)$ is concave for any $n \in \mathbb{N}$, we see that (33) is still satisfied by putting any convex combination of p_n and p_0 in place of p_n; in particular it holds for some sequence $(\overline{p}_n)_n$ with $|\overline{p}_n - p_0| = r$, where r is an arbitrarily chosen positive constant. Up to subsequences, we can assume that $(\overline{p}_n)_n$ converges to some $\overline{p}$. Sending n to $+\infty$, we obtain

$$L(x_0,q_0) = \langle p_0,q_0\rangle - H(x_0,p_0) \le \langle \overline{p},q_0\rangle - H(x_0,\overline{p}),$$

which implies that $\overline{p} \in D_q^- L(x_0,q_0)$ by Theorem A.2. This is in contrast to $D_q^- L(x_0,q_0)$ being bounded because $|\overline{p} - p_0| = r$ and r is arbitrarily large. □

The argument used for the proof of item (ii) in the previous theorem also gives the following corollary (cf. Remark 2.1).

COROLLARY A.4. *The set-valued map* $(x,q) \mapsto D_q^- L(x,q)$ *is locally uniformly bounded in* $\mathrm{int}(\mathrm{dom}L)$.

Proof of Proposition 2.7. (i) The lower semicontinuous and convex character of L has already been pointed out at the beginning of the appendix.

(ii) By item (i), we need only show that L is upper semicontinuous in Ω. Hence, let $(x_0,q_0) \in \Omega$ be the limit of some sequence $((x_n,q_n))_n$ contained in Ω. Given $p_n \in D_q^- L(x_n,q_n)$, we have that $(p_n)_n$ is bounded by Corollary A.4, and so convergent, up to a subsequence, to some p_0. Thanks to Theorem A.2 we know that

$$L(x_n,q_n) = \langle p_n,q_n\rangle - H(x_n,p_n),$$

and by sending n to infinity we get

$$\limsup_{n\to+\infty} L(x_n,q_n) = \lim_{n\to+\infty} \langle p_n,q_n\rangle - H(x_n,p_n) = \langle p_0,q_0\rangle - H(x_0,\ p_0) \le L(x_0,q_0),$$

which proves the claim.

(iii) Fix $x \in \mathbb{T}^N$. The C^1-regularity of the function $L(x,\cdot)$ in $\mathrm{int}(\mathrm{dom}L(x,\cdot))$ is equivalent to the strict convexity of $H(x,\cdot)$ on $\mathbb{R}^N$ (cf. [7]). In particular,

$$L(x,q) = \langle D_qL(x,q),q\rangle - H(x,D_qL(x,q)) \qquad \text{for all } (x,q) \in \Omega,$$

and $p = D_qL(x,q)$ is the unique maximizer of the function $\langle\cdot\ ,q\rangle - H(x,\cdot)$. To prove the continuity of $D_pL(x,q)$ in Ω, it suffices to show, by Corollary A.4, that

$(x_n, q_n) \to (x_0, q_0)$ in Ω and $D_q L(x_n, q_n) \to \overline{p}$ in $\mathbb{R}^N$ imply $\overline{p} = D_q L(x_0, q_0)$. This actually follows from the continuity of L in Ω, since we can pass to the limit in the equality $L(x_n, q_n) = \langle D_q L(x_n, q_n), q_n\rangle - H(x_n, D_q L(x_n, q_n))$ to obtain $L(x_0, q_0) = \langle \overline{p}, q_0\rangle - H(x_0, \overline{p})$, which gives $\overline{p} = D_p L(x_0, q_0)$ by our previous remarks.

(iv) If the set of maximizers of $p \mapsto \langle p, q\rangle - H(x, p)$ is nonempty, then it reduces to a singleton by the strict convexity of H with respect to p. The assertion thus follows from Theorems A.1 and A.3(ii). □

Let us now define, for each $n \in \mathbb{N}$,

$$H_n(x, p) := H(x, p) + \max\{|p|^2 - n^2, 0\} \qquad \text{for every } (x, p) \in \mathbb{T}^N \times \mathbb{R}^N,$$

and denote by L_n the Fenchel transform of H_n. Note that $(H_n)_n$ is a decreasing sequence of superlinear Hamiltonians, satisfying assumptions (H1), (H2), (H3), uniformly converging to H on compact subsets of $\mathbb{T}^N \times \mathbb{R}^N$. This, in turn, implies that $(L_n)_n$ is an increasing sequence of Lagrangians, defined and continuous on $\mathbb{T}^N \times \mathbb{R}^N$, converging pointwise to L on $\mathbb{T}^N \times \mathbb{R}^N$, and uniformly superlinear at infinity in q, as well (see, e.g., [7]).

THEOREM A.5. *The representation formula* (9) *holds for every* $\phi \in \mathrm{C}(\mathbb{T}^N)$, $t > 0$.

The proof of the theorem is based on a Γ-convergence result (cf. [9]) that we show first. For this, we employ a classical sequential weak compactness criterion in $W^{1,1}$ (see, for instance, Theorem 2.13 of [5]), which is in turn a consequence of the Dunford–Pettis theorem (cf. Theorem 2.11 in [5]).

PROPOSITION A.6. *For any fixed* $x \in T^N$ *and* $t > 0$, *denote by* $X_t(x)$ *the space*

$$\{\gamma \in W^{1,1}([0, t], \mathbb{T}^N),\ \gamma(t) = x\}$$

endowed with the strong topology of $L^1([0, t], \mathbb{T}^N)$. *For any* $\phi \in \mathrm{C}(\mathbb{T}^N)$, *let us set*

$$\mathbb{L}^t(\gamma) := \phi(\gamma(0)) + \int_0^t L(\gamma(s), \dot\gamma(s))\,\mathrm{d}s,$$
$$\mathbb{L}^t_n(\gamma) := \phi(\gamma(0)) + \int_0^t L_n(\gamma(s), \dot\gamma(s))\,\mathrm{d}s.$$

Then the functionals $\mathbb{L}^t_n$ Γ-*converge to* $\mathbb{L}^t$ *on* $X_t(x)$. *Moreover,*

$$\min_{\gamma \in X_t(x)} \mathbb{L}^t(\gamma) = \lim_{n \to +\infty} \min_{\gamma \in X_t(x)} \mathbb{L}^t_n(\gamma).$$

Proof. We first set

$$\Theta(t) := \inf_{x \in \mathbb{T}^N} \left(\inf_{|q| \geq t} L_1(x, q) \right) \qquad \text{for every } t \geq 0$$

and observe that

$$\lim_{t \to +\infty} \frac{\Theta(t)}{t} = +\infty, \qquad \Theta(|q|) \leq L_n(x, q) \leq L(x, q) \tag{34}$$

for any $n \in \mathbb{N}$ and $(x, q) \in \mathbb{T}^N \times \mathbb{R}^N$. We claim that the functionals $\mathbb{L}^t_n$ and $\mathbb{L}^t$ are lower semicontinuous on $X_t(x)$. In fact, any sequence $(\gamma_n)_n$ in $X_t(x)$ with $\lim_n \mathbb{L}^t(\gamma_n) < +\infty$ also satisfies $\sup_n \int_0^t \Theta(\dot\gamma_n)\,\mathrm{d}s < +\infty$ by (34), and this in turn implies that $(\gamma_n)_n$ is weakly convergent in $W^{1,1}([0, t], \mathbb{T}^N)$, up to subsequences (cf. Theorem 2.13 of [5]).

This shows the sequential lower semicontinuity of $\mathbb{L}^t$ in $X_t(x)$, in view of Theorem 2.9; the lower semicontinuity follows as $X_t(x)$ is a metric space. The same argument gives the claim for each $\mathbb{L}^t_n$.

The Γ-convergence result is then assured by [9, Proposition 5.4], since $(\mathbb{L}^t_n)_n$ is, in addition, an increasing sequence of functionals converging pointwise to $\mathbb{L}^t$ on $X_t(x)$. To prove the asserted convergence of the minima, we remark that the set

$$K_t(x) := \left\{ \gamma \in X_t(x) \,:\, \int_0^t \Theta(|\dot\gamma|)\,\mathrm{d}s \le \|\phi\|_\infty + \Bbbk\, t \right\},$$

with $\Bbbk := \sup_{y\in\mathbb{T}^N} L(y,0)$, is sequentially weakly compact in $W^{1,1}([0,t],\mathbb{T}^N)$, and hence compact in $X_t(x)$ because the weak convergence implies the uniform convergence (cf. [5, Theorem 2.13]). Notice also that

$$\int_0^t \Theta(|\dot\gamma_x|)\,\mathrm{d}s \le \mathbb{L}^t_n(\gamma_x) \le \mathbb{L}^t(\gamma_x) \le \|\phi\|_\infty + \Bbbk\, t$$

for any n, where γ_x denotes the curve in $X_t(x)$ constantly equal to x. Consequently $K_t(x)$ is nonempty and

$$\inf\{\, \mathbb{L}^t_n(\gamma) \,:\, \gamma \in X_t(x)\,\} = \min\{\, \mathbb{L}^t_n(\gamma) \,:\, \gamma \in K_t(x)\}$$

for each n, so the assertion follows in view of [9, Theorem 7.4]. □

Proof of Theorem A.5. We first notice that it is enough to show the assertion for $\phi \in \mathrm{Lip}(\mathbb{T}^N)$. The general case of a continuous initial datum may in fact be recovered by density, thanks to the nonexpansiveness property of the Lax–Oleinik semigroup.

We denote by $\mathcal{S}_n(t)$ the semigroup associated with the Cauchy problem (6), with H_n in place of H. Since $\phi \in \mathrm{Lip}(\mathbb{T}^N)$, we have by Theorem 2.5

$$\mathcal{S}(t)\phi = \mathcal{S}_n(t)\phi \tag{35}$$

for n sufficiently large. By Remark 2.8 each $\mathcal{S}_n(t)\phi$ admits an integral representation of the form (9), with L_n in place of L. This fact can be equivalently expressed, using the symbols introduced in Proposition A.6, by

$$\big(\mathcal{S}_n(t)\phi\big)(x) = \min_{\gamma\in X_t(x)} \mathbb{L}^t_n(\gamma)$$

for every $x \in \mathbb{T}^N$ and $t > 0$. In view of Proposition A.6, the assertion follows by sending n to $+\infty$ in (35). □

Acknowledgments. The first author gratefully acknowledges the hospitality and support of the Department of Mathematics of the University of Rome "La Sapienza," where this research was initiated.

REFERENCES

[1] L. Ambrosio, O. Ascenzi, and G. Buttazzo, *Lipschitz regularity for minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1989), pp. 301–316.

[2] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations. With Appendices by Maurizio Falcone and Pierpaolo Soravia*, Systems Control Found. Appl., Birkhäuser Boston, Inc., Boston, MA, 1997.

[3] G. Barles, *Solutions de viscosité des équations de Hamilton–Jacobi*, Math. Appl. (Berlin) 17, Springer-Verlag, Paris, 1994.
[4] G. Barles and P. E. Souganidis, *On the large time behavior of solutions of Hamilton–Jacobi equations*, SIAM J. Math. Anal., 31 (2000), pp. 925–939.
[5] G. Buttazzo, M. Giaquinta, and S. Hildebrandt, *One-dimensional Variational Problems. An Introduction*, Oxford Lecture Ser. Math. Appl. 15, The Clarendon Press, Oxford University Press, New York, 1998.
[6] F. Camilli and A. Siconolfi, *Time-dependent measurable Hamilton–Jacobi equations*, Comm. Partial Differential Equations, 30 (2005), pp. 813–847.
[7] P. Cannarsa and C. Sinestrari, *Semiconcave Functions, Hamilton–Jacobi Equations, and Optimal Control*, Prog. Nonlinear Differential Equations Appl. 58, Birkhäuser Boston, Inc., Boston, MA, 2004.
[8] G. Contreras and R. Iturriaga, *Global Minimizers of Autonomous Lagrangians*, 22nd Brazilian Mathematics Colloquium, Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 1999.
[9] G. Dal Maso, *An Introduction to* Γ*-convergence*, Prog. Nonlinear Differential Equations Appl. 8, Birkäuser Boston Inc., Boston, MA, 1993.
[10] G. Dal Maso and H. Frankowska, *Uniqueness of solutions to Hamilton–Jacobi equations arising in the calculus of variations*, in Optimal Control and Partial Differential Equations, J. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, Amsterdam, 2000, pp. 335–345.
[11] A. Fathi, *Sur la convergence du semi-groupe de Lax–Oleinik*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 267–270.
[12] A. Fathi, *Weak Kam Theorem in Lagrangian Dynamics*, Cambridge University Press, Cambridge, UK, 2004.
[13] A. Fathi and A. Siconolfi, *PDE aspects of Aubry–Mather theory for quasiconvex Hamiltonians*, Calc. Var. Partial Differential Equations, 22 (2005), pp. 185–228.
[14] P. L. Lions, *Generalized solutions of Hamilton–Jacobi equations*, Res. Notes in Math. 69, Pitman (Advanced Publishing Program), Boston, MA; London, 1982.
[15] P. L. Lions, G. Papanicolau, and S. R. S. Varadhan, *Homogenization of Hamilton–Jacobi Equations*, preprint, 1987.
[16] G. Namah and J. M. Roquejoffre, *Remarks on the time-asymptotic behaviour of the solutions of Hamilton–Jacobi equations*, Comm. Partial Differential Equations, 24 (1999), pp. 883–893.
[17] R. T. Rockafellar, *Convex Analysis*, Princeton Math. Ser., Princeton University Press, Princeton, NJ, 1970.
[18] J. M. Roquejoffre, *Convergence to steady states or periodic solutions in a class of Hamilton–Jacobi equations*, J. Math. Pures Appl. (9), 80 (2001), pp. 85–104.
[19] M. Rorro, *An approximation scheme for the effective Hamiltonian and applications*, Appl. Numer. Math., to appear.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 503–529

COARSENING RATES FOR A DROPLET MODEL: RIGOROUS UPPER BOUNDS*

FELIX OTTO†, TOBIAS RUMP†, AND DEJAN SLEPČEV‡

Abstract. Certain liquids on solid substrates form a configuration of droplets connected by a precursor layer. This configuration coarsens: The average droplet size grows while the number of droplets decreases and the characteristic distance between them increases. We study this type of coarsening behavior in a model given by an evolution equation for the film height on an n-dimensional substrate. Heuristic arguments based on the asymptotic analysis of Glasner and Witelski [*Phys. Rev. E,* 67 (2003), p. 016302, *Phys. D.,* 209 (2005), pp. 80–104] and numerical simulations suggest a statistically self-similar behavior characterized by a single exponent which determines the coarsening rate. In this paper, we establish rigorously an upper bound on the coarsening rate in a time-averaged sense. We use the fact that the evolution is a gradient flow, i.e., a steepest descent in an energy landscape. Coarse information on the geometry of the energy landscape serves to obtain coarse information on the dynamics. This robust method was proposed in [R. V. Kohn and F. Otto, *Comm. Math. Phys.,* 229 (2002), pp. 375–395]. Our main analytical contribution is an interpolation inequality involving the Wasserstein distance, which characterizes the coarse shape of the energy landscape. The upper bound we obtain is in agreement with heuristic arguments and numerical simulations.

Key words. thin film equation, Wasserstein distance, coarsening

AMS subject classifications. 76A20, 35K55, 35Q35

DOI. 10.1137/050630192

1. Introduction.

1.1. Statement of the result.

Thin layers of viscous liquid are well described by the lubrication approximation, which capitalizes on the separation of horizontal and vertical length scales. It yields a single equation for the time-dependent film height $h = h(x,t) > 0$, [20]. We now introduce this equation in its nondimensionalized form.

Let $Q \subset \mathbb{R}^n$ parametrize the substrate. We consider smooth solutions $h : (0,\infty) \times Q \to [0,\infty)$ of

$$\partial_t h - \nabla \cdot \left(M(h) \nabla \left(\frac{\partial E}{\partial h} \right) \right) = 0 \quad \text{in } (0,\infty) \times Q. \tag{1.1}$$

Here, $\frac{\partial E}{\partial h}$ denotes the L^2-gradient of the energy functional with respect to h. The total energy E is given by

$$E(h) = \int \frac{1}{2} |\nabla h|^2 + \mathcal{U}(h)\, dx, \tag{1.2}$$

where the gradient term describes the linearized contribution of the liquid-air surface energy, while $\mathcal{U}$ models the intermolecular forces between the substrate and the film; see section 1.3. We shall always write $\int$ for $\int_Q$.

*Received by the editors April 28, 2005; accepted for publication (in revised form) January 10, 2006; published electronically June 9, 2006.

http://www.siam.org/journals/sima/38-2/63019.html

†Institute for Applied Mathematics, University of Bonn, Wegelerstraße 10, 53115 Bonn, Germany (otto@iam.uni-bonn.de, rump@iam.uni-bonn.de). The authors acknowledge partial support by the Sonderforschungsbereich 611 Singular Phenomena and Scaling in Mathematical Models at the University of Bonn.

‡Mathematics Department, UCLA, Box 951555, Los Angeles, CA 90095 (slepcev@math.ucla.edu). This author acknowledges support by NSF grant DMS-0244498 and ONR grant N000140410078.

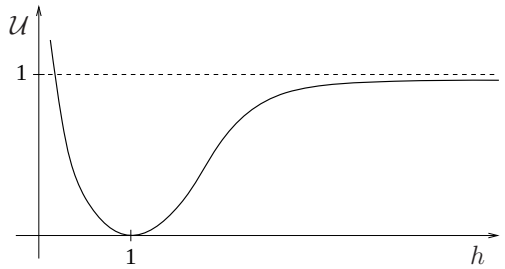


FIG. 1. *The intermolecular potential $\mathcal{U}$ combines repulsive and attractive forces.*

Qualitatively, the potential $\mathcal{U}$ is of the form sketched in Figure 1. We normalize $\mathcal{U}$ so that $\mathcal{U}(\infty) - \min \mathcal{U} = 1$ and $\operatorname{argmin} \mathcal{U} = 1$. The well-known Lennard–Jones potential [9] corresponds to $\mathcal{U}(h) = \frac{1}{3}h^{-8} - \frac{4}{3}h^{-2} + 1$, but other potentials are also physically relevant [24]. In fact, the exact expression for the potential is often not essential; as in [2, 6, 7], our results hold for a large class of potentials.

Bertozzi, Grün, and Witelski [2] studied (1.1) for potentials of the form $\mathcal{U}(h) = h^{-q} - h^{-p}$, $0 < p < q$ on a bounded one-dimensional substrate. If $q \geq 2$—that is, if the potential blew up at zero fast enough—boundedness of the energy was used to show that solutions with positive initial data stay positive for all time. That implied that the equation has unique, classical solutions for given initial data.

In our model (1.1), we consider the linear mobility function

$$M(h) = h. \tag{1.3}$$

Let us clearly state that the appropriate mobility function for a liquid film governed by the Stokes equations with no-slip boundary condition at the substrate would be $M(h) = h^3$. In case of boundary conditions which allow a finite slip (Navier condition), the mobility would be $M(h) = h^2$ provided that film heights are small compared to the slip length; see [16, 17, 20]. The mobility function (1.3) is appropriate for a liquid film governed by the Darcy equation, as in a porous medium ($n = 2$) or a Hele–Shaw cell ($n = 1$). From the applied point of view, $M(h) = h$ is thus a rather artificial choice. It is motivated both by a technical and a conceptual consideration: The technical consideration is that only for $M(h) = h$ is the induced distance in the gradient flow structure (see Appendix B) known explicitly. The conceptual consideration is that for other mobilities the coarse-grained slope of the energy landscape (see section 2) overestimates the heuristically derived coarsening rates: The "collision pathways" (see section 1.3) in the energy landscape are shortcuts not taken by the actual dynamics. Hence the straightforward application of our method would yield suboptimal results for mobilities other than (1.3), even though we believe that the coarsening rate for $M(h) = h^3$ is the same as for $M(h) = h$.

In view of (1.2) and (1.3), (1.1) turns into

$$\partial_t h + \nabla \cdot (h \nabla (\Delta h - \mathcal{U}'(h))) \;=\; 0 \quad \text{in } (0,\infty) \times Q. \tag{1.4a}$$

As suitable boundary conditions, we take equilibrium and no-flux boundary conditions:

$$\nu \cdot \nabla h = \nu \cdot \nabla \left(\frac{\partial E}{\partial h} \right) \;=\; 0 \quad \text{on } (0,\infty) \times \partial Q. \tag{1.4b}$$

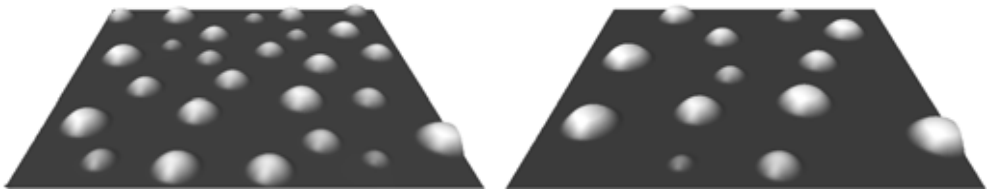


FIG. 2. *Plots of numerically computed height of the liquid on a two-dimensional substrate at two different times.*

These boundary conditions ensure that the total mass $\int h\ dx$ is conserved and that the total energy $E(h)$ decreases over time:

$$\frac{d}{dt}\int h\ dx\ =\ 0,$$

$$\frac{d}{dt}E\ =\ -\int h\Big|\nabla\left(\frac{\partial E}{\partial h}\right)\Big|^2\ dx\ \leq\ 0.$$

A glance at the energy functional E, and $\mathcal{U}$ in particular, reveals that there is a characteristic scale for x and h, both normalized to order one. We are interested in very large systems, i.e.,

$$Q=(0,\Lambda)^n \quad \text{with} \quad \Lambda\gg 1.$$

We focus on initial data which are within the unstable range,

$$h\ \approx\ \text{const} \quad \text{with} \quad \mathcal{U}''(h)<0,$$

and whose height is of the order of the precursor layer, i.e.,

$$h\ \sim\ 1.$$

Our numerical simulations reveal the generic behavior of the evolution. After an initial stage, a configuration of well-defined droplets connected by a precursor layer of height $h\approx 1=\operatorname{argmin}\mathcal{U}$ emerges. From then onward, the large droplets grow at the expense of the small ones via mass exchange through the precursor layer. Eventually, the smaller droplets disappear. Figure 2 shows a typical evolution of the film height. In a sufficiently large system, this coarsening process seems statistically self-similar. It is driven by the reduction of total energy E. Our main result gives a lower bound on the rate by which the energy decreases.

We shall see that E not only is a Lyapunov functional for (1.4), but that (1.4) can, in fact, be interpreted as a gradient flow of (1.2); see Appendix B. For our analysis, we need a measure of the distance in the configuration space, i.e., a way to express how far two droplet configurations h_0, h_1 are. It is natural to take the distance which is given by the gradient flow structure. As motivated in Appendix B, that distance is the Wasserstein distance $\mathcal{W}(h_0,h_1)$:

(1.5)
$$\mathcal{W}(h_0,h_1)^2\ :=\ \inf\left\{\iint|x-y|^2d\pi(x,y)\ \Big|\ \int\ d\pi(\cdot,y)=h_0,\ \int\ d\pi(x,\cdot)=h_1\right\}.$$

The so-called transportation plan, π, is a measure on the product space $Q \times Q$. It is admissible if its projections to first and second coordinates are measures with densities h_0 and h_1, respectively. The transportation cost is measured by the squared Euclidean distance $|x-y|^2$. For properties of $\mathcal{W}$, see section 7.1 in [32].

We are now in a position to formulate our result.

THEOREM 1. *Let $\mathcal{U}$ satisfy*

$$\begin{aligned} &\mathcal{U}(h) \geq 0 \quad \text{on } (0,\infty), \\ &\mathcal{U}(h) \geq 1 \quad \text{on } (2,\infty). \end{aligned} \tag{1.6}$$

Let h be a smooth solution of (1.4)*, which has the same total volume as a constant layer of thickness* 3*:*

$$\int h\,dx \;=\; \int h^*\,dx, \quad h^* \equiv 3. \tag{1.7}$$

Then for $\sigma \in (1, \frac{3n+2}{n})$,

$$\int_0^T \left(\Lambda^{-n} E(h(t))\right)^\sigma \,dt \quad \gtrsim \quad \int_0^T (t^{-\frac{n}{3n+2}})^\sigma \,dt, \tag{1.8}$$

provided $T \gg (\Lambda^{-\frac{n}{2}} \mathcal{W}(h^*, h(0)))^{\frac{3n+2}{n+1}}$ and $\Lambda^{-n} E(h(0)) \ll 1$.

For the precise meaning of the notation "$\gg$" and "$\gtrsim$" we refer to Remark 1 in section 2. The result states that the energy per volume cannot decrease faster than $t^{-\frac{n}{3n+2}}$. For further interpretation of the result in terms of the droplet configuration we refer to section 1.4. The assumption $\Lambda^{-n} E(h(0)) \ll 1$ of small energy densities encodes that $h(0)$ energetically behaves as a configuration of droplets connected by a precursor layer of height $\operatorname{argmin} \mathcal{U} = 1$; see Figure 2. This means that we "start the clock" once the system has entered such a regime. The assumption $T \gg \left(\Lambda^{-\frac{n}{2}} \mathcal{W}(h^*, h(0))\right)^{\frac{3n+2}{n+1}}$, on the other hand, ensures that the initial data $h(0)$ are not too far from a constant film thickness. The values 1, 2, and 3 in (1.6) and (1.7) are set purely for convenience. In particular, given a potential satisfying (1.6), any constant layer of thickness above 2 is admissible. However, the constants in the estimates would depend on the thickness.

The framework for proving lower bounds on energy decay was introduced by Kohn and Otto [10] for the constant-mobility and the degenerate-mobility Cahn–Hilliard equation. The basic idea is to use the gradient flow structure of (1.4). A gradient flow structure is determined by the energy functional E and a (Riemannian) geometry of the state space (the space of all droplet configurations h). The metric tensor encodes the relevant dissipation mechanism, see Appendix B. Following [10], we use coarse information on the geometry of the energy landscape to derive coarse information on the gradient flow dynamics. The coarse information on geometry is the rate at which E can decrease as a function of the distance to a reference configuration h^*, where the function is given by a power law with the "geometric exponent" α; see Figure 3. On the other hand, the coarse information on dynamics limits how fast E can decrease as a function of time, where the function is given by a power law with the "dynamic exponent" γ. Proposition 1 in section 2 relates the dynamic exponent γ to the geometric exponent α.

Let us point out that studying *lower* bounds on the coarsening rate is a more complicated question. In fact, there are solutions which do not coarsen at all; for

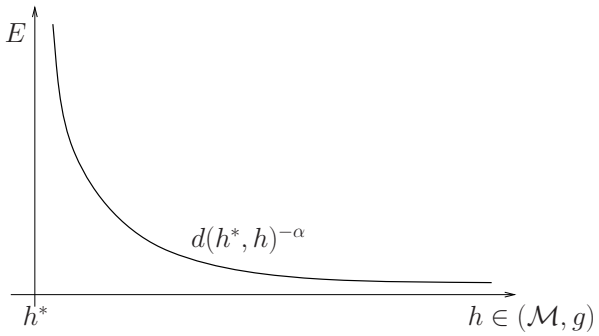


FIG. 3. *The energy $E(h)$ is bounded below by the distance between h and the reference state h^*.*

example, the unstable, periodic steady states. Hence a lower bound can hold only in some generic sense—a statement for all trajectories cannot hold. We do not address this issue.

Other applications of the method proposed in [10] can be found in the literature. In [3], coarsening in off-critical mixtures within the Mullins–Sekerka evolution is studied. The authors of [4] and [5] study the coarsening behavior in mean-field models of phase transitions and in a phase-field model that deals with both temperature and phase fields. In [11] and [12], rigorous bounds on coarsening rates are proven for an epitaxial growth model and for models of multicomponent phase separation.

1.2. Outline. The following subsections provide background information: In section 1.3, we describe the underlying physical processes. Heuristics, which further interpret the result in Theorem 1, are presented in section 1.4. Formal asymptotics, e.g., done for the one-dimensional case in [6], are briefly reviewed in section 1.5. In section 1.6, we present our numerical experiments which support that the power law bounds we obtain are optimal for $n = 1$ and $n = 2$. The proof of Theorem 1 does not depend on the results presented in these subsections.

In section 2, we review the abstract framework, which exploits the gradient flow structure to obtain a bound on the energy, proposed in [10], and give modified proofs of the main results. Our main contribution is the interpolation inequality, which is essential for the framework. From the mathematical point of view, it is an extension of the one established in [3]. We rigorously state and prove the inequality in section 3. Section 4 provides the proof of the main result. Appendix A contains the analysis of stationary droplets, while Appendix B explains the gradient flow structure of the thin-film equation. The Wasserstein metric, as the induced distance, is introduced heuristically.

1.3. Physics. The statics are determined by capillarity (the surface tension between liquid and vapor) and the short range forces between the film and the solid substrate. These intermolecular forces are the combination of a very short-range repulsive (Born-type) force and a moderately short-range attractive (van der Waals) force. Their competition stabilizes a precursor layer of well-defined height, which covers the entire substrate. In particular, our choice of $\mathcal{U}$ with $\operatorname{argmin} \mathcal{U} = 1$ yields a precursor layer of height ≈ 1. On a more mesoscopic level, the liquid is partially wetting: It allows for equilibrium droplets of a well-defined apparent (i.e., mesoscopic) contact angle. The apparent nonzero contact angle arises from the competition between the short range forces and capillarity.

Intermolecular forces of the form discussed above are relevant for a range of liquids: polymers [1, 26, 27, 28, 29, 33], liquid crystal films and liquid metals [8, 31], evaporating films [19], and others.

The dynamics is driven by the reduction of total energy (capillary and intermolecular). The reduction of energy is limited by viscous friction. This dissipation mechanism is rather pronounced in the thin liquid film. This means that inertial effects are negligible and the dynamics is determined by a quasi-stationary balance of thermodynamic driving and viscous frictional forces. Notice that the presence of the precursor layer removes the singularity of the moving contact line which arises from the no-slip boundary condition.

We think of the initial condition as a flat film of height sufficiently large with respect to the equilibrium height of the precursor layer. More precisely, we think of a perturbation of this configuration; due to the intermolecular forces, sufficiently long-wavelength perturbations grow. In analogy with spinodal decomposition of binary mixtures (as described by the Cahn–Hilliard equation) this process is often referred to as spinodal dewetting. The liquid film almost ruptures and holes (with film thickness of the precursor layer) surrounded by a network of ridges form. This initial process may have various morphologies [1, 30]. As time passes, the ridges break up and relax into droplets sitting on the precursor layer. In this paper, we are not concerned with this initial stage of droplet formation.

We are interested in the late-stage *coarsening* behavior. A configuration of well-separated droplets connected by a precursor layer coarsens in time; see Figure 2. We have in mind a scenario with clearly separated time scales: The time scale of the coarsening process is slow compared to the scale on which the droplets relax into equilibrium shape. Hence the configuration is essentially described by the radius and the position of the center of mass of the individual droplets. These quantities evolve slowly. This is sometimes called quasi-static evolution or quasi-stationary motion.

The coarsening process can be mediated by two mechanisms: collapse or collision. Collapse relies on mass exchange between the droplets through the precursor layer. In this scenario, the large droplets grow at the expense of the small ones which eventually collapse. This is a particular instance of Ostwald ripening. The basic difference with respect to the traditional Ostwald ripening for binary mixtures (as described by the late stages of the Cahn–Hilliard equation with strongly off-critical initial data; see for instance [18] for the two-dimensional case) lies in the mixed dimensionality: Ripening of droplets on an n-dimensional substrate is $(n+1)$-dimensional with respect to mass and energy, but n-dimensional when it comes to the kinetics.

We now address collision: As do the particles in traditional Ostwald ripening, the droplets drift. The difference with respect to traditional Ostwald ripening lies in the fact that droplets are much more mobile than particles (since the mobility strongly depends on height). In [6], it has been argued that on a one-dimensional substrate, this effect may lead to collision and thus to "accidental" coarsening.

Unlike for the initial instability and subsequent dewetting, there are few experimental studies of coarsening in liquid films. The only long-time results we are aware of are studies of coarsening for certain polymers [14, 15].

1.4. Dynamical scaling: Heuristics. Numerical simulations suggest that the coarsening, although rather complex in detail, has a simple statistical behavior (see section 1.6). In particular, the time-dependence of averaged quantities, like the average distance L between droplets, in sufficiently large systems appears to be a power

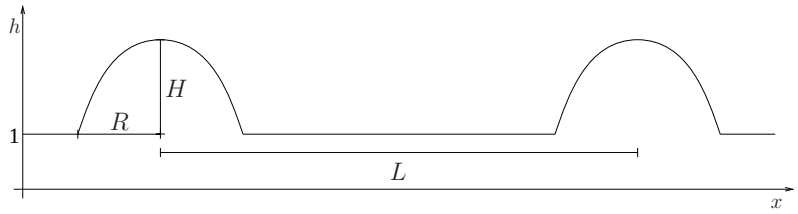


FIG. 4. *Droplet configuration on one-dimensional substrate. The typical length scales H, R, and L are not independent; cf.* (1.10) *and* (1.11).

law:

$$L \sim t^{\beta}.$$

We are interested in understanding the mechanisms which determine the characteristic exponent β. In order to discern what information about β is contained in the energy bound (1.8), we first analyze and relate the length scales present in the problem. We consider configurations of well-separated, equilibrium-shaped droplets with fixed average height greater than the height of the precursor layer. The typical averaged length scales are the typical distance L, the height H, and radius R of a droplet. Figure 4 sketches a typical droplet configuration we have in mind.

From this point of view, we infer the following scaling relations:

(i) On a mesoscopic level, the potential $\mathcal{U}$ acts as the characteristic function of $\{h > 1\}$. Thus the individual droplets are governed by the mesoscopic energy

$$\overline{E}(h) = \int \tfrac{1}{2}|\nabla h|^2 \, dx + \text{ vol}\,(\{h > 1\}),$$

which enforces an apparent equilibrium contact angle of

$$\tfrac{1}{2}|\nabla h|^2 = 1 \quad \text{ on } \partial\{h > 1\}; \tag{1.9}$$

see Appendix A. Notice that (1.9) implies that the average height H and the average radius R of the droplets scale the same:

$$H \sim R. \tag{1.10}$$

(ii) We infer from mass conservation, i.e.,

$$\Lambda^{-n} \int h \, dx \sim 1,$$

and

$$\begin{aligned} \Lambda^{-n} \int h \, dx &\sim \begin{pmatrix} \text{number density} \\ \text{of droplets} \end{pmatrix} \times \begin{pmatrix} \text{volume of} \\ \text{individual droplet} \end{pmatrix} \\ &\sim L^{-n} \times HR^n \\ &\overset{(1.10)}{\sim} L^{-n} \times R^{n+1}, \end{aligned}$$

that R and L are related by

$$L \sim R^{\frac{n+1}{n}}. \tag{1.11}$$

Now let us explain why a lower bound on the energy heuristically yields an upper bound on the average droplet distance L. Indeed, the energy density (i.e., the energy per system volume) is related to L via

$$\begin{aligned}\Lambda^{-n}E &\sim \left(\begin{array}{c}\text{number density}\\ \text{of droplets}\end{array}\right)\times\left(\begin{array}{c}\text{energy of}\\ \text{individual droplet}\end{array}\right)\\ &\overset{(1.10)}{\sim} L^{-n}\times R^n\\ &\overset{(1.11)}{\sim} L^{-\frac{n}{n+1}}.\end{aligned}\tag{1.12}$$

Thus a lower bound on E heuristically yields an upper bound on L. The energy bound (that we show in a time-averaged form in Theorem 1)

$$\Lambda^{-n}E \gtrsim t^{-\frac{n}{3n+2}}$$

in view of (1.12) heuristically amounts to the upper bound

$$\beta \le \frac{n+1}{3n+2} = \begin{cases}\frac{2}{5} & \text{for } n=1,\\ \frac{3}{8} & \text{for } n=2.\end{cases}\tag{1.13}$$

Based on numerical and heuristic predictions (which are solid for $n=1$ and somewhat less so for $n=2$) this bound is optimal (up to logarithmic correction in two dimensions). In the nonphysical case of $n\ge 3$, we do not expect the bound to be optimal.

1.5. Formal asymptotics. The coarsening process on one-dimensional substrates was analytically studied by Glasner and Witelski [6, 7] for physically relevant mobility h^3. Based on the quasi-static assumption, they derive a system of coupled ODEs for droplet pressures $\{P_i\}$ (which are in one-to-one correspondence to their radius) and droplet positions $\{X_i\}$. Using the scale separation $H\sim R\ll L$ they argue heuristically that

$$L \sim t^{\frac{2}{5}}.$$

Their numerical experiments confirm this scaling. Furthermore, Glasner and Witelski show that in certain regimes both collapse and collision of droplets are possible mechanisms of coarsening.

We verified that their heuristics extends to our mobility $M(h)=h$ and yields the same scaling. It is a particularity of the mobility $M(h)=h^3$ that both processes, collapse and collision, yield the same exponent (modulo a logarithm). For $M(h)=h$, however, collapse is eventually the faster process.

A back-of-the-envelope argument for two-dimensional substrates based on the collapse scenario suggests that

$$L \sim t^{\frac{3}{8}}$$

(modulo a logarithm specific to the fundamental solution of a Laplace equation in two dimensions; see [18]). The importance of collisions, as it depends on $M(h)$, is not yet well understood, in our opinion, despite the investigation of droplet mobility in [25].

Our analysis does not rely on a derivation of a reduced model based on a quasi-static assumption. Likewise, it does not presuppose that coarsening is simple on a statistical level.

1.6. Numerics. To gain some understanding of the coarsening dynamics we carried out several numerical experiments. Let us first address the discretization of the thin-film equation (1.4). To solve (1.4) in one and two dimensions we use a modification of the discretization of a Cahn–Hilliard-type equation proposed in [21]. This discretization approach is guided by two features of the model (1.4):

(i) The total mass $\int h$ is preserved due to the continuity equation

$$\partial_t h + \nabla \cdot J \;=\; 0, \tag{1.14}$$

where $J = -h\nabla\left(\frac{\partial E}{\partial h}\right)$.

(ii) The energy E is a Lyapunov functional:

$$\frac{d}{dt}E \;=\; -\int \frac{1}{h}|J|^2 \;\le\; 0.$$

We use a semi-implicit time discretization (explicit in the mobility). Since the time-discrete equation is nonlinear we apply a single Newton step:

$$h^{k+1} - h^k - \tau\nabla\cdot\left[h^k\nabla\left(\frac{\partial^2 E}{\partial h^2}(h^k)(h^{k+1}-h^k) + \frac{\partial E}{\partial h}(h^k)\right)\right] \;=\; 0.$$

By introducing the flux J,

$$h^{k+1} \;=\; h^k - \tau\nabla\cdot J^{k+1},$$

we obtain a symmetric problem

$$\left[\frac{1}{h^k}\mathrm{id} - \tau\nabla\left(\frac{\partial^2 E}{\partial h^2}(h^k)\right)\nabla\cdot\right] J^{k+1} = -\nabla\left(\frac{\partial E}{\partial h}(h^k)\right).$$

We use a finite difference scheme for the fourth-order problem and solve the linear system by the conjugate gradient method, preconditioned by the constant coefficient operator $\mathrm{id} - \tau\nabla(\frac{\partial^2 E}{\partial h^2}(1))\nabla\cdot$, which we invert by FFT.

Now we turn to the numerical experiments. We use the potential

$$\mathcal{U}(h) = 2h^{-3} - 3h^{-2} + 1,$$

which was used in [6, 19], and take as initial data

$$h(t=0) = h^* + \text{perturbation},$$

where we choose $h^* \equiv 2$, which is in the concave part of $\mathcal{U}$ but still of the order of the precursor layer thickness. System sizes are chosen $\Lambda = 10000$ for $n = 1$ and $\Lambda = 1000$ for $n = 2$.

Figure 5 shows a logarithmic plot of the energy density of the configurations versus time both in one and two dimension. Furthermore, we measure the droplets; see Figure 6. Note that the data is averaged over ten runs.

Our numerical experiments reveal the scaling exponent $\beta = \frac{2}{5}$ for $n = 1$ (see Figure 6, left), since the number density $\Lambda^{-n}N$ scales like L^{-n}. This exponent is equal to the upper bound we obtained in (1.13). Experiments for $n = 2$ (see Figure 6, right) suggest a faster decrease of the number density than for $n = 1$, as predicted. The coarsening exponent is in agreement with the bound (1.13) but appears to be

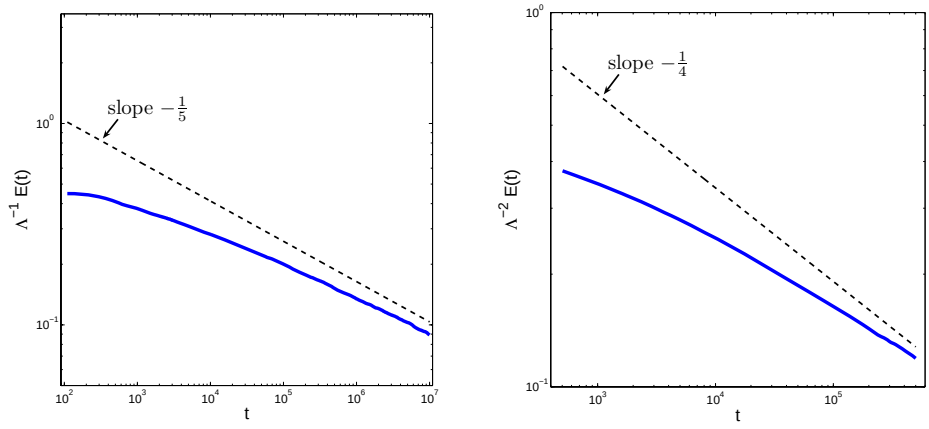


FIG. 5. *Energy density of droplet configuration for $n = 1$ (left) and for $n = 2$ (right) versus time.*

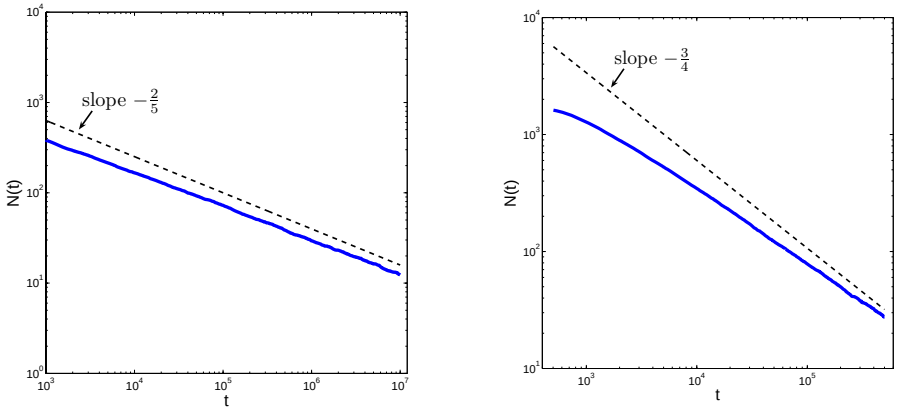


FIG. 6. *Number of droplets for $n = 1$ (left) and for $n = 2$ (right) versus time.*

slightly different than the bound itself. Let us comment on this apparent discrepancy. The heuristics in section 1.4 are based on the assumption that

$$\begin{aligned}\Lambda^{-n}E &\overset{(1.12)}{\sim} L^{-\frac{n}{n+1}}\\ &\sim \left(\Lambda^{-n}N\right)^{\frac{1}{n+1}}.\end{aligned}$$

We therefore monitor the system averaged quantity

$$\Lambda^{-n}E\left(\Lambda^{-n}N\right)^{-\frac{1}{n+1}}$$

over time. For an infinite system this number should reach an asymptotic value if coarsening is statistically self-similar. Figure 7 shows that for $n = 2$ the numerical simulations have barely reached an asymptotic state. Hence for $n = 2$ a numerical confirmation of the optimality of our result would require much larger time horizons and thus much larger system sizes.

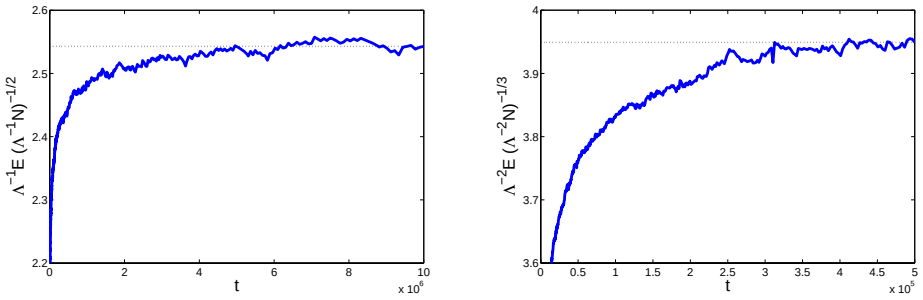


FIG. 7. *The system averaged quantity* $\Lambda^{-n}E\left(\Lambda^{-n}N\right)^{-\frac{1}{n+1}}$ *for* $n=1$ *(left) and for* $n=2$ *(right) versus time.*

2. Abstract framework: From geometry to dynamics. In this section we show how coarse information on the *geometry* of the energy landscape (see Figure 3) leads to coarse information on the gradient flow *dynamics*. Proposition 1 relates the dynamic exponent γ to the geometric exponent α:

$$\gamma \;=\; \frac{\alpha}{\alpha+2}.$$

This insight is essentially from [10]. We give a somewhat different presentation here and include the modified proofs for the convenience of the reader. For the clarity of presentation, we adopt an abstract framework: Let $\mathcal{M}$ be a manifold endowed with a metric tensor g and a function E. We denote by d the induced distance on the Riemannian manifold $(\mathcal{M}, g)$.

PROPOSITION 1 (see [10]). *Let* $h^* \in \mathcal{M}$. *Let* $h : \mathbb{R}_+ \to \mathcal{M}$ *be a solution of*

$$\partial_t h \;=\; -\mathrm{grad}E(h), \tag{2.1}$$

and $h(0) = h_0$.

Assume that for some $\alpha > 0$ *the interpolation inequality*

$$E(h)d(h,h^*)^\alpha \geq 1 \quad \textit{for all } h \in \mathcal{M} \textit{ with } E(h) \leq 1 \tag{2.2}$$

holds. Then for $\sigma \in (1, 1+\frac{2}{\alpha})$

$$\int_0^T E(h(t))^\sigma \, dt \gtrsim \int_0^T (t^{-\frac{\alpha}{\alpha+2}})^\sigma \, dt \tag{2.3}$$

provided $T \gg d(h_0,h^*)^{\alpha+2}$ *and* $E(h(0)) \leq 1$.

Remark 1. The notation $\gtrsim$ and $\gg$ stands for the following.

For all $\sigma \in (1, 1+\frac{2}{\alpha})$ there exists a constant $C = C(\alpha,\sigma)$ such that for all $\delta > 0$ $\exists C_\delta = C(\alpha,\sigma,\delta)$,

$$\int_0^T E(h(t))^\sigma \, dt \geq (1-\delta)C \int_0^T (t^{-\frac{\alpha}{\alpha+2}})^\sigma \, dt \tag{2.4}$$

provided $T \geq C_\delta d(h_0,h^*)^{\alpha+2}$.

Remark 2. It is not true that (2.1) and (2.2) imply the pointwise estimate

$$E(t) \;\gtrsim\; t^{-\frac{\alpha}{\alpha+2}}. \tag{2.5}$$

Indeed, let $\mathcal{M} = \mathbb{R}_+$, $h^* = 0$, and $h(0) = 1$. For a given $b \gg 1$ let E_b be equal to $h^{-\alpha}$ outside the interval $(1, b)$ and let it be linear on $[1, b]$ so that E_b is continuous:

$$E_b(h) := \begin{cases} 1 + \frac{b^{-\alpha}-1}{b-1}(h-1) & \text{on } [1,b], \\ h^{-\alpha} & \text{otherwise.} \end{cases}$$

Then

$$\partial_t h \;=\; -\frac{dE_b}{dh}(h(t)) = -\frac{b^{-\alpha}-1}{b-1}$$

as long as $h(t) \le b$. Hence

$$h(t) = 1 - \frac{b^{-\alpha}-1}{b-1}t$$

and thus $h(t_b) = b$ for $t_b := \frac{(b-1)^2}{1-b^{-\alpha}}$. Therefore

$$\frac{E_b(h(t_b))}{t_b^{-\frac{\alpha}{\alpha+2}}} \;=\; \frac{(b-1)^{\frac{2\alpha}{\alpha+2}}}{b^\alpha (1-b^{-\alpha})^{\frac{\alpha}{\alpha+2}}} \;\le\; 2b^{-\frac{\alpha^2}{\alpha+2}} \to 0$$

as $b \to \infty$.

Remark 3. The range for $1 < \sigma < 1 + \frac{2}{\alpha}$ is (almost) optimal.

The example above can be used to show that for $0 < \sigma < 1$ the statement cannot hold. An elementary (but lengthy) calculation shows that for $T = b^{2+\alpha-\eta}$, with $0 < \eta < \alpha(1-\sigma)$,

$$\int_0^T (E_b(h(t)))^\sigma dt \sim b^{2+\alpha(1-\sigma)-\eta},$$
$$\int_0^T t^{-\frac{\alpha\sigma}{\alpha+2}} dt \sim b^{2+\alpha(1-\sigma)-\eta+\frac{\alpha\sigma\eta}{\alpha+2}}.$$

Thus inequality (2.3) cannot hold when $0 < \sigma < 1$. The case $\sigma = 1$ remains open. The proof we present for $\sigma > 1$ does not extend to $\sigma = 1$ since the constant C in (2.4) vanishes as σ approaches 1; see (2.12).

For $\sigma = 1 + \frac{2}{\alpha}$ the statement would hold if the lower bound 0 in the integrals was replaced by 1. This follows from the continuity of the functionals with respect to σ and from the fact that constant C is bounded away from 0 as σ approaches $1 + \frac{2}{\alpha}$; see (2.12). The range $\sigma > 1 + \frac{2}{\alpha}$ is not of interest, since the integral from 1 to infinity of the right-hand side is finite, and hence the inequality (2.3) would contain no information on the decay rate of E.

We now consider an arbitrary but fixed trajectory $h(t)$ of (2.1). The following lemma restricts the rate at which the distance $d(h, h^*)$ between $h(t)$ and any fixed $h^* \in \mathcal{M}$ can change. We set for convenience

$$E(t) := E(h(t)),$$
$$D(t) := d(h(t), h^*).$$

LEMMA 1. *Let h be a solution of* (2.1) *and $h^* \in \mathcal{M}$.*
Then

$$\left|\tfrac{d}{dt} D(t)\right|^2 \;\le\; -\tfrac{d}{dt} E(t). \tag{2.6}$$

Proof of Lemma 1. The triangle inequality and the definition of d imply for any $\delta \in \mathbb{R}$

$$\begin{aligned}\left|\frac{1}{\delta}(D(t+\delta)-D(t))\right| &\le \frac{1}{\delta} d(h(t+\delta),h(t)) \\ &\le \frac{1}{\delta}\int_0^1 \sqrt{g_{h(t+s\delta)}(\partial_t h(t+s\delta)\delta, \partial_t h(t+s\delta)\delta)}\, ds \\ &= \int_0^1 \sqrt{g_{h(t+s\delta)}(\partial_t h(t+s\delta), \partial_t h(t+s\delta))}\, ds.\end{aligned}$$

Hence we obtain in the limit $\delta \to 0$

$$\left|\frac{d}{dt}D(t)\right| \;\le\; \sqrt{g_{h(t)}(\partial_t h(t), \partial_t h(t))}.$$

Furthermore

$$\tfrac{d}{dt}E(t) = g_{h(t)}(\mathrm{grad}E(t), \partial_t h(t)) \overset{(2.1)}{=} -g_{h(t)}(\partial_t h(t), \partial_t h(t)),$$

so that we conclude

$$\left|\tfrac{d}{dt}D(t)\right|^2 \;\le\; -\tfrac{d}{dt}E(t). \qquad \square$$

Proof of Proposition 1. Since E is a monotone function of time, $D(t)$ can be viewed as a function of $E(t)$. To distinguish the argument of this function from the actual value of the energy we write $D = D(e)$. Hence (2.6) turns into

$$1 \;\ge\; \left(\frac{dD}{de}\right)^2 |\dot{E}|. \tag{2.7}$$

Multiplying (2.7) by $E(t)^\sigma$ and integrating in t yield

$$\int_0^T E(t)^\sigma \, dt \ge \int_0^T E(t)^\sigma \left(\frac{dD}{de}\right)^2 |\dot{E}| \, dt = \int_{E_T}^{E_0} e^\sigma \left(\frac{dD}{de}\right)^2 \, de, \tag{2.8}$$

where we have set $E_0 = E(0)$ and $E_T = E(T)$.

From the Cauchy–Schwarz inequality we obtain

$$\left(\int_{E_T}^{E_0} e^\sigma \left(\frac{dD}{de}\right)^2 \, de \int_{E_T}^{E_0} e^{-\sigma} \, de\right)^{\frac{1}{2}} \;\ge\; \left|\int_{E_T}^{E_0} \frac{dD}{de} \, de\right| \;=\; |D_0 - D_T|,$$

where we define $D_0 = D(0)$ and $D_T = D(T)$. Substituting in (2.8) and integrating $\int_{E_T}^{E_0} e^{-\sigma} \, de = (\sigma-1)^{-1}\left(E_T^{1-\sigma} - E_0^{1-\sigma}\right)$ implies

$$\begin{aligned}\int_0^T E(t)^\sigma \, dt &\ge (\sigma-1)\left(E_T^{1-\sigma} - E_0^{1-\sigma}\right)^{-1} (D_0 - D_T)^2 \\ &\ge (\sigma-1) E_T^{\sigma-1} (D_0 - D_T)^2. \end{aligned} \tag{2.9}$$

Here we have used the assumption $\sigma > 1$. We rewrite the right-hand side of (2.9) as

$$(\sigma-1) E_T^{\sigma-1-\frac{2}{\alpha}} (E_T D_T^\alpha)^{\frac{2}{\alpha}} (1 - \tfrac{D_0}{D_T})^2$$

and apply the interpolation inequality (2.2), so that

$$\int_0^T E(t)^\sigma \, dt \geq (\sigma-1) E_T^{\sigma-1-\frac{2}{\alpha}} (1-\tfrac{D_0}{D_T})^2. \tag{2.10}$$

Set for abbreviation

$$f(T) := \int_0^T E(t)^\sigma \, dt.$$

Then (2.10) turns into

$$\begin{aligned} f(T) &\geq (\sigma-1) f'(T)^{(\sigma-1-\frac{2}{\alpha})/\sigma} (1-\tfrac{D_0}{D_T})^2 \\ &= (\sigma-1) f'(T)^{\frac{\sigma\alpha-\alpha-2}{\sigma\alpha}} (1-\tfrac{D_0}{D_T})^2, \end{aligned}$$

or equivalently,

$$f(T)^{\frac{\sigma\alpha}{\alpha+2-\sigma\alpha}} f'(T) \geq \Big((\sigma-1)(1-\tfrac{D_0}{D_T})^2 \Big)^{\frac{\sigma\alpha}{\alpha+2-\sigma\alpha}}$$

provided $\sigma < 1+\frac{2}{\alpha}$. Note that

$$f(T)^{\frac{\sigma\alpha}{\alpha+2-\sigma\alpha}} f'(T) = \frac{d}{dt} \left(\frac{f(T)^{\frac{\sigma\alpha}{\alpha+2-\sigma\alpha}+1}}{\frac{\sigma\alpha}{\alpha+2-\sigma\alpha}+1} \right) = \frac{d}{dt} \left(\frac{f(T)^{\frac{\alpha+2}{\alpha+2-\sigma\alpha}}}{\frac{\alpha+2}{\alpha+2-\sigma\alpha}} \right).$$

Then we get by integration in time

$$f(T) \geq (\sigma-1)^{\frac{\sigma\alpha}{\alpha+2}} \left(\tfrac{\alpha+2}{\alpha+2-\sigma\alpha} \right)^{1-\sigma\frac{\alpha}{\alpha+2}} \, (1-\tfrac{D_0}{D_T})^{2\frac{\sigma\alpha}{\alpha+2}} \, T^{1-\sigma\frac{\alpha}{\alpha+2}}. \tag{2.11}$$

When T is such that $\frac{D_0}{D_T} \leq 1-(1-\delta)^{\frac{\alpha+2}{2\sigma\alpha}} =: \varepsilon(\delta)$, (2.11) yields

$$f(T) \geq C(\alpha,\sigma)(1-\delta) \, T^{1-\sigma\frac{\alpha}{\alpha+2}}$$

with

$$C(\alpha,\sigma) := (\sigma-1)^{\frac{\sigma\alpha}{\alpha+2}} \left(\tfrac{\alpha+2}{\alpha+2-\sigma\alpha} \right)^{1-\sigma\frac{\alpha}{\alpha+2}}. \tag{2.12}$$

For the case $\frac{D_0}{D_T} > \varepsilon(\delta)$ the interpolation inequality yields

$$E_T^\sigma > \varepsilon(\delta)^{\sigma\alpha} D_0^{-\sigma\alpha}.$$

Since the energy decreases in time this inequality holds for all $t \leq T$, so that

$$\int_0^T E(t)^\sigma \, dt \geq \varepsilon(\delta)^{\sigma\alpha} D_0^{-\sigma\alpha} T = \varepsilon(\delta)^{\sigma\alpha} D_0^{-\sigma\alpha} T^{\sigma\frac{\alpha}{\alpha+2}} \, T^{1-\sigma\frac{\alpha}{\alpha+2}}.$$

Hence

$$f(T) \;\geq\; (1-\delta) C(\alpha,\sigma) \, T^{1-\sigma\frac{\alpha}{\alpha+2}}$$

provided $T \geq C(\alpha,\sigma)^{\frac{\alpha+2}{\sigma\alpha}} \varepsilon(\delta)^{-\alpha-2} \, D_0^{\alpha+2}$, which proves (2.4). □

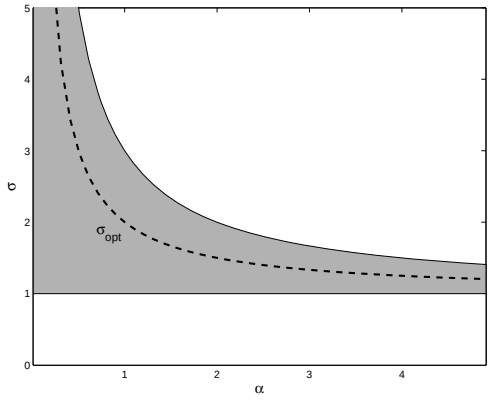


FIG. 8. *Permitted values of σ (grey region) as a function of α. The dotted line $\sigma_{opt} := 1 + \frac{1}{\alpha}$ indicates the values of σ for which the coefficient in* (2.3) *is optimal.*

Remark 4. The coefficient C is optimal for $\sigma = 1 + \frac{1}{\alpha}$; see Figure 8.

To see that the coefficient is optimal for $\sigma = 1 + \frac{1}{\alpha}$, consider in one dimension the energy $E(h) := h^{-\alpha}$. Obviously this energy obeys the interpolation inequality (2.2) for $h^* = 0$. The gradient flow of E with $h_0 = 0$ is given by

$$h(t) = (\alpha(\alpha+2)t)^{\frac{1}{\alpha+2}}.$$

Hence

$$E(h(t)) = h(t)^{-\alpha} = (\alpha(\alpha+2)t)^{-\frac{\alpha}{\alpha+2}}$$

and furthermore

$$\int_0^T E(t)^\sigma \, dt = (\alpha(\alpha+2))^{-\sigma\frac{\alpha}{\alpha+2}} \frac{\alpha+2}{\alpha+2-\sigma\alpha} \, T^{1-\sigma\frac{\alpha}{\alpha+2}}.$$

The coefficient coincides with the coefficient in (2.11) provided $\sigma = 1 + \frac{1}{\alpha}$.

3. An interpolation inequality. From section 2 and Appendix B we learn that relating the energy (1.2) and the induced distance (1.5) by an interpolation inequality of the form (1.12) provides a main ingredient for the proof of a lower bound on the energy. In view of Proposition 1, the geometric exponent α determines the dynamic exponent γ. As mentioned before, we fix $h^* \equiv 3$ to focus on ideas. Proposition 2 shows that the geometric exponent is $\alpha = \frac{n}{n+1}$.

PROPOSITION 2. *There exists a constant $C > 0$ depending only on n such that*

$$\Lambda^{-n} E(h) \, \left(\Lambda^{-\frac{n}{2}} \mathcal{W}(h,3)\right)^{\frac{n}{n+1}} \geq \tfrac{1}{C} \qquad \textit{provided} \quad E(h) \leq \tfrac{1}{C}, \; \Lambda \geq C.$$

Before giving the rigorous proof, let us motivate the result. The exponent $\alpha = \frac{n}{n+1}$ can be heuristically inferred from the following argument:

(i) From (1.12) we have

$$\Lambda^{-n} E \; \sim \; L^{-\frac{n}{n+1}}.$$

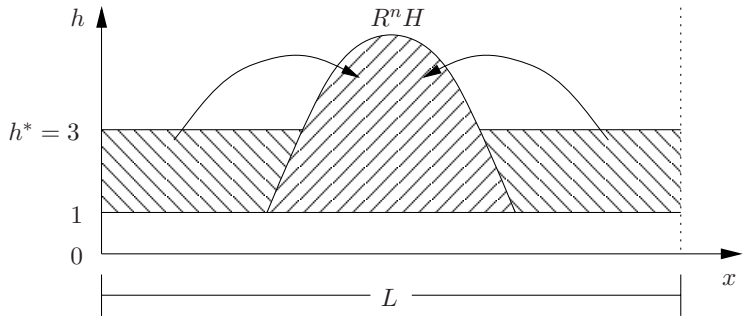


FIG. 9. *Scaling of* $\mathcal{W}(h,3)$.

(ii) From the definition (1.5) of the Wasserstein distance $\mathcal{W}$ we obtain the scaling

$$\begin{aligned}\Lambda^{-n}\mathcal{W}(h,3)^2 &\sim \Lambda^{-n}\iint |x-y|^2\,d\pi(x,y)\\ &\sim \Lambda^{-n}\times L^2\iint d\pi(x,y)\\ &\sim \Lambda^{-n}\times L^2\times\int 3\,dx\\ &\sim L^2,\end{aligned}$$

(see Figure 9); that is,

$$\Lambda^{-\frac{n}{2}}\mathcal{W}(h,3)\ \sim\ L. \tag{3.1}$$

These two scaling relations yield

$$\Lambda^{-n}E\ \left(\Lambda^{-\frac{n}{2}}\mathcal{W}(h,3)\right)^{\frac{n}{n+1}}\ \sim\ L^{-\frac{n}{n+1}}\times L^{\frac{n}{n+1}}\ \sim\ 1.$$

We turn to the rigorous proof of Proposition 2. Set

$$R := (\Lambda^{-n}E(h))^{-1}\qquad\text{and}\qquad \widetilde{h} := (h-2)_+. \tag{3.2}$$

Note that the definition of R is motivated by the scaling (1.11) and (1.12). The proof is done in several lemmas.

(i) Lemma 2 shows that the average droplet height H scales like the average droplet radius in accordance with (1.10). The radius is expressed in terms of the energy; cf. (3.2).

(ii) Lemma 3, applied to $\widetilde{h}$, shows that most of the droplet mass lies in a "small" set in the sense that the volume of the thickened set is controlled.

(iii) Lemmas 2 and 3 imply Lemma 4, which shows that the typical droplet distance L scales like $R^{\frac{n+1}{n}}$, as suggested by the heuristic arguments in (1.11).

(iv) Finally, Lemma 5 reveals that for sufficiently distant droplets the volume-averaged Wasserstein distance between h and the average height scales like L in accordance with (3.1).

LEMMA 2.

(i) *The typical droplet height H is at least of order R in the sense that*

$$\int_{\{\widetilde{h}>H\}}\widetilde{h}\,dx \ge \tfrac{1}{2}\int\widetilde{h}\,dx\qquad \textit{for } H=\tfrac{R}{2}. \tag{3.3}$$

(ii) *The typical droplet radius is at least of the order R in the sense that*

$$R \int |\nabla \widetilde{h}| \, dx \le \int \widetilde{h} \, dx. \tag{3.4}$$

Proof. We first notice that $h \le \widetilde{h} + 2$ implies

$$\int h \, dx \le \int \widetilde{h} \, dx + \tfrac{2}{3} \int 3 \, dx \overset{(1.7)}{=} \int \widetilde{h} \, dx + \tfrac{2}{3} \int h \, dx,$$

so that

$$\int \widetilde{h} \, dx \ge \tfrac{1}{3} \int h \, dx. \tag{3.5}$$

Next we notice that

$$\Lambda^{-n} E(h) \overset{(1.7)}{=} \tfrac{3}{\int h \, dx} \int \tfrac{1}{2} |\nabla h|^2 + \mathcal{U}(h) \, dx \overset{(1.6)}{\ge} \tfrac{3}{\int h \, dx} \operatorname{vol}(\{h > 2\}),$$

so that by (3.2)

$$\operatorname{vol}(\{h > 2\}) \le \tfrac{1}{3R} \int h \, dx. \tag{3.6}$$

This implies

$$\begin{aligned} \int_{\{\widetilde{h} \le H\}} \widetilde{h} \, dx &= \int_{\{2 < h \le H+2\}} (h-2) \, dx \\ &\le H \operatorname{vol}(\{h > 2\}) \\ &\overset{(3.6)}{\le} \tfrac{H}{3R} \int h \, dx \\ &\overset{(3.5)}{\le} \tfrac{H}{R} \int \widetilde{h} \, dx. \end{aligned}$$

Hence we obtain (3.3):

$$\int_{\{\widetilde{h} > H\}} \widetilde{h} \, dx = \int \widetilde{h} \, dx - \int_{\{\widetilde{h} \le H\}} \widetilde{h} \, dx \ge \left(1 - \tfrac{H}{R}\right) \int \widetilde{h} \, dx.$$

This motivates the choice of H.

Now we turn to (3.4):

$$\begin{aligned} \int |\nabla \widetilde{h}| \, dx &= \int_{\{h > 2\}} |\nabla h| \, dx \\ &\overset{(1.6)}{\le} \int |\nabla h| \sqrt{\mathcal{U}(h)} \, dx \\ &\le \int \tfrac{1}{2} |\nabla h|^2 + \mathcal{U}(h) \, dx \\ &\overset{(1.7)}{=} \tfrac{\int h \, dx}{3\Lambda^n} \int \tfrac{1}{2} |\nabla h|^2 + \mathcal{U}(h) \, dx \\ &\overset{(3.5)}{\le} \Lambda^{-n} E(h) \int \widetilde{h} \, dx. \end{aligned}$$

According to (3.2) this turns into

$$R\int|\nabla\widetilde{h}|\,dx\le\int\widetilde{h}\,dx.\qquad\square$$

The next lemma is strongly inspired by [3, Lemma 2.1].

LEMMA 3. *Let $R\le\Lambda$. Assume $g:(0,\Lambda)^n\to[0,\infty)$*

(i) *has height H in the sense that*

$$\int_{\{g\ge H\}}g\,dx\ge\tfrac{1}{2}\int g\,dx \tag{3.7}$$

(ii) *and radius R in the sense that*

$$R\int|\nabla g|\,dx\le\int g\,dx. \tag{3.8}$$

Then there exists a set $A_R\subset\{g\ge H\}$

(i) *which contains substantial mass in the sense that*

$$\int_{A_R}g\,dx\ge\tfrac{1}{4}\int g\,dx$$

(ii) *and is small in the sense that the volume of the thickened sets*

$$A_R^d:=\{x\in(0,\Lambda)^n\mid dist(x,A_R)<d\}$$

is controlled by

$$vol(A_R^d)\le 3^n2^{n+1}\left(1+4\tfrac{d}{R}\right)^n\tfrac{1}{H}\int g\,dx\qquad\text{for all } d>0.$$

Proof. Extend $g:[0,\Lambda]^n\to\mathbb{R}$ to $\overline{g}:\mathbb{R}^n\to\mathbb{R}$ such that

(i) $[0,\Lambda]^n\rightsquigarrow[-\Lambda,\Lambda]^n$ by even reflection and

(ii) $[-\Lambda,\Lambda]^n\rightsquigarrow\mathbb{R}^n$ by periodic continuation.

Set for convenience $\overline{A}:=\{\overline{g}\ge H\}$. Define

$$\overline{A}_R:=\left\{x\in\overline{A}\;\middle|\;\int_{B(x,\frac{R}{8})}\overline{g}\,dy\ge\tfrac{H}{2}\,\mathrm{vol}\left(B\left(x,\frac{R}{8}\right)\right)\right\}. \tag{3.9}$$

With the help of the convolution of $\overline{g}$,

$$\overline{g}_R(x):=\frac{1}{\mathrm{vol}\,(B(x,\frac{R}{8}))}\int_{B(x,\frac{R}{8})}\overline{g}\,dy,$$

$\overline{A}_R$ can be written as $\overline{A}_R=\{x\in\overline{A}\mid\overline{g}_R(x)\ge\frac{H}{2}\}$.

We use the standard estimate

$$\int_{(-\Lambda,\Lambda)^n}|\overline{g}-\overline{g}_R|\,dx\le\tfrac{R}{8}\int_{(-\Lambda,\Lambda)^n}|\nabla\overline{g}|\,dx\overset{(3.8)}{\le}\tfrac{1}{8}\int_{(-\Lambda,\Lambda)^n}\overline{g}\,dx.$$

Since the integrands are even functions this yields

$$\int|g-\overline{g}_R|\,dx\le\tfrac{1}{8}\int g\,dx. \tag{3.10}$$

We now define $A := \overline{A} \cap (0,\Lambda)^n$ and $A_R := \overline{A}_R \cap (0,\Lambda)^n$. Then we have

$$g \geq H \geq 2\overline{g}_R \quad \text{on } A - A_R$$

and thus

$$g \leq 2(g - \overline{g}_R) \quad \text{on } A - A_R.$$

Therefore

$$\int_{A-A_R} g\,dx \leq 2\int_{A-A_R} (g-\overline{g}_R)\,dx \leq 2\int |g-\overline{g}_R|\,dx \overset{(3.10)}{\leq} \tfrac{1}{4}\int g\,dx. \tag{3.11}$$

Notice that by assumption (3.7), $\int_A g\,dx \geq \frac{1}{2}\int g\,dx$. Hence

$$\begin{aligned}\int g\,dx &\leq 2\int_A g\,dx \\ &\leq 2\left(\int_{A_R} g\,dx + \int_{A-A_R} g\,dx\right) \\ &\overset{(3.11)}{\leq} 2\int_{A_R} g\,dx + \tfrac{1}{2}\int g\,dx,\end{aligned}$$

which yields the first assertion.

Let $J \subset A_R$ be maximal with the property

$$\left\{B(x,\tfrac{R}{8})\right\}_{x\in J} \text{ are disjoint.} \tag{3.12}$$

Then necessarily

$$A_R \subset \bigcup_{x\in J} B(x,\tfrac{R}{4}). \tag{3.13}$$

Thus

$$\begin{aligned}\#J\ \mathrm{vol}\,(B(0,\tfrac{R}{8})) &= \sum_{x\in J} \mathrm{vol}\,(B(x,\tfrac{R}{8})) \\ &\overset{(3.9)}{\leq} \tfrac{2}{H}\sum_{x\in J}\int_{B(x,\frac{R}{8})} \overline{g}\,dx \\ &\overset{(3.12)}{\leq} \tfrac{2}{H}\int_{(-\Lambda,2\Lambda)^n} \overline{g}\,dx \\ &= 3^n\tfrac{2}{H}\int g\,dx. \qquad (3.14)\end{aligned}$$

Here we used the assumption $R \leq \Lambda$.

Now (3.13) implies $A_R^d \subset \bigcup_{x\in J} B(x,\frac{R}{4}+d)$, so that

$$\begin{aligned}\mathrm{vol}\,(A_R^d) &\leq \#J\ \mathrm{vol}\,(B(0,\tfrac{R}{4}+d)) \\ &= \frac{\mathrm{vol}\,(B(0,\frac{R}{4}+d))}{\mathrm{vol}\,(B(0,\frac{R}{8}))}\,\#J\ \mathrm{vol}\,(B(0,\tfrac{R}{8})) \\ &\overset{(3.14)}{\leq} 3^n(\tfrac{8}{R}(\tfrac{R}{4}+d))^n\tfrac{2}{H}\int g\,dx \\ &= 3^n 2^{n+1}(1+4\tfrac{d}{R})^n\tfrac{1}{H}\int g\,dx,\end{aligned}$$

which proves the second assertion. ∎

LEMMA 4. *Let $\Lambda \geq R \geq 3^n 2^{8n}$. Then the typical droplet distance L is at least of the order $R^{\frac{n+1}{n}}$ in the following sense: There exists a set $A_R \subset \mathbb{R}^n$ such that*

(i)

$$\int_{A_R} h\,dx \geq \tfrac{1}{12} \int h\,dx \tag{3.15}$$

and

(ii)

$$3\ vol(A_R^L) \leq \tfrac{3}{4} \int_{A_R} h\,dx \quad \text{for } L = 3^{-1}2^{-10}R^{\frac{n+1}{n}}. \tag{3.16}$$

Proof. According to Lemma 3 there exists a set A_R such that

$$\int_{A_R} \widetilde{h}\,dx \geq \tfrac{1}{4} \int \widetilde{h}\,dx, \tag{3.17}$$

which by (3.5) turns into (3.15), and

$$\mathrm{vol}\,(A_R^L) \leq 3^n 2^{n+1} \left(1 + 4\tfrac{L}{R}\right)^n \tfrac{1}{H} \int \widetilde{h}\,dx. \tag{3.18}$$

By the definition of H and R in Lemma 2, (3.18) gives rise to

$$\begin{aligned} \mathrm{vol}\,(A_R^L) &\leq 3^n 2^{n+2} \left(1 + 4\tfrac{L}{R}\right)^n \tfrac{1}{R} \int \widetilde{h}\,dx \\ &\overset{(3.17)}{\leq} 3^n 2^{n+4} \left(1 + 4\tfrac{L}{R}\right)^n \tfrac{1}{R} \int_{A_R} \widetilde{h}\,dx \\ &\leq 3^n 2^{n+4} \left(1 + 4\tfrac{L}{R}\right)^n \tfrac{1}{R} \int_{A_R} h\,dx. \end{aligned}$$

Now L in (3.16) is defined such that

$$3^n 2^{n+4} \left(1 + 4\tfrac{L}{R}\right)^n \tfrac{1}{R} \leq \tfrac{1}{4}$$

provided $R \geq 3^n 2^{8n}$. Hence the inequality turns into

$$3\ \mathrm{vol}\,(A_R^L) \leq \tfrac{3}{4} \int_{A_R} h\,dx. \qquad \square$$

LEMMA 5. *Let $h : Q \to [0, \infty)$ with $h^* := \Lambda^{-n} \int h(x)\,dx$ and let$A \subset \mathbb{R}^n$ and $L > 0$ be given with*

$$h^*\ vol(\{dist(\cdot, A) < L\}) \leq \tfrac{3}{4} \int_A h(x)\,dx. \tag{3.19}$$

Then

$$\mathcal{W}(h, h^*)^2 \geq \tfrac{1}{4} L^2 \int_A h(x)\,dx.$$

Proof. Set for abbreviation $A^L := \{\mathrm{dist}(.,A) < L\}$. Let π be any admissible transportation plan in the definition of $\mathcal{W}$. We conclude that

$$\begin{aligned}\int_{\mathbb{R}^n\times\mathbb{R}^n} |x-y|^2\,d\pi(x,y) &\geq \int_{A\times(\mathbb{R}^n-A^L)} |x-y|^2\,d\pi(x,y)\\ &\geq L^2\pi(A\times(\mathbb{R}^n-A^L))\\ &\geq L^2(\pi(A\times\mathbb{R}^n)-\pi(\mathbb{R}^n\times A^L))\\ &= L^2\left(\int_A h\,dx-\int_{A^L} h^*\,dx\right)\\ &= L^2\left(\int_A h\,dx-h^*\,\mathrm{vol}\,(A^L)\right)\\ &\overset{(3.19)}{\geq} \tfrac{1}{4}L^2\int_A h\,dx. \qquad \square\end{aligned}$$

Proof of Proposition 2. According to Lemma 5 (applied to $h^*=3$ and $A=A_R$) it follows from Lemma 4 for the L defined in (3.16) that

$$\mathcal{W}(h,3)^2 \geq \tfrac{1}{4}L^2\int_{A_R} h\,dx \overset{(3.15)}{\geq} \tfrac{1}{48}L^2\int h\,dx.$$

In view of (1.7), this turns into

$$\Lambda^{-n}\mathcal{W}(h,3)^2 \geq 2^{-4}L^2.$$

In view of the definition (3.2) of R and the definition (3.16) of L this yields

$$\begin{aligned}\Lambda^{-n}E(h)\ \left(\Lambda^{-\frac{n}{2}}\mathcal{W}(h,3)\right)^{\frac{n}{n+1}} &\geq R^{-1}(2^{-4}L^2)^{\frac{n}{2(n+1)}}\\ &= R^{-1}(3^{-2}2^{-24}R^{2\frac{n+1}{n}})^{\frac{n}{2(n+1)}}\\ &= 3^{-\frac{n}{n+1}}2^{-\frac{12n}{n+1}}. \qquad \square\end{aligned}$$

4. Proof of Theorem 1. We cannot apply Proposition 1 right away since the argument for the (infinite-dimensional) gradient flow structure introduced in Appendix B is formal. An inspection of the proof of Proposition 1 reveals that it is only necessary to find a substitute for Lemma 1. In fact, one can directly prove the equivalent of Lemma 1 for the Wasserstein metric as defined in (1.5) and a smooth solution of (1.4).

LEMMA 6. *Let h be a smooth solution of* (1.4). *Then*

$$\left|\frac{d}{dt}\mathcal{W}(3,h(t))\right|^2 \leq \left(-\frac{d}{dt}E(t)\right). \tag{4.1}$$

Proof. We follow [22].

Note that

$$\frac{d}{dt}E(t) = -\int h\left|\nabla\frac{\partial E}{\partial h}\right|^2 dx.$$

It thus is sufficient to establish the inequality

$$\left|\frac{d}{dt}\mathcal{W}(3,h(t))\right|^2 \leq \int h|u|^2\,dx \tag{4.2}$$

for the transport equation

$$\partial_t h + \nabla \cdot (hu) \;=\; 0. \tag{4.3}$$

Due to the triangle inequality we only need to show

$$\lim_{\delta \to 0} \frac{1}{\delta} \mathcal{W}(h_t, h_{t+\delta}) \;\leq\; \sqrt{\int h_t |u_t|^2 \, dx}. \tag{4.4}$$

Here, the indices t and $t+\delta$ denote the time argument of h and u.

First we show that $h_{t+\delta}$ is the push-forward of h_t under the flow map Φ_δ generated by $u_{t+\delta}$, i.e.,

$$\partial_\delta \Phi_\delta = u_{t+\delta} \circ \Phi_\delta, \qquad \Phi_0 = \mathrm{id}. \tag{4.5}$$

Note that by the push-forward one understands that

$$\int \zeta \; h_{t+\delta} \, dx \;=\; \int (\zeta \circ \Phi_\delta) \; h_t \, dx \qquad \text{for all } \zeta \in C_0^0(\mathbb{R}^n). \tag{4.6}$$

For given ζ define $\zeta_\delta := \zeta \circ \Phi_\delta^{-1}$; ζ_δ satisfies

$$\partial_\delta \zeta_\delta + u_{t+\delta} \cdot \nabla \zeta_\delta \;=\; 0.$$

Furthermore, recall that $h_{t+\delta}$ solves the transport equation

$$\partial_\delta h_{t+\delta} + \nabla \cdot (h_{t+\delta} u_{t+\delta}) \;=\; 0.$$

Hence we obtain

$$\frac{d}{d\delta} \int \zeta_\delta \, h_{t+\delta} \, d\delta \;=\; \int (\partial_\delta \zeta_\delta \, h_{t+\delta} + \partial_\delta h_{t+\delta} \, \zeta_\delta) \, dx \;=\; 0,$$

which proves (4.6).

Next we define a product measure π_δ by

$$d\pi_\delta(x,y) \;=\; dh_t(x) \, \delta[y = \Phi_\delta(x)].$$

According to (4.6), π_δ defines an admissible transportation plan in the sense of the definition of $\mathcal{W}$. Thus we get by definition

$$\frac{1}{\delta} \mathcal{W}(h_t, h_{t+\delta}) \leq \sqrt{\int \frac{1}{\delta^2} |x - \Phi_\delta(x)|^2 h_t(x) \, dx}.$$

We obtain from the definition (4.5) of the flow map that $\frac{1}{\delta^2}|x - \Phi_\delta(x)|^2$ converges pointwise to $|u_t|^2$. The dominated convergence theorem yields

$$\lim_{\delta \to 0} \int \frac{1}{\delta^2} |x - \Phi_\delta(x)|^2 h_t(x) \, dx \;=\; \int h_t |u_t|^2 \, dx,$$

which establishes (4.4). ∎

Appendix A. Stationary droplet shape.

In this section, we analyze the shape of a stationary droplet. For this purpose we consider a single droplet with prescribed mass $V > 0$ on top of the precursor layer of equilibrium thickness $h \equiv 1$ on an n-dimensional substrate. We are interested in the stationary droplet shape on a mesoscopic scale. Hence we focus on the mesoscopic energy

$$\overline{E} = \int \tfrac{1}{2}|\nabla h|^2 \, dx + \text{ vol}\,(\{h > 1\}) \tag{A.1}$$

(see section 1.4) for all h, which fulfill the mass constraint

$$\int (h-1)_+ \, dx = V. \tag{A.2}$$

Since the precursor layer of height $h \equiv 1$ has no contribution to the energy $\overline{E}$, we shift h by -1 and consider the problem

$$\text{minimize } \int \tfrac{1}{2}|\nabla h|^2 \, dx + \text{ vol}\,(\{h > 0\}) \text{ subject to } \int h \, dx = V. \tag{A.3}$$

For convenience, we keep the notation h for the shifted film height.

PROPOSITION 3. *Let $V \geq 0$. Then there exists a constant $H > 0$ depending on V such that*

$$\bar{h}(x) \;=\; -\tfrac{1}{2H}|x|^2 + H \qquad \text{on } B(0, \sqrt{2}H) \tag{A.4}$$

is the unique (up to translations) minimizer of problem (A.3).

Proof. We proceed in three steps.

(i) Any minimizer $\bar{h}$ of (A.3) is radially symmetric and monotonically decreasing.

(ii) Any minimizer $\bar{h}$ satisfies

$$-\Delta \bar{h} \;=\; \text{const} \qquad \text{in } \{\bar{h} > 0\}. \tag{A.5}$$

(iii) A unique (up to translations) minimizer $\bar{h}$ exists and satisfies

$$\tfrac{1}{2}|\nabla \bar{h}|^2 \;=\; 1 \qquad \text{on } \partial\{\bar{h} > 0\}. \tag{A.6}$$

Argument for (i)*:* The proof is based on the symmetric decreasing rearrangement $h^{\#}$ of the function h. It is well-known [13, Lemma 7.17] that

$$\int |\nabla h^{\#}|^2 \, dx \;\leq\; \int |\nabla h|^2 \, dx$$

with equality if and only if h is radially symmetric and monotone decreasing. Furthermore, the second contribution to the energy is conserved:

$$\text{vol}\,(\{h^{\#} > 0\}) \;=\; \text{vol}\,(\{h > 0\}).$$

Argument for (ii)*:* The first variation of $\overline{E}$ yields

$$\int (-\Delta h)\, \delta h \, dx \;=\; 0$$

for all variation δh with

$$\int \delta h \, dx = 0 \quad \text{and} \quad \operatorname{supp} \delta h \subset \{h > 0\}.$$

Hence we obtain (A.5).

Argument for (iii)*:* From (1) and (2), we deduce that any minimizer centered at the origin must be of the form

$$h(x) = -A|x|^2 + H \qquad \text{on } B(0, \sqrt{H/A}). \tag{A.7}$$

Note that the family of candidates is invariant under the volume-conserving homothetic variation:

$$h_\lambda(x) \; = \; \lambda^{-n} h\left(\frac{x}{\lambda}\right).$$

Then the first variation of $\overline{E}$ in λ at $\lambda = 1$ yields that at a critical point $\bar{h}$

$$(n+2) \int \tfrac{1}{2} |\nabla \bar{h}|^2 \, dx \; = \; n \; \operatorname{vol}(\{\bar{h} > 0\}). \tag{A.8}$$

Note that the critical point is in fact a minimum since the function

$$\lambda \; \mapsto \; E(h_\lambda) = \lambda^{-(n+2)} \int \tfrac{1}{2} |\nabla \bar{h}|^2 \, dx + \lambda^n \operatorname{vol}(\{\bar{h} > 0\})$$

is convex. We compute

$$\int_{\{\bar{h}>0\}} \tfrac{1}{2} |\nabla \bar{h}|^2 \, dx = \tfrac{2}{n+2} \omega_n A^2 (\sqrt{H/A})^{n+2}$$

and

$$\operatorname{vol}(\{\bar{h} > 0\}) \; = \; \tfrac{1}{n} \omega_n (\sqrt{H/A})^n,$$

where ω_n denotes the $(n-1)$-dimensional measure of $\mathbb{S}^{n-1}$. Thus we obtain from (A.8) that for the minimizer $\bar{h}$ A is determined by

$$2AH = 1. \tag{A.9}$$

Hence at $\partial\{\bar{h} > 0\}$, where $|x|^2 = H/A$,

$$|\nabla \bar{h}|^2 \; = \; 2,$$

which proves (3). Furthermore, we obtain from (A.7) and (A.9) that

$$\bar{h}(x) \; = \; -\tfrac{1}{2H}|x|^2 + H \qquad \text{on } B(0, \sqrt{2}H). \qquad \square$$

Remark 5. Proposition 3 reveals the scaling (1.10)

$$H \; \sim \; R,$$

since the radius of the droplet is given by $\sqrt{2}H$.

Appendix B. The gradient flow structure.

In this section, we specify in which sense the evolution (1.4) is a gradient flow of E defined in (1.2). This heuristic section serves purely as a motivation which guides our analysis. The rigorous result is independent of this section.

The mathematical structure required for a gradient flow,

$$\partial_t h \;=\; -\mathrm{grad} E(h), \tag{B.1}$$

is determined by a smooth function $\mathcal{M} \ni h \mapsto E(h)$ on a Riemannian manifold $(\mathcal{M}, g)$. A trajectory $[0,\infty) \ni t \mapsto h(t) \in \mathcal{M}$ of (B.1) is characterized by the fact that for any tangent vector field $[0,\infty) \ni t \mapsto \delta h(t) \in T_{h(t)}\mathcal{M}$, one has

$$g_{h(t)}\left(\partial_t h(t), \delta h(t)\right) + \left\langle \mathrm{diffE}_{h(t)}, \delta h(t) \right\rangle \;=\; 0 \quad \text{for all } t \geqslant 0. \tag{B.2}$$

In our case $\mathcal{M}$ corresponds to the space of all possible film heights which take the overall mass constraint into account:

$$\mathcal{M} \;= \left\{ h \geq 0 \;\middle|\; \int h(x)\, dx = \int 3\, dx \right\}.$$

The metric tensor encodes the limiting dissipation mechanism by (viscous) friction. Given the continuity equation $\partial_t h + \nabla \cdot (h\, u) = 0$ for the film height $h \geq 0$ by a (vertically averaged horizontal) velocity field $u \in \mathbb{R}^n$, the rate of energy dissipation by friction is given by $\int h\, |u|^2\, dx$ in the case of Darcy-type friction. (It would be $\int \frac{1}{h} |u|^2\, dx$ for Stokes friction with no-slip boundary conditions.) The (quadratic part of) the metric tensor is given by

$$g_h(\delta h, \delta h) \;=\; \inf_u \left\{ \int h|u|^2\, dx \;\middle|\; \delta h + \nabla \cdot (hu) = 0 \right\}. \tag{B.3}$$

For the sake of simplicity, we do not state the boundary conditions such as $\nu \cdot u = 0$ on ∂Q. The squared size, $g_h(\delta h, \delta h)$, of an infinitesimal perturbation δh is the minimal rate of energy dissipation by friction which is necessary to generate δh.

Writing down the Euler–Lagrange equation for (B.3) yields the following representation in terms of the velocity potential φ:

$$g_h(\delta h, \delta h) = \int h|\nabla\varphi|^2\, dx, \qquad \text{where} \quad \delta h + \nabla \cdot (h \nabla \varphi) = 0.$$

By polarization, this yields

$$g_h(\delta h_1, \delta h_2) \;=\; \int h\, \nabla\varphi_1 \cdot \nabla\varphi_2\, dx, \tag{B.4}$$

where the functions φ_i are defined by

$$\delta h_i + \nabla \cdot (h \nabla \varphi_i) = 0.$$

It is easy to check that indeed (1.4) is the gradient flow of (1.2) in the sense of (B.2) with respect to the metric tensor (B.4) defined on $\mathcal{M}$.

Any Riemannian manifold $(\mathcal{M}, g)$ is endowed with a natural distance function d between two points h_0 and h_1 by means of minimizing the action of curves from h_0 to h_1. In view of the definition of the metric tensor (B.3), d turns into

$$d(h_0, h_1)^2 \;=\; \inf_{(h,u)} \left\{ \int_0^1 \int h|u|^2\, dx ds \;\middle|\; \partial_s h + \nabla \cdot (hu) = 0, \left\{ \begin{array}{l} h(0,\cdot) = h_0 \\ h(1,\cdot) = h_1 \end{array} \right\} \right\}.$$

It is shown in [23] that d, in fact, coincides with the Wasserstein distance $\mathcal{W}$ defined in (1.5).

Acknowledgments. The authors thank Patrick Penzler and Maria Reznikoff for fruitful discussions. Dejan Slepčev would also like to thank the University of Bonn for hospitality and Sonderforschungsbereich 611 for supporting his visits.

REFERENCES

[1] J. Becker, G. Grün, R. Seeman, H. Mantz, K. Jacobs, K. R. Mecke, and R. Blossey, *Complex dewetting scenarios captured by thin-film models*, Nature Materials, 2 (2003), pp. 59–63.

[2] A. L. Bertozzi, G. Grün, and T. P. Witelski, *Dewetting films: bifurcations and concentrations*, Nonlinearity, 14 (2001), pp. 1569–1592.

[3] S. Conti, B. Niethammer, and F. Otto, *Coarsening rates in off-critical mixtures*, SIAM J. Math. Anal., 37 (2006), pp. 1732–1741.

[4] S. Dai and R. L. Pego, *Universal bounds on coarsening rates for mean-field models of phase transition*, SIAM J. Math. Anal., 37 (2005), pp. 347–371.

[5] S. Dai and R. L. Pego, *An upper bound on the coarsening rate for mushy zones in a phase-field model*, Interfaces Free Bound., 7 (2005), pp. 187–197.

[6] K. B. Glasner and T. P. Witelski, *Coarsening dynamics of dewetting films*, Phys. Rev. E., 67 (2003), p. 016302.

[7] K. B. Glasner and T. P. Witelski, *Collision vs. collapse of droplets in coarsening of dewetting thin films*, Phys. D., 209 (2005), pp. 80–104.

[8] S. Herminghaus, K. Jacobs, K. Mecke, J. Bischof, A. Fery, M. Ibn-Elhaj, and S. Schlagowski, *Spinodal dewetting in liquid cristal and liquid metal films*, Science, 282 (1998), pp. 916–919.

[9] J. N. Israelachvili, *Intermolecular and Surface Forces*, Academic Press, New York, 2nd ed., 1992.

[10] R. V. Kohn and F. Otto, *Upper bounds on coarsening rates*, Comm. Math. Phys., 229 (2002), pp. 375–395.

[11] R. V. Kohn and X. Yan, *Upper bound on the coarsening rate for an epitaxial growth model*, Comm. Pure Appl. Math., 56 (2003), pp. 1549–1564.

[12] R. V. Kohn and X. Yan, *Coarsening rates for models of multicomponent phase separation*, Interfaces Free Bound., 6 (2004), pp. 135–149.

[13] E. H. Lieb and M. Loss, *Analysis*, Graduate Studies in Mathematics, 14, 2nd ed., AMS, Providence, RI, 2001.

[14] R. Limary and P. F. Green, *Late-stage coarsening of an unstable structured liquid film*, Phys. Rev. E, 66 (2002), p. 021601.

[15] R. Limary and P. F. Green, *Dynamics of droplets on the surface of a structured fluid film: Late-stage coarsening*, Langmuir, 19 (2003), pp. 2419–2424.

[16] A. Münch, *Dewetting rates of thin liquid films*, J. Phys. Condens. Matter, 17 (2005), pp. S309–S318.

[17] A. Münch and B. Wagner, *Contact-line instability of dewetting thin films*, Phys. D., 209 (2005), pp. 178–190.

[18] B. Niethammer and F. Otto, *Domain coarsening in thin films*, Comm. Pure Appl. Math., 54 (2001), pp. 361–384.

[19] A. Oron and S. G. Bankoff, *Dewetting of a heated surface by an evaporating liquid film under conjoining/disjoining pressures*, J. Colloid Interface Sci., 218 (1999), pp. 152–166.

[20] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Mod. Phys., 69 (1997), pp. 932–977.

[21] F. Otto, P. Penzler, and T. Rump, *Discretisation and numerical tests of a diffuse–interface model with Ehrlich–Schwoebel barrier*, in Multiscale Modeling in Epitaxial Growth, International Series of Numerical Mathematics, 149, Birkhäuser, Basel, Switzerland, 2005, pp. 127–158.

[22] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.

[23] F. Otto and M. Westdickenberg, *Eulerian calculus for the contraction in the Wasserstein distance*, SIAM J. Math. Anal., 37 (2005), pp. 1227–1255.

[24] L. M. Pismen, *Spinodal dewetting in volatile liquid film*, Phys. Rev. E, 70 (2004), pp. 021601-1–021601-9.

[25] L. M. Pismen and Y. Pomeau, *Mobility and interactions of weakly nonwetting droplets*, Phys. Fluids, 16 (2004), pp. 2604–2612.

[26] G. Reiter, *Dewetting of thin polymer films*, Phys. Rev. Lett., 68 (1992), pp. 75–78.

[27] G. Reiter and R. Khanna, *Kinetics of autophobic dewetting of polymer films*, Langmuir, 16 (2000), pp. 6351–6357.
[28] R. Seeman, S. Herminghaus, and K. Jacobs, *Gaining control of pattern formation of dewetting liquid films*, J. Phys. Condens. Matter, 13 (2001), pp. 4925–4938.
[29] A. Sharma and R. Khanna, *Pattern formation in unstable thin liquid films*, Phys. Rev. Lett., 81 (1998), pp. 3463–3466.
[30] U. Thiele, M. G. Velarde, and K. Neuffer, *Dewetting: film rupture by nucleation in the spinodal regime*, Phys. Rev. Lett., 87 (2001), p. 4.
[31] F. Vandenbrouck, M. P. Valignat, and A. M. Cazabat, *Thin nematic films: metastability and spinodal dewetting*, Phys. Rev. Lett., 82 (1999), pp. 2693–2696.
[32] C. Villani, *Topics in Optimal Transportation*, Graduate Studies in Mathematics, 58, AMS, Providence, RI, 2003.
[33] R. Xie, A. Karim, J. F. Douglas, C. C. Han, and R. A. Weiss, *Spinodal dewetting of thin polymer films*, Phys. Rev. Lett., 81 (1998), pp. 1251–1254.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 530–556

WAVELETS WITH SHORT SUPPORT*

BIN HAN† AND ZUOWEI SHEN‡

Abstract. This paper is to construct Riesz wavelets with short support. Riesz wavelets with short support are the objective of interest in both theory and application. In theory, it is known that a B-spline of order m has the shortest support among all compactly supported refinable functions with the same regularity. However, it remained open whether a Riesz wavelet with the shortest support and m vanishing moments can be constructed from the multiresolution analysis generated by the B-spline of order m. In various applications, a Riesz wavelet with a short support, a high order of regularity, and vanishing moments is often desirable in signal and image processing, since they have a good time frequency localization and approximation property, as well as fast algorithms. This paper presents a theory for the construction of Riesz wavelets with short support and gives various examples. In particular, from the multiresolution analysis whose underlying refinable function is the B-spline of order m, we are able to construct the shortest supported Riesz wavelet with m vanishing moments. The support of the wavelet functions can be made even shorter by reducing their orders of vanishing moments. The study here also provides a new insight into the structures of the spline tight frame systems constructed in [A. Ron and Z. Shen, *J. Funct. Anal.*, 148 (1997), pp. 408–447, I. Daubechies, B. Han, A. Ron, and Z. Shen, *Appl. Comput. Harmon. Anal.*, 14 (2003), pp. 1–46, B. Han and Q. Mo, *Proc. Amer. Math. Soc.*, 132 (2004), pp. 77–86] and bi-frame systems in [I. Daubechies, B. Han, A. Ron, and Z. Shen, *Appl. Comput. Harmon. Anal.*, 14 (2003), pp. 1–46, I. Daubechies and B. Han, *Constr. Approx.*, 20 (2004), pp. 325–352].

Key words. Riesz wavelet bases, B-spline functions, wavelet frames

AMS subject classifications. 42C40, 41A15, 41A05

DOI. 10.1137/S0036141003438374

1. Introduction.

The objective of this paper is to design Riesz wavelets with short support from a multiresolution analysis. We start with some basic notions and definitions. A function ϕ is *refinable* if it satisfies the refinement equation

$$\phi = 2\sum_{k\in\mathbb{Z}} a(k)\phi(2\cdot -k), \tag{1.1}$$

where $a : \mathbb{Z} \mapsto \mathbb{C}$ is a sequence on $\mathbb{Z}$, called the *refinement mask* for ϕ.

By $\ell_2(\mathbb{Z})$ we denote the set of all sequences $u : \mathbb{Z} \mapsto \mathbb{C}$ such that

$$\|u\|_{\ell_2(\mathbb{Z})} := \left(\sum_{k\in\mathbb{Z}} |u(k)|^2\right)^{1/2} < \infty.$$

The *Fourier series* $\hat{u}$ of a sequence u in $\ell_2(\mathbb{Z})$ is defined as

$$\hat{u}(\xi) := \sum_{k\in\mathbb{Z}} u(k)e^{-ik\xi}, \qquad \xi \in \mathbb{R}, \tag{1.2}$$

*Received by the editors December 8, 2003; accepted for publication (in revised form) January 18, 2006; published electronically June 9, 2006.
http://www.siam.org/journals/sima/38-2/43837.html

†Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2G1 (bhan@math.ualberta.ca, http://www.ualberta.ca/∼bhan). The research of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC Canada) under grant G121210654.

‡Department of Mathematics, National University of Singapore, Singapore (matzuows@math.nus.edu.sg, http://www.math.nus.edu.sg/∼matzuows). The research of this author was supported in part by several grants at the National University of Singapore.

where i is the imaginary unit such that $i^2 = -1$. Similarly, the Fourier transform of a function $f \in L_1(\mathbb{R})$ is defined as

$$\hat{f}(\xi) := \int_{\mathbb{R}} f(t) e^{-i\xi t}\, dt, \qquad \xi \in \mathbb{R},$$

which can be extended naturally to functions in $L_2(\mathbb{R})$ and tempered distributions. Now, refinement equation (1.1) can be rewritten, in terms of the Fourier transform, as

$$\hat{\phi}(\xi) = \hat{a}(\xi/2)\hat{\phi}(\xi/2), \qquad \text{a.e. } \xi \in \mathbb{R}.$$

Using the above definition, one extends the concept of refinable functions to that of refinable distributions. Throughout the paper we assume that $\hat{\phi}(0) \neq 0$ and $\hat{a}(0) = 1$. We also call $\hat{a}$ a refinement mask for convenience.

For a given compactly supported refinable function $\phi \in L_2(\mathbb{R})$, define V_0 to be the smallest closed subspace of $L_2(\mathbb{R})$ generated by $\phi(\cdot - k)$, $k \in \mathbb{Z}$. Then V_0 is a shift (integer translate) invariant subspace of $L_2(\mathbb{R})$. Let $V_j := \{f(2^j \cdot) : f \in V_0\}$, for $j \in \mathbb{Z}$. Then the sequence of subspaces V_j, $j \in \mathbb{Z}$, forms a multiresolution analysis (MRA) in $L_2(\mathbb{R})$, which is generated by ϕ, i.e., (i) $V_j \subset V_{j+1}$, $j \in \mathbb{Z}$; (ii) $\overline{\cup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R})$ and $\cap_{j\in\mathbb{Z}} V_j = \{0\}$ (see, e.g., [1] and [15]). In this paper, a function $\psi \in V_1$ (or more precisely, $\psi \in V_1 \backslash V_0$) is called an MRA-based wavelet function, or simply a wavelet, derived from the MRA $\{V_j\}_{j\in\mathbb{Z}}$.

In this paper, we are interested in finding a function ψ such that the wavelet system

$$X(\psi) := \{\psi_{j,k} := 2^{j/2}\psi(2^j \cdot -k) \ : \ j, k \in \mathbb{Z}\}$$

forms a Riesz basis for $L_2(\mathbb{R})$. The set $X(\psi)$ is called the wavelet system generated by ψ. Recall that the system $X(\psi)$ is a *Riesz basis* of $L_2(\mathbb{R})$ if the linear span of $X(\psi)$ is dense in $L_2(\mathbb{R})$ and $X(\psi)$ is a Riesz sequence; that is, there exist two positive constants C_1 and C_2 such that

$$C_1\|\{c_{j,k}\}\|_{\ell_2(\mathbb{Z}^2)} \leqslant \left\|\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} c_{j,k}\psi_{j,k}\right\|_{L_2(\mathbb{R})} \leqslant C_2\|\{c_{j,k}\}\|_{\ell_2(\mathbb{Z}^2)} \qquad \forall\ \{c_{j,k}\}_{j,k\in\mathbb{Z}} \in \ell_2(\mathbb{Z}^2). \tag{1.3}$$

If $X(\psi)$ is a Riesz basis of $L_2(\mathbb{R})$, then ψ is called a *Riesz wavelet*. To construct a compactly supported MRA-based Riesz wavelet ψ, one starts with a compactly supported refinable function ϕ with stable shifts. Recall that the shifts of a function ϕ are *stable* if $\{\phi(\cdot - k) \ : \ k \in \mathbb{Z}\}$ is a Riesz sequence; that is, there exist two positive constants C_1 and C_2 such that $C_1\|\{c_k\}\|_{\ell_2(\mathbb{Z})} \leqslant \|\sum_{k\in\mathbb{Z}} c_k\phi(\cdot - k)\|_{L_2(\mathbb{R})} \leqslant C_2\|\{c_k\}\|_{\ell_2(\mathbb{Z})}$ for all $\{c_k\}_{k\in\mathbb{Z}} \in \ell_2(\mathbb{Z})$. Then a compactly supported Riesz wavelet ψ is obtained by selecting some desirable finitely supported sequence b that is called a *wavelet mask* or a high-pass filter in the language of engineering. With the wavelet mask b, the wavelet function ψ is obtained from b and the refinable function ϕ via

$$\psi = 2\sum_{k\in\mathbb{Z}} b(k)\phi(2\cdot -k), \quad \text{or equivalently,} \quad \hat{\psi}(\xi) = \hat{b}(\xi/2)\hat{\phi}(\xi/2). \tag{1.4}$$

When $\{\phi(\cdot-k) : k\in\mathbb{Z}\}$ forms an orthonormal system, a natural choice of b is

$$b(k)=(-1)^{k-1}\overline{a(1-k)}, \qquad k\in\mathbb{Z}.$$

Its Fourier series can be written as

$$\hat{b}(\xi)=e^{-i\xi}\overline{\hat{a}(\xi+\pi)}. \tag{1.5}$$

Then it is well known that the wavelet system $X(\psi)$ forms an orthonormal basis of $L_2(\mathbb{R})$. The wavelet function ψ has the same length of support as that of the corresponding refinable function ϕ. Thus, once a compactly supported refinable function ϕ whose shifts form an orthonormal system is given, the corresponding orthonormal wavelet ψ can be obtained easily. Hence, the major task and difficulty in constructing compactly supported dyadic orthonormal wavelets in dimension one are to design refinement masks whose corresponding refinable functions have the required smoothness order and whose shifts form an orthonormal system. This was discussed in detail in [6].

On the other hand, compactly supported refinable functions whose shifts form a Riesz sequence are much easier to get. One such class of refinable functions are B-splines.

The B-spline function of order m ($m\in\mathbb{N}$), denoted by B_m, can be obtained via the following recursive formula: $B_1=\chi_{[0,1]}$, the characteristic function of the interval $[0,1]$, and

$$B_m(x):=\int_0^1 B_{m-1}(x-t)\,dt, \qquad x\in\mathbb{R},\ m\in\mathbb{N}. \tag{1.6}$$

The B-spline function $B_m\in C^{m-2}(\mathbb{R})$ is a function of piecewise polynomials of degree less than m, vanishes outside the interval $[0,m]$ and is symmetric about the point $m/2$, that is, $B_m(m-x)=B_m(x)$ for all $x\in\mathbb{R}$. It is well known that the B-spline function B_m is a refinable function satisfying the refinement equation

$$\widehat{B_m}(2\xi)=\Big(\frac{1+e^{-i\xi}}{2}\Big)^m\widehat{B_m}(\xi), \qquad \xi\in\mathbb{R}. \tag{1.7}$$

When $m=1$, the shifts of B_1 form an orthonormal basis of the shift invariant space V_0 generated by B_1. The shifts of B_m, $m>1$, form a Riesz, but not an orthonormal, basis of the shift invariant space V_0 generated by B_m. When m is even, $B_m(\cdot-m/2)$ is symmetric about the origin with the refinement mask

$$\hat{a}(\xi)=\cos^{2m}(\xi/2).$$

While compactly supported refinable functions with stable shifts are easier to obtain, the construction of compactly supported Riesz wavelets from an MRA generated by a B-spline of order m is not straightforward. A compactly supported Riesz wavelet ψ from the MRA generated by B_m was first constructed in [3]. While $X(\psi)$ forms a Riesz basis of $L_2(\mathbb{R})$ and the system keeps the orthogonality between different dilation levels, the support of the prewavelet ψ in [3] is $[0,2m-1]$ (therefore, almost two times that of B_m) and ψ has m vanishing moments. Recently, [16] derived a Riesz wavelet ψ from B_m such that the Riesz wavelet system $X(\psi)$ forms a Riesz basis of $L_2(\mathbb{R})$. When ψ is required to have m vanishing moments, the construction of [16] gives the prewavelet of [3] and hence its support is $[0,2m-1]$; efforts are made in [16] to shorten

the support of the Riesz wavelets at the cost of a reduced order of vanishing moments of the Riesz wavelets. For example, the support of the Riesz wavelet ψ can be reduced to $[0, m]$ when m is odd (or $[0, m+1]$ when m is even) with 1 or 2 vanishing moments. It is a bit surprising to us that there are no discussions in the literature whether the natural choice of a function,

$$\psi := 2\sum_{k\in\mathbb{Z}} b(k)B_m(2\cdot -k) \tag{1.8}$$

with

$$b(k) = (-1)^{k-1}\overline{a(1-k)}, \qquad k \in \mathbb{Z}, \tag{1.9}$$

where $\hat{a}(\xi) = 2^{-m}(1+e^{-i\xi})^m$ is the refinement mask of B_m, is a Riesz wavelet. There are several other motivations that lead to the discussions here. First, the mask defined in (1.9) works for the case $m = 1$. It is clear that when $m = 1$, the corresponding wavelet function ψ is the Haar wavelet. Hence, $X(\psi)$ is an orthonormal basis of $L_2(\mathbb{R})$. In fact, it works for an arbitrary compactly supported refinable function whose shifts form an orthonormal system. It is natural to ask whether the function ψ in (1.8) with the wavelet mask defined in (1.9) is a Riesz wavelet. Second, ψ in (1.8) has the same length of support as that of B_m. Further, as will be shown in section 2, in some sense, ψ is the shortest supported Riesz wavelet of regularity $m - 1/2$ with m vanishing moments. Recall that a function ψ has *m vanishing moments* if

$$\hat{\psi}^{(j)}(0) = 0 \qquad \forall\, j = 0, \dots, m-1,$$

where $\hat{\psi}^{(j)}$ denotes the jth derivative of $\hat{\psi}$. This means that it has good time frequency localization and can lead to efficient algorithms in applications. One of the main objectives of this paper is to prove that the system $X(\psi)$, where ψ is defined in (1.8), forms a Riesz basis of $L_2(\mathbb{R})$, which will be established in section 2. The spline-based Riesz wavelets in this paper with short support and high vanishing moments may be of interest in wavelet-based numerical algorithms, since the wavelet functions are piecewise polynomials, while it is well known that most other wavelets do not have explicit analytic forms.

In many applications, it is not only important to have Riesz wavelets with short support, but also desirable to have short supported Riesz wavelets with a small condition number, namely, a small ratio of the upper and lower Riesz bounds in (1.3). The condition number of the spline Riesz wavelet suggested here cannot be smaller than that of the system $\{B_m(\cdot - k) : k \in \mathbb{Z}\}$. However, it is well known that the condition number of the system $\{B_m(\cdot - k) : k \in \mathbb{Z}\}$ increases as m goes to ∞. In this regard, it is of interest to construct Riesz wavelets with short support, which are as close as possible to some orthonormal wavelets or tight frame wavelets, for a given order of regularity or vanishing moments. Such wavelet systems will then have small condition numbers. This is of interest and importance in various applications, although it is not a topic addressed in this paper.

The general theory needed for this paper is given in section 3. The theory also provides a new insight into the systematic constructions of the tight frame systems from the B-spline of order m by using the unitary extension principle of [20] and the oblique extension principle of [9]. In those constructions, for a given B_m, a set $\Psi := \{\psi^1, \dots, \psi^L\}$ of functions is obtained so that the system

$$X(\Psi) := \{\psi^\ell_{j,k} := 2^{j/2}\psi^\ell(2^j\cdot -k) \ : \ \ell = 1, \dots, L \ \text{ and } \ j, k \in \mathbb{Z}\}$$

forms a tight frame for $L_2(\mathbb{R})$. That is,

$$f = \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \langle f, \psi_{j,k}^{\ell}\rangle \psi_{j,k}^{\ell} \qquad \forall\, f \in L_2(\mathbb{R}).$$

The system $X(\Psi)$ is a redundant system. However, we discover from the study here that in all spline constructions, there exists a function $\psi \in \Psi$ such that $X(\psi)$ (in some cases $X(\psi(\cdot - 1/2))$ forms a Riesz basis for $L_2(\mathbb{R})$. This finding roughly says that one of the functions in the set Ψ already can generate a Riesz basis for $L_2(\mathbb{R})$, while the other functions in Ψ are there just to either improve the condition number determined by the upper and lower frame bounds or provide a better dual system.

More generally, a system $X(\psi)$ is *Bessel* in $L_2(\mathbb{R})$ if there exists a positive constant C such that

$$\sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} |\langle f, \psi_{j,k}\rangle|^2 \leqslant C\|f\|^2 \qquad \forall\, f \in L_2(\mathbb{R}). \tag{1.10}$$

A system $X(\psi)$ is Bessel if both functions $\sum_{k\in\mathbb{Z}} |\hat{\psi}(\cdot + 2\pi k)|$ and $\sum_{j\in\mathbb{Z}} |\hat{\psi}(2^j\cdot)|$ are in $L_\infty(\mathbb{R})$ (see [22, Corollary 15]). This will hold whenever ψ has a sufficient smoothness order. For example, it is known [11, Propositions 2.6 and 3.5] that if, for some $\varepsilon > 0$, there exists a positive constant C such that $|\hat{\psi}(\xi)| \leqslant C(1+|\xi|)^{-1/2-\varepsilon}$ and $|\hat{\psi}(\xi)| \leqslant C|\xi|^{\varepsilon}$ for all $\xi \in \mathbb{R}$, then the corresponding system $X(\psi)$ is Bessel in $L_2(\mathbb{R})$. It is clear that for a finite set Ψ, $X(\Psi)$ is Bessel if and only if $X(\psi)$ is Bessel for each $\psi \in \Psi$.

A system $X(\Psi)$ with $\Psi = \{\psi^1, \dots, \psi^L\}$ is a *frame* for $L_2(\mathbb{R})$ if there exist positive constants C_1 and C_2 such that

$$C_1\|f\|^2 \leqslant \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} |\langle f, \psi_{j,k}^{\ell}\rangle|^2 \leqslant C_2\|f\|^2 \qquad \forall\, f \in L_2(\mathbb{R}). \tag{1.11}$$

Let $\Psi := \{\psi^1, \dots, \psi^L\}$ and $\tilde{\Psi} := \{\tilde{\psi}^1, \dots, \tilde{\psi}^L\}$ be two sets of functions in $L_2(\mathbb{R})$. We say that $(X(\Psi), X(\tilde{\Psi}))$ is a pair of *bi-frames* in $L_2(\mathbb{R})$ if each of $X(\Psi)$ and $X(\tilde{\Psi})$ is Bessel in $L_2(\mathbb{R})$, and if $X(\Psi)$ and $X(\tilde{\Psi})$ satisfy

$$\langle f, g\rangle = \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \langle f, \tilde{\psi}_{j,k}^{\ell}\rangle \langle \psi_{j,k}^{\ell}, g\rangle \qquad \forall\, f, g \in L_2(\mathbb{R}), \tag{1.12}$$

where $\langle f, g\rangle := \int_{\mathbb{R}} f(t)\overline{g(t)}\, dt$. If (1.12) holds with $\tilde{\psi}^{\ell} = \psi^{\ell}$ for all $\ell = 1, \dots, L$, then $X(\Psi)$ is a tight wavelet frame in $L_2(\mathbb{R})$.

A similar phenomenon also can be found for pairs of bi-frames constructed from B-splines in [9, 8]. All of these phenomena will be discussed in detail in section 4. More examples of Riesz wavelets with short support will be given in the last section. Some of our examples are even from nonspline refinable functions.

2. Riesz spline wavelet bases with short support. In order to show that the system $X(\psi)$ with ψ defined in (1.8) forms a Riesz basis for $L_2(\mathbb{R})$, we need the following lemma that is a special case of Corollary 3.3.

LEMMA 2.1. *Let a be a finitely supported refinement mask for a compactly supported refinable function $\phi \in L_2(\mathbb{R})$ with $\hat{a}(0) = 1$ and $\hat{a}(\pi) = 0$ such that $\hat{\phi}(0) \neq 0$ and $\hat{a}$ can be factorized into the form*

$$\hat{a}(\xi) = \Big(\frac{1 + e^{-i\xi}}{2}\Big)^m \hat{A}(\xi),$$

where $\hat{A}$ is the Fourier series of a finitely supported sequence A with $\hat{A}(\pi) \neq 0$. Suppose that

$$|\hat{a}(\xi)|^2 + |\hat{a}(\xi + \pi)|^2 \neq 0 \quad \forall \quad \xi \in \mathbb{R}.$$

Define

$$\hat{\psi}(2\xi) := e^{-i\xi}\overline{\hat{a}(\xi + \pi)}\hat{\phi}(\xi)$$

and

$$\hat{\tilde{A}}(\xi) := \frac{\hat{A}(\xi)}{|\hat{a}(\xi)|^2 + |\hat{a}(\xi + \pi)|^2}.$$

Assume that

$$\rho_A := \inf_{n\in\mathbb{N}} \|\widehat{A_n}\|^{1/n}_{L_\infty(\mathbb{R})} < 2^{m-1/2} \qquad \text{and} \qquad \rho_{\tilde{A}} := \inf_{n\in\mathbb{N}} \|\widehat{\tilde{A}_n}\|^{1/n}_{L_\infty(\mathbb{R})} < 2^{m-1/2},$$

where $\widehat{A_n}(\xi) := \hat{A}(2^{n-1}\xi)\cdots\hat{A}(2\xi)\hat{A}(\xi)$ and $\widehat{\tilde{A}_n}(\xi) := \hat{\tilde{A}}(2^{n-1}\xi)\cdots\hat{\tilde{A}}(2\xi)\hat{\tilde{A}}(\xi)$. Then $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$.

Recall that a function f is in the Sobolev space $W^\beta(\mathbb{R})$ if

$$\int |\hat{f}(\xi)|^2(1 + |\xi|^2)^\beta d\xi < \infty.$$

We say that f has the *regularity* α if $f \in W^\beta(\mathbb{R})$ for all $\beta < \alpha$. It is well known that the B-spline B_m of order m has the regularity $m - 1/2$. A compactly supported function ϕ satisfies the Strang–Fix condition of order m if

$$\hat{\phi}(0) \neq 0 \quad \text{and} \qquad \hat{\phi}^{(j)}(2\pi k) = 0 \qquad \forall\, j = 0, 1, \ldots, m-1, \quad k \in \mathbb{Z}\backslash\{0\}.$$

Applying Lemma 2.1 to the B-spline functions, we obtain the following result.

THEOREM 2.2. *Let B_m be the B-spline function of order m with the refinement mask*

$$\hat{a}(\xi) = 2^{-m}(1 + e^{-i\xi})^m.$$

Define

$$\hat{\psi}(2\xi) = 2^{-m}e^{-i\xi}(1 - e^{i\xi})^m\widehat{B_m}(\xi). \tag{2.1}$$

Then the following hold:

(i) *The function ψ has the regularity $m - 1/2$ and has m vanishing moments. It is either symmetric or antisymmetric satisfying $\psi = (-1)^m\psi(1 - \cdot)$ and it is supported on the interval $[1/2 - m/2, 1/2 + m/2]$.*

(ii) *$X(\psi)$ forms a Riesz basis for $L_2(\mathbb{R})$.*

(iii) *Among all wavelets (redundant or nonredundant) which have m vanishing moments and are based on an MRA whose underlying refinable function has the regularity $m-1/2$, the Riesz wavelet ψ has the shortest support.*

Proof. Conclusion (i) follows directly from the properties of B-splines and the definition of ψ.

For (ii), we apply Lemma 2.1. First, it is easy to check that $|\hat{a}(\xi)|^2+|\hat{a}(\xi+\pi)|^2 \neq 0$ for all $\xi \in \mathbb{R}$. Since $\hat{A}$ in Lemma 2.1 is 1, clearly, $\rho_A = 1 < 2^{m-1/2}$ for all $m \in \mathbb{N}$. The corresponding $\hat{\tilde{A}}$ in Lemma 2.1 is

$$\hat{\tilde{A}}(\xi) = \frac{1}{|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2} = \frac{1}{\cos^{2m}(\xi/2) + \sin^{2m}(\xi/2)}.$$

To prove that $\rho_{\tilde{A}} < 2^{m-1/2}$, we observe that the function $f_m(x) := x^m + (1-x)^m$ decreases on $[0,1/2]$ and increases on $[1/2,1]$ by $f_m'(x) = m[x^{m-1} - (1-x)^{m-1}]$. Consequently, we have $f_m(x) \geqslant f(1/2) = 2^{1-m}$ for all $x \in [0,1]$ and $m \in \mathbb{N}$. Since $\hat{\tilde{A}}(\xi) = [f(\cos^2(\xi/2))]^{-1}$, we conclude that $\rho_{\tilde{A}} \leqslant \|\hat{\tilde{A}}\|_{L_\infty(\mathbb{R})} \leqslant [f(1/2)]^{-1} = 2^{m-1} < 2^{m-1/2}$. By Lemma 2.1, $X(\psi)$ is a Riesz wavelet basis for $L_2(\mathbb{R})$.

For (iii), since the corresponding refinable function ϕ has the regularity $m-1/2$, ϕ must satisfy the Strang–Fix condition of order m (see [19, 14]). Hence, ϕ must be the convolution of B_m with some function/distribution (see [18, Theorem 3.7]). Hence, B_m has the shortest support among all refinable functions of the regularity $m-1/2$.

For any given MRA-based wavelet with m vanishing moments, since the refinable function ϕ satisfies $\hat{\phi}(0) \neq 0$, the wavelet mask must have the factor

$$\Big(\frac{1-e^{i\xi}}{2}\Big)^m.$$

This says that in order to have m vanishing moments, the wavelet mask cannot be shorter than $(\frac{1-e^{i\xi}}{2})^m$. Altogether, we conclude that ψ defined in (2.1) has the shortest support among all wavelets (redundant or nonredundant) which have m vanishing moments and are based on an MRA whose underlying refinable function has the regularity $m-1/2$. □

Since it is rare to be able to derive a wavelet of regularity $m-1/2$ from a multiresolution whose underlying refinable function has the regularity smaller than $m-1/2$, (iii) in the above theorem essentially says that ψ defined in (2.1) is the shortest supported MRA-based wavelet having the regularity $m-1/2$ and m vanishing moments.

Example 2.3. Let $m=2$. Then by (2.1), $\psi = \frac{1}{2}B_2(2\cdot-1) - B_2(2\cdot) + \frac{1}{2}B_2(2\cdot+1)$. By Theorem 2.2, $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$. The Riesz wavelet ψ has 2 vanishing moments and the regularity 3/2. See Figure 1 for the graphs of the functions B_2 and ψ.

Example 2.4. Let $m=3$. Then by (2.1), $\psi = \frac{1}{4}B_3(2\cdot-1) - \frac{3}{4}B_3(2\cdot) + \frac{3}{4}B_3(2\cdot+1) - \frac{1}{4}B_3(2\cdot+2)$. By Theorem 2.2, $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$. The Riesz wavelet ψ has 3 vanishing moments and the regularity 5/2. See Figure 2 for the graphs of the functions B_3 and ψ.

3. Biorthogonal wavelets with infinite masks. In this section, we give a general form of Lemma 2.1. This not only leads to a proof of Lemma 2.1, but also leads to a result in a more general setting. This further allows us to connect the

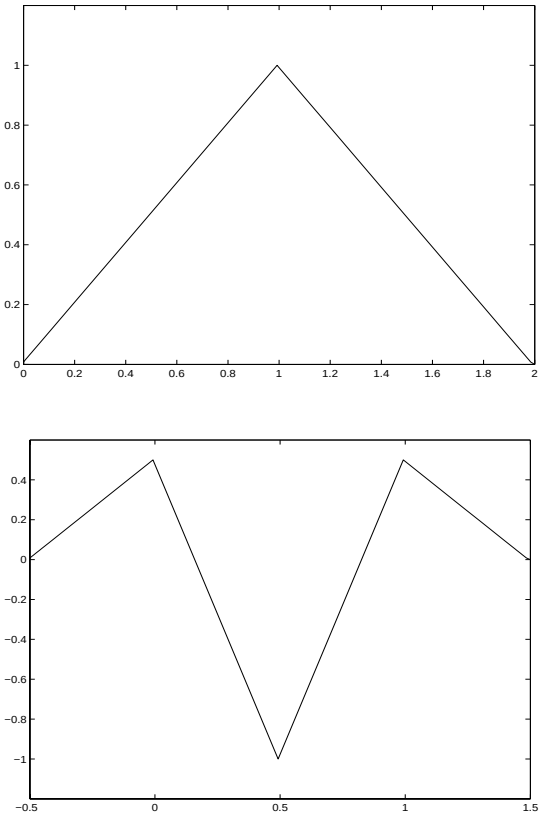


FIG. 1. *The graph of the B-spline B_2 (top) and the graph of the wavelet function ψ (bottom) in Example* 2.3. *The Riesz wavelet ψ has* 2 *vanishing moments and the regularity* 3/2. *The wavelet system $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$.*

discussions here to the spline tight frame wavelet systems given in [20, 9, 12] and bi-frame systems in [9, 8], as we will discuss in section 4.

We start with some basic notions. Recall that a function f on $\mathbb{R}$ has *polynomial decay* if

$$(1 + |\cdot|)^j f \in L_\infty(\mathbb{R}) \qquad \forall\, j \in \mathbb{N}$$

and has *exponential decay* if there exists a positive number β such that

$$e^{\beta|\cdot|} f \in L_\infty(\mathbb{R}).$$

If a function f has polynomial decay or exponential decay, then clearly $f \in L_p(\mathbb{R})$ for all $1 \leqslant p \leqslant \infty$.

Similarly, a sequence a on $\mathbb{Z}$ has polynomial decay if

$$\sum_{k\in\mathbb{Z}} (1 + |k|)^j |a(k)| < \infty \qquad \forall\, j \in \mathbb{N},$$

or equivalently, $\sup_{k\in\mathbb{Z}}(1 + |k|)^j |a(k)| < \infty$ for all $j \in \mathbb{N}$. It is easy to see that a sequence a has polynomial decay if and only if $\hat{a} \in C^\infty(\mathbb{R})$.

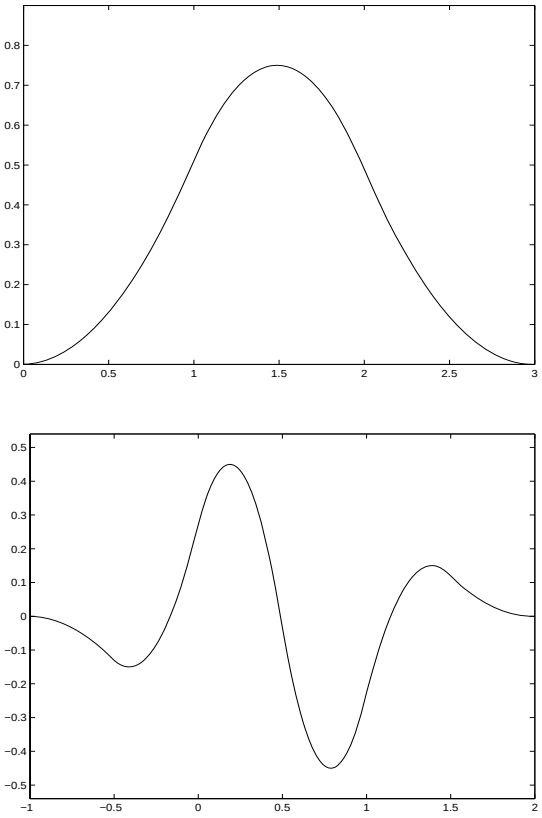


FIG. 2. *The graph of the B-spline B_3 (top) and the graph of the function ψ (bottom) in Example* 2.4. *The function ψ has* 3 *vanishing moments and the regularity* 5/2. *The wavelet system $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$.*

For a sequence u on $\mathbb{Z}$ and a function f on $\mathbb{R}$, we define

$$\nabla u := u - u(\cdot - 1) \quad \text{and} \quad \nabla f := f - f(\cdot - 1). \tag{3.1}$$

In general, $\nabla^m u = \sum_{k=0}^m (-1)^k \frac{m!}{k!(m-k)!} u(\cdot - k)$.

The following lemma is similar to [10, Theorem 3.6] and will be needed later.

LEMMA 3.1. *Let $f \in L_2(\mathbb{R})$ be a function with polynomial decay and let m be an arbitrary given positive integer. Then the following statements are equivalent:*

(i) $\hat{f}^{(j)}(2\pi k) = 0$ *for all* $k \in \mathbb{Z}$ *and* $j = 0, \ldots, m-1$.

(ii) *The identity* $f = \nabla^m h$ *holds where*

$$h := \sum_{k=0}^{\infty} \frac{(k+m-1)!}{k!(m-1)!} f(\cdot - k) \tag{3.2}$$

has polynomial decay. In particular, h is in $L_2(\mathbb{R})$.

Proof. Assume that (i) holds. It is easy to see that (i) is equivalent to $\sum_{k\in\mathbb{Z}} k^j f(\cdot - k) = 0$ for all $j = 0, \ldots, m-1$. Let h be the function as given in (3.2). Since f has

polynomial decay, it is easy to see that h is well defined on $\mathbb{R}$. Note that

$$\begin{aligned}\nabla h &= h - h(\cdot - 1)\\ &= \sum_{k=0}^{\infty} \frac{(k+m-1)!}{k!(m-1)!} f(\cdot - k) - \sum_{k=1}^{\infty} \frac{(k+m-2)!}{(k-1)!(m-1)!} f(\cdot - k)\\ &= f + \sum_{k=1}^{\infty} \left[\frac{(k+m-1)!}{k!(m-1)!} - \frac{(k+m-2)!}{(k-1)!(m-1)!} \right] f(\cdot - k)\\ &= f + \sum_{k=1}^{\infty} \frac{(k+m-2)!}{k!(m-2)!} f(\cdot - k)\\ &= \sum_{k=0}^{\infty} \frac{(k+m-2)!}{k!(m-2)!} f(\cdot - k).\end{aligned}$$

Hence, $\nabla^m h = f$ by induction on m. Next, we show that h has polynomial decay. Since f has polynomial decay, there exist positive constants $C_\ell, \ell \in \mathbb{N}$, such that

$$|f(t)| \leqslant C_\ell (1 + |t|)^{-\ell} \qquad \forall\, t \in \mathbb{R}.$$

Let $t \leqslant 0$ and $j \in \mathbb{N}$. Then

$$\begin{aligned}(1+|t|)^j |h(t)| &\leqslant \sum_{k=0}^{\infty} \left| \frac{(k+m-1)\cdots(k+1)}{(m-1)!} \right| |f(t-k)|(1+|t|)^j\\ &\leqslant C_\ell \sum_{k=0}^{\infty} \frac{(k+m-1)^m}{(m-1)!} (1+|t-k|)^{-\ell} (1+|t|)^j\\ &\leqslant C_\ell \sum_{k=0}^{\infty} \frac{m^m (k+1)^m}{(m-1)!} (1+|t|+k)^{-\ell} (1+|t|)^j\\ &\leqslant C_\ell \frac{m^m}{(m-1)!} \sum_{k=0}^{\infty} (1+|t|+k)^{m+j-\ell}\\ &\leqslant C_\ell \frac{m^m}{(m-1)!} \sum_{k=0}^{\infty} (1+k)^{m+j-\ell},\end{aligned}$$

which is finite whenever $\ell > m + j + 1$. (Note here that one always can choose such an ℓ by the definition of the polynomial decay of f.) For $t > 0$, we first note that

$$\frac{(k+m-1)!}{k!} = (k+m-1)\cdots(k+1)$$

(when $m = 1$, by convention it takes value 1) is a polynomial of degree $m-1$ for the variable k. Then (i) asserts that

$$0 = \sum_{k=-\infty}^{-1} \frac{(k+m-1)\cdots(k+1)}{(m-1)!} f(\cdot - k) + \sum_{k=0}^{\infty} \frac{(k+m-1)\cdots(k+1)}{(m-1)!} f(\cdot - k).$$

Hence,

$$h = -\sum_{k=-\infty}^{-1} \frac{(k+m-1)\cdots(k+1)}{(m-1)!} f(\cdot - k).$$

By a similar argument applying to the above identity, we conclude that h must have polynomial decay for $t > 0$.

Assume that (ii) holds. This implies that $\hat{f}(\xi) = (1-e^{-i\xi})^m \hat{h}(\xi)$, which gives (i). This completes the proof. □

By δ we denote the *Dirac sequence* on $\mathbb{Z}$ such that $\delta(0) = 1$ and $\delta(k) = 0$ for all $k \in \mathbb{Z}\backslash\{0\}$. The *bracket product* of two functions f and g in $L_2(\mathbb{R})$ is defined to be (see [13])

$$[f,g](\xi) := \sum_{k\in\mathbb{Z}} f(\xi+2\pi k)\overline{g(\xi+2\pi k)}, \qquad \xi \in \mathbb{R}. \tag{3.3}$$

It is well known that $\langle f(\cdot-k), g\rangle = \delta(k)$ for all $k \in \mathbb{Z}$ if and only if $[\hat{f}, \hat{g}] = 1$.

Assume that $(X(\Psi), X(\tilde{\Psi}))$ is a pair of bi-frames. If the system $(X(\Psi), X(\tilde{\Psi}))$ further satisfies $\langle \psi^\ell_{j,k}, \psi^{\ell'}_{j',k'}\rangle = \delta(\ell-\ell')\delta(j-j')\delta(k-k')$ for all $\ell, \ell' = 1, \ldots, L$ and $j, j', k, k' \in \mathbb{Z}$, then $(X(\Psi), X(\tilde{\Psi}))$ forms a pair of *biorthogonal wavelet bases* in $L_2(\mathbb{R})$. Clearly, if $(X(\Psi), X(\tilde{\Psi}))$ forms a pair of biorthogonal wavelet bases in $L_2(\mathbb{R})$, then both systems $X(\Psi)$ and $X(\tilde{\Psi})$ form a Riesz wavelet basis in $L_2(\mathbb{R})$ (see, e.g., [21]).

With Lemma 3.1, we prove the following result on biorthogonal wavelets with infinite masks.

THEOREM 3.2. *Let a and b be two sequences on $\mathbb{Z}$ satisfying the following two conditions:*

(i) *There are positive integers m and $\tilde{m}$ such that $\hat{a}(\xi) = (\frac{1+e^{-i\xi}}{2})^m \hat{A}(\xi)$ and $\hat{b}(\xi) = (\frac{1-e^{i\xi}}{2})^{\tilde{m}} \hat{B}(\xi)$, where A and B are sequences on $\mathbb{Z}$ with polynomial decay satisfying $\hat{A}(0) = 1$ and $\hat{B}(\pi) \neq 0$.*

(ii) *The function $\hat{d}(\xi) := \hat{a}(\xi)\hat{b}(\xi+\pi) - \hat{a}(\xi+\pi)\hat{b}(\xi)$ does not vanish for all $\xi \in \mathbb{R}$.*

Let

$$\begin{aligned} &\hat{\tilde{a}}(\xi) := \left(\frac{1+e^{-i\xi}}{2}\right)^{\tilde{m}} \hat{\tilde{A}}(\xi), \quad \textit{where} \quad \hat{\tilde{A}}(\xi) := \frac{\overline{\hat{B}(\xi+\pi)}}{\overline{\hat{d}(\xi)}}, \quad \textit{and} \\ &\hat{\tilde{b}}(\xi) := -\frac{\overline{\hat{a}(\xi+\pi)}}{\overline{\hat{d}(\xi)}}. \end{aligned} \tag{3.4}$$

Define

$$\begin{aligned} &\hat{\phi}(\xi) := \prod_{j=1}^{\infty} \hat{a}(2^{-j}\xi) \quad \textit{and} \quad \hat{\tilde{\phi}}(\xi) := \prod_{j=1}^{\infty} \hat{\tilde{a}}(2^{-j}\xi), \\ &\hat{\psi}(\xi) := \hat{b}(\xi/2)\hat{\phi}(\xi/2) \quad \textit{and} \quad \hat{\tilde{\psi}}(\xi) := \hat{\tilde{b}}(\xi/2)\hat{\tilde{\phi}}(\xi/2). \end{aligned}$$

Assume that

$$\limsup_{n\to\infty} \|A_n\|^{1/n}_{\ell_2(\mathbb{Z})} < 2^{m-1/2} \quad \textit{and} \quad \limsup_{n\to\infty} \|\tilde{A}_n\|^{1/n}_{\ell_2(\mathbb{Z})} < 2^{\tilde{m}-1/2}, \tag{3.5}$$

where

$$\widehat{A_n}(\xi) := \hat{A}(2^{n-1}\xi)\cdots\hat{A}(2\xi)\hat{A}(\xi) \quad \textit{and} \quad \widehat{\tilde{A}_n}(\xi) := \hat{\tilde{A}}(2^{n-1}\xi)\cdots\hat{\tilde{A}}(2\xi)\hat{\tilde{A}}(\xi). \tag{3.6}$$

Then all the functions $\phi, \tilde{\phi}, \psi, \tilde{\psi}$ belong to $L_2(\mathbb{R})$ satisfying

$$\begin{aligned} &\langle \phi, \tilde{\phi}(\cdot-k)\rangle = \langle \psi, \tilde{\psi}(\cdot-k)\rangle = \delta(k), \\ &\langle \phi, \tilde{\psi}(\cdot-k)\rangle = \langle \psi, \tilde{\phi}(\cdot-k)\rangle = 0 \qquad \forall\, k \in \mathbb{Z}. \end{aligned} \tag{3.7}$$

If we further assume that

(iii) $[\hat{\phi}, \hat{\phi}] \in L_\infty(\mathbb{R})$ *and* $[\hat{\tilde{\phi}}, \hat{\tilde{\phi}}] \in L_\infty(\mathbb{R})$,

(iv) $X(\psi)$ *and* $X(\tilde{\psi})$ *are Bessel in* $L_2(\mathbb{R})$,

then $(X(\psi), X(\tilde{\psi}))$ *forms a pair of biorthogonal wavelet bases in* $L_2(\mathbb{R})$. *In particular,* $X(\psi)$ *is a Riesz basis of* $L_2(\mathbb{R})$.

Proof. The essential part of the proof is to show that the corresponding cascade algorithms to obtain ϕ and $\tilde{\phi}$ as defined below converge in $L_2(\mathbb{R})$. We shall use some ideas here from the proof of [10, Theorem 4.3], which deals with convergence of vector cascade algorithms in Sobolev spaces. We use the compactly supported orthonormal refinable function η that has a support $[0, 2\max(m, \tilde{m}) + 1]$ (see [7]) satisfying

$$\hat{\eta}(0) = 1,\ \hat{\eta}^{(j)}(2\pi k) = 0 \quad \forall\, k \in \mathbb{Z}\backslash\{0\} \quad \text{and} \quad j = 0, \dots, \max(m, \tilde{m}) \tag{3.8}$$

to obtain the initial seed in the cascade algorithm. Since η and its shifts form an orthonormal system, $[\hat{\eta}, \hat{\eta}] = 1$, it can be easily verified that $(\hat{\eta}\overline{\hat{\eta}})^{(j)}(0) = \delta(j)$ for all $j = 0, \dots, 2\max(m, \tilde{m})$. Since $\hat{\tilde{a}}(0) = \hat{\eta}(0) = 1$, by [10, Lemmas 2.2 and 3.4], there exists a finitely supported sequence c on $\mathbb{Z}$ such that

$$\hat{c}(0) = 1, \qquad 2^{-j}(\hat{\tilde{a}}\hat{c}\hat{\eta})^{(j)}(0) = (\hat{c}\hat{\eta})^{(j)}(0) \qquad \forall\, j = 0, \dots, \max(m, \tilde{m}) \tag{3.9}$$

and $\hat{c}(\xi) \neq 0$ for all $\xi \in \mathbb{R}$. In fact, as shown in [10, Lemmas 2.2 and 3.4], the values $\hat{c}^{(j)}(0), j = 1, \dots, \max(m, \tilde{m})$ are uniquely determined by the system of linear equations given in (3.9).

Now, pick the initial seeds ϕ_0 and $\tilde{\phi}_0$ by $\widehat{\phi_0}(\xi) := \hat{\eta}(\xi)/\overline{\hat{c}(\xi)}$ and $\widehat{\tilde{\phi}_0}(\xi) := \hat{c}(\xi)\hat{\eta}(\xi)$. Since the sequence c is finitely supported and $\hat{c}(\xi) \neq 0$ for all $\xi \in \mathbb{R}$, we see that $\phi_0, \tilde{\phi}_0 \in L_2(\mathbb{R})$ have exponential decay. Moreover, it is easy to check that $[\widehat{\phi_0}, \widehat{\tilde{\phi}_0}] = [\hat{\eta}/\bar{\hat{c}}, \hat{\eta}\hat{c}] = [\hat{\eta}, \hat{\eta}] = 1$. The corresponding cascade operators Q_a and $Q_{\tilde{a}}$ defined by a and $\tilde{a}$ are

$$Q_a f := 2\sum_{k\in\mathbb{Z}} a(k) f(2\cdot - k) \quad \text{and} \quad Q_{\tilde{a}} f := 2\sum_{k\in\mathbb{Z}} \tilde{a}(k) f(2\cdot - k), \qquad f \in L_2(\mathbb{R}). \tag{3.10}$$

Let $g := Q_a\phi_0 - \phi_0$ and $\tilde{g} = Q_{\tilde{a}}\tilde{\phi}_0 - \tilde{\phi}_0$. Then

$$\hat{g}(\xi) = \hat{a}(\xi/2)\widehat{\phi_0}(\xi/2) - \widehat{\phi_0}(\xi) = \hat{a}(\xi/2)\hat{\eta}(\xi/2)/\overline{\hat{c}(\xi/2)} - \hat{\eta}(\xi)/\overline{\hat{c}(\xi)} \tag{3.11}$$

and

$$\hat{\tilde{g}}(\xi) = \hat{\tilde{a}}(\xi/2)\widehat{\tilde{\phi}_0}(\xi/2) - \widehat{\tilde{\phi}_0}(\xi) = \hat{\tilde{a}}(\xi/2)\hat{c}(\xi/2)\hat{\eta}(\xi/2) - \hat{c}(\xi)\hat{\eta}(\xi). \tag{3.12}$$

Since a and $\tilde{a}$ have polynomial decay, by the fact that the function η is compactly supported, it follows from (3.11) and (3.12) that both g and $\tilde{g}$ have polynomial decay.

Next, we prove that

$$\hat{g}^{(j)}(2\pi k) = 0 \quad \forall\, j = 0, \dots, m-1 \quad \text{and} \quad k \in \mathbb{Z} \tag{3.13}$$

and

$$\hat{\tilde{g}}^{(j)}(2\pi k) = 0 \quad \forall\, j = 0, \dots, \tilde{m}-1 \quad \text{and} \quad k \in \mathbb{Z}. \tag{3.14}$$

First, when $k \in 2\mathbb{Z}\backslash\{0\}$, applying (3.8), we obtain $\hat{g}^{(j)}(2\pi k) = \hat{\tilde{g}}^{(j)}(2\pi k) = 0$ for all $j = 0, \ldots, \max(m, \tilde{m})$. Second, when $k \in 2\mathbb{Z}+1$, since $\hat{a}(\xi) = 2^{-m}(1+e^{-i\xi})^m \hat{A}(\xi)$ and $\hat{\tilde{a}}(\xi) = 2^{-\tilde{m}}(1+e^{-i\xi})^{\tilde{m}} \hat{\tilde{A}}(\xi)$, it is easy to see that $\hat{g}^{(j)}(2\pi k) = 0$ for all $j = 0, \ldots, m-1$ and $\hat{\tilde{g}}^{(j)}(2\pi k) = 0$ for all $j = 0, \ldots, \tilde{m} - 1$.

Hence, in order to prove (3.13) and (3.14), it suffices to prove the case $k = 0$. For (3.14), applying (3.9), one obtains that

$$\begin{aligned} \hat{\tilde{g}}^{(j)}(0) &= [\hat{\tilde{a}}(\cdot/2)\hat{c}(\cdot/2)\hat{\eta}(\cdot/2) - \hat{c}\hat{\eta}]^{(j)}(0) \\ &= 2^{-j}[\hat{\tilde{a}}\hat{c}\hat{\eta}]^{(j)}(0) - [\hat{c}\hat{\eta}]^{(j)}(0) = 0 \qquad \forall\, j = 0, \ldots, \max(m, \tilde{m}). \end{aligned} \tag{3.15}$$

This gives (3.14).

Next, we prove (3.13). It follows from the definition of $\tilde{a}$ that

$$\overline{\hat{a}(\xi)}\hat{\tilde{a}}(\xi) + \overline{\hat{a}(\xi+\pi)}\hat{\tilde{a}}(\xi+\pi) = 1.$$

This, together with

$$\hat{a}(\xi) = 2^{-m}(1+e^{-i\xi})^m \hat{A}(\xi), \quad \text{and} \quad \hat{\tilde{a}}(\xi) = 2^{-\tilde{m}}(1+e^{-i\xi})^{\tilde{m}} \hat{\tilde{A}}(\xi),$$

leads to

$$\overline{\hat{a}(\xi)}\hat{\tilde{a}}(\xi) = 1 + O(|\xi|^{m+\tilde{m}-1}) \quad \text{as} \quad \xi \to 0$$

and

$$\hat{a}(\xi) = (\overline{\hat{\tilde{a}}(\xi)})^{-1} + O(|\xi|^{m+\tilde{m}-1}).$$

Hence, as $\xi \to 0$,

$$\begin{aligned} \hat{g}(\xi) &= \hat{a}(\xi/2)\hat{\eta}(\xi/2)/\overline{\hat{c}(\xi/2)} - \hat{\eta}(\xi)/\overline{\hat{c}(\xi)} \\ &= \hat{\eta}(\xi/2)\overline{\hat{\eta}(\xi/2)}\,(\overline{\hat{\tilde{a}}(\xi/2)\hat{\eta}(\xi/2)\hat{c}(\xi/2)})^{-1} - \hat{\eta}(\xi)\overline{\hat{\eta}(\xi)}\,(\overline{\hat{c}(\xi)\hat{\eta}(\xi)})^{-1} + O(|\xi|^{m+\tilde{m}-1}) \\ &= (\overline{\hat{\tilde{a}}(\xi/2)\hat{\eta}(\xi/2)\hat{c}(\xi/2)})^{-1} - (\overline{\hat{c}(\xi)\hat{\eta}(\xi)})^{-1} + O(|\xi|^{\max(m,\tilde{m})}) \\ &= (\overline{\hat{\tilde{a}}(\xi/2)\hat{\eta}(\xi/2)\hat{c}(\xi/2)})^{-1}(\overline{\hat{c}(\xi)\hat{\eta}(\xi)})^{-1}\overline{\Big[\hat{c}(\xi)\hat{\eta}(\xi) - \hat{\tilde{a}}(\xi/2)\hat{c}(\xi/2)\hat{\eta}(\xi/2)\Big]} \\ &\qquad + O(|\xi|^{\max(m,\tilde{m})}) \\ &= (\overline{\hat{\tilde{a}}(\xi/2)\hat{\eta}(\xi/2)\hat{c}(\xi/2)})^{-1}(\overline{\hat{c}(\xi)\hat{\eta}(\xi)})^{-1}\overline{\hat{\tilde{g}}(\xi)} + O(|\xi|^{\max(m,\tilde{m})}) \\ &= O(|\xi|^{\max(m,\tilde{m})}). \end{aligned}$$

The third equality follows from $|\hat{\eta}(\xi)|^2 = 1 + O(|\xi|^{\max(m,\tilde{m})})$, and the last equality holds by (3.15). Therefore, (3.13) holds.

With (3.13) and (3.14), Lemma 3.1 says that there exist two functions $h, \tilde{h} \in L_2(\mathbb{R})$ with polynomial decay such that $g = \nabla^m h$ and $\tilde{g} = \nabla^{\tilde{m}} \tilde{h}$.

Let $f_n := Q_a^n \phi_0$ and $\tilde{f}_n := Q_{\tilde{a}}^n \tilde{\phi}_0$. Then their Fourier transforms are

$$\widehat{f_n}(\xi) = \widehat{\phi_0}(2^{-n}\xi) \prod_{j=1}^{n} \hat{a}(2^{-j}\xi) \quad \text{and} \quad \widehat{\tilde{f}_n}(\xi) = \widehat{\tilde{\phi}_0}(2^{-n}\xi) \prod_{j=1}^{n} \hat{\tilde{a}}(2^{-j}\xi).$$

One can prove inductively that

(3.16)
$$f_{n+1} - f_n = Q_a^n g = 2^n \sum_{k\in\mathbb{Z}} a_n(k)[\nabla^m h](2^n\cdot - k) = 2^n \sum_{k\in\mathbb{Z}} [\nabla^m a_n](k) h(2^n\cdot - k)$$

and

(3.17)
$$\tilde{f}_{n+1} - \tilde{f}_n = Q_{\tilde{a}}^n \tilde{g} = 2^n \sum_{k\in\mathbb{Z}} \tilde{a}_n(k)[\nabla^{\tilde{m}} \tilde{h}](2^n\cdot - k) = 2^n \sum_{k\in\mathbb{Z}} [\nabla^{\tilde{m}} \tilde{a}_n](k) \tilde{h}(2^n\cdot - k),$$

where $\widehat{a_n}(\xi) := \hat{a}(2^{n-1}\xi)\cdots\hat{a}(2\xi)\hat{a}(\xi)$ and $\widehat{\tilde{a}_n}(\xi) := \hat{\tilde{a}}(2^{n-1}\xi)\cdots\hat{\tilde{a}}(2\xi)\hat{\tilde{a}}(\xi)$.

Since both h and $\tilde{h}$ are $L_2(\mathbb{R})$ functions with polynomial decay, we conclude that both $[\hat{h}, \hat{h}]$ and $[\hat{\tilde{h}}, \hat{\tilde{h}}]$ are in $L_\infty(\mathbb{R})$. Hence, identities (3.16) and (3.17) imply that there exists a positive constant C such that, for all $n \in \mathbb{N}$,

$$\begin{aligned} &\|f_{n+1} - f_n\|_{L_2(\mathbb{R})} \leqslant C 2^{n/2} \|\nabla^m a_n\|_{\ell_2(\mathbb{Z})} \quad \text{and} \\ &\|\tilde{f}_{n+1} - \tilde{f}_n\|_{L_2(\mathbb{R})} \leqslant C 2^{n/2} \|\nabla^{\tilde{m}} \tilde{a}_n\|_{\ell_2(\mathbb{Z})}. \end{aligned} \tag{3.18}$$

Since

$$\widehat{\nabla^m a_n}(\xi) = (1 - e^{-i\xi})^m \widehat{a_n}(\xi) = 2^{-mn}(1 - e^{-i2^n\xi})^m \widehat{A_n}(\xi)$$

and

$$\widehat{\nabla^{\tilde{m}} \tilde{a}_n}(\xi) = (1 - e^{-i\xi})^{\tilde{m}} \widehat{\tilde{a}_n}(\xi) = 2^{-\tilde{m}n}(1 - e^{-i2^n\xi})^{\tilde{m}} \widehat{\tilde{A}_n}(\xi),$$

we have

$$\|\nabla^m a_n\|_{\ell_2(\mathbb{Z})} \leqslant 2^{m-mn} \|A_n\|_{\ell_2(\mathbb{Z})}, \quad \text{and} \quad \|\nabla^{\tilde{m}} \tilde{a}_n\|_{\ell_2(\mathbb{Z})} \leqslant 2^{\tilde{m}-\tilde{m}n} \|\tilde{A}_n\|_{\ell_2(\mathbb{Z})}.$$

However, (3.5) says that there exist two positive constants ρ with $0 < \rho < 1$ and C_1 such that

$$\|A_n\|_{\ell_2(\mathbb{Z})} \leqslant C_1 \rho^n 2^{(m-1/2)n} \quad \text{and} \quad \|\tilde{A}_n\|_{\ell_2(\mathbb{Z})} \leqslant C_1 \rho^n 2^{(\tilde{m}-1/2)n} \qquad \forall\, n \in \mathbb{N}. \tag{3.19}$$

Therefore, we deduce from (3.18) that

$$\|f_{n+1} - f_n\|_{L_2(\mathbb{R})} \leqslant 2^m C C_1 \rho^n \quad \text{and} \quad \|\tilde{f}_{n+1} - \tilde{f}_n\|_{L_2(\mathbb{R})} \leqslant 2^{\tilde{m}} C C_1 \rho^n \qquad \forall\, n \in \mathbb{N}.$$

Since $0 < \rho < 1$, both $\{f_n\}_{n\in\mathbb{N}}$ and $\{\tilde{f}_n\}_{n\in\mathbb{N}}$ are Cauchy sequences in $L_2(\mathbb{R})$.

Since $\hat{a}(0) = 1$ and $\hat{b}(\pi) \neq 0$, we have $\hat{\tilde{a}}(0) = 1$. Hence, $\lim_{n\to\infty} \widehat{f_n}(\xi) = \hat{\phi}(\xi)$ and $\lim_{n\to\infty} \widehat{\tilde{f}_n}(\xi) = \hat{\tilde{\phi}}(\xi)$ for $\xi \in \mathbb{R}$. This leads to $\lim_{n\to\infty} \|f_n - \phi\|_{L_2(\mathbb{R})} = 0$ and $\lim_{n\to\infty} \|\tilde{f}_n - \tilde{\phi}\|_{L_2(\mathbb{R})} = 0$, since both $\{f_n\}_{n\in\mathbb{N}}$ and $\{\tilde{f}_n\}_{n\in\mathbb{N}}$ are Cauchy sequences in $L_2(\mathbb{R})$.

Next, we prove that $[\hat{\phi}, \hat{\tilde{\phi}}] = 1$. It is clear that $[\widehat{f_0}, \widehat{\tilde{f}_0}] = [\widehat{\phi_0}, \widehat{\tilde{\phi}_0}] = 1$. Using $\overline{\hat{a}(\xi)}\hat{\tilde{a}}(\xi) + \overline{\hat{a}(\xi+\pi)}\hat{\tilde{a}}(\xi+\pi) = 1$, one can prove inductively that $[\widehat{f_n}, \widehat{\tilde{f}_n}] = 1$ for all $n \in \mathbb{N}$. Finally, since $\lim_{n\to\infty} \|f_n - \phi\|_{L_2(\mathbb{R})} = 0$ and $\lim_{n\to\infty} \|\tilde{f}_n - \tilde{\phi}\|_{L_2(\mathbb{R})} = 0$, we must have $[\hat{\phi}, \hat{\tilde{\phi}}] = 1$. Therefore, $(\phi, \tilde{\phi})$ is a pair of refinable functions whose shifts form a pair of biorthogonal systems.

We further note that by the definition of $\tilde{a}$ and $\tilde{b}$, it is easy to verify that

$$\begin{bmatrix} \hat{a}(\xi) & \hat{a}(\xi+\pi) \\ \hat{b}(\xi) & \hat{b}(\xi+\pi) \end{bmatrix} \overline{\begin{bmatrix} \hat{\tilde{a}}(\xi) & \hat{\tilde{a}}(\xi+\pi) \\ \hat{\tilde{b}}(\xi) & \hat{\tilde{b}}(\xi+\pi) \end{bmatrix}}^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.20}$$

With all these relations, (3.7) and the rest of the conclusions of this theorem follow directly from a standard argument in wavelet analysis on biorthogonal wavelets (see [5, 7, 21]. □

In the above proof, the assumption that both the sequences A and B have polynomial decay is used only to apply Lemma 3.1 and to show that $[\hat{h}, \hat{h}] \in L_\infty(\mathbb{R})$ and $[\hat{\tilde{h}}, \hat{\tilde{h}}] \in L_\infty(\mathbb{R})$. Checking the proof of Lemma 3.1, we see that such polynomial decay condition on both A and B can be further weakened.

As a direct consequence of Theorem 3.2, we have the following result.

COROLLARY 3.3. *Let sequences a and b be given in Theorem* 3.2*, and sequences $\tilde{a}, \tilde{b}, A$, and $\tilde{A}$ and functions $\phi, \tilde{\phi}, \psi, \tilde{\psi}$ be defined as in Theorem* 3.2*. Define*

$$\rho_A := \inf_{n\in\mathbb{N}} \|\widehat{A_n}\|^{1/n}_{L_\infty(\mathbb{R})} \quad \text{and} \quad \rho_{\tilde{A}} := \inf_{n\in\mathbb{N}} \|\widehat{\tilde{A}_n}\|^{1/n}_{L_\infty(\mathbb{R})}, \tag{3.21}$$

where A_n and $\tilde{A}_n$ are defined in (3.6)*. Then for any $\varepsilon > 0$, there exists a positive constant C such that*

$$\max(|\hat{\phi}(\xi)|, |\hat{\psi}(\xi)|) \leqslant C(1+|\xi|)^{-m+\varepsilon+\log_2 \rho_A} \qquad \forall\, \xi \in \mathbb{R} \tag{3.22}$$

and

$$\max(|\hat{\tilde{\phi}}(\xi)|, |\hat{\tilde{\psi}}(\xi)|) \leqslant C(1+|\xi|)^{-\tilde{m}+\varepsilon+\log_2 \rho_{\tilde{A}}} \qquad \forall\, \xi \in \mathbb{R}. \tag{3.23}$$

Consequently, if $\rho_A < 2^{m-1/2}$ and $\rho_{\tilde{A}} < 2^{\tilde{m}-1/2}$, then $(X(\psi), X(\tilde{\psi}))$ forms a pair of biorthogonal wavelet bases in $L_2(\mathbb{R})$. In particular, $X(\psi)$ is a Riesz basis of $L_2(\mathbb{R})$.

Proof. The proof of (3.22) and (3.23) follows from the proof of [7, Lemmas 7.1.1 and 7.1.2]. Note that

$$\limsup_{n\to\infty} \|A_n\|^{1/n}_{\ell_2(\mathbb{Z})} \leqslant \limsup_{n\to\infty} \|\widehat{A_n}\|^{1/n}_{L_\infty(\mathbb{R})} = \inf_{n\in\mathbb{N}} \|\widehat{A_n}\|^{1/n}_{L_\infty(\mathbb{R})}$$

by $\|A_n\|_{\ell_2(\mathbb{Z})} \leqslant \|\widehat{A_n}\|_{L_\infty(\mathbb{R})}$. Therefore, if $\rho_A < 2^{m-1/2}$ and $\rho_{\tilde{A}} < 2^{\tilde{m}-1/2}$, then it follows from (3.22) and (3.23) that there exist $\varepsilon > 0$ and $C > 0$ such that

$$\max(|\hat{\phi}(\xi)|, |\hat{\psi}(\xi)|, |\hat{\tilde{\phi}}(\xi)|, |\hat{\tilde{\psi}}(\xi)|) \leqslant C(1+|\xi|)^{-1/2-\varepsilon}$$

and $\max(|\hat{\psi}(\xi)|, |\hat{\tilde{\psi}}(\xi)|) \leqslant C|\xi|^\varepsilon$ for all $\xi \in \mathbb{R}$. Therefore, by [11, Propositions 2.6 and 3.5], both $X(\psi)$ and $X(\tilde{\psi})$ are Bessel. Moreover, it is evident that $[\hat{\phi}, \hat{\phi}] \in L_\infty(\mathbb{R})$ and $[\hat{\tilde{\phi}}, \hat{\tilde{\phi}}] \in L_\infty(\mathbb{R})$. The proof is completed by Theorem 3.2. □

For a refinement mask a, define b as

$$\hat{b}(\xi) = e^{-i\xi}\overline{\hat{a}(\xi+\pi)}.$$

Then

$$\hat{d}(\xi) := \hat{a}(\xi)\hat{b}(\xi+\pi) - \hat{a}(\xi+\pi)\hat{b}(\xi) = e^{-i(\xi+\pi)}(|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2). \tag{3.24}$$

Further, the masks $\tilde{a}$ and $\tilde{b}$ defined in (3.4) take the following form:

$$\hat{\tilde{a}}(\xi) = \frac{\hat{a}(\xi)}{|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2} \quad \text{and} \quad \hat{\tilde{b}}(\xi) = \frac{e^{-i\xi}\overline{\hat{a}(\xi+\pi)}}{|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2}. \tag{3.25}$$

Now, it is clear that Lemma 2.1 becomes a special case of Corollary 3.3. The mask for the B-spline of order m is $\hat{a}(\xi) = 2^{-m}(1+e^{-i\xi})^m$. It follows from (3.24) that $\hat{d}(\xi) = e^{-i(\xi+\pi)}[\cos^{2m}(\xi/2) + \sin^{2m}(\xi/2)] \neq 0$ for all $\xi \in \mathbb{R}$.

The following result is a direct consequence and a slight modification of Theorem 3.2.

COROLLARY 3.4. *Under the same notations and conditions as in Theorem* 3.2, *for any finitely supported sequence* c *on* $\mathbb{Z}$ *such that* $\hat{c}(\xi) \neq 0$ *for all* $\xi \in \mathbb{R}$, *we redefine the functions* ψ *and* $\tilde{\psi}$ *in Theorem* 3.2 *by*

$$\hat{\psi}(\xi) := \hat{c}(\xi)\hat{b}(\xi/2)\hat{\phi}(\xi/2) \quad \textit{and} \quad \hat{\tilde{\psi}}(\xi) := \hat{\tilde{b}}(\xi/2)\hat{\tilde{\phi}}(\xi/2)/\overline{\hat{c}(\xi)}.$$

Then all the claims in Theorem 3.2 *hold.*

Proof. Denote $\hat{b}^{new} := \hat{c}(2\cdot)\hat{b}$ and $\hat{\tilde{b}}^{new} = \hat{\tilde{b}}/\overline{\hat{c}(2\cdot)}$. To apply Theorem 3.2, one observes that

$$\begin{aligned}\hat{d}^{new}(\xi) &:= \hat{a}(\xi)\hat{b}^{new}(\xi+\pi) - \hat{a}(\xi+\pi)\hat{b}^{new}(\xi)\\ &= \hat{c}(2\xi)[\hat{a}(\xi)\hat{b}(\xi+\pi) - \hat{a}(\xi+\pi)\hat{b}(\xi)] \neq 0 \qquad \forall\, \xi \in \mathbb{R}.\end{aligned}$$

It is easy to see that (3.20) holds with $\hat{b}$ and $\hat{\tilde{b}}$ being replaced by $\hat{b}^{new}$ and $\hat{\tilde{b}}^{new}$, respectively. The rest of the proof follows directly from Theorem 3.2. □

Remark. As shown in section 2, Riesz spline wavelets ψ, constructed using Theorem 3.2 and its corollaries, have, in some sense, the shortest support for a given order of smoothness and vanishing moments. However, its dual wavelet systems normally are not compactly supported and have low smoothness orders. Hence, Riesz wavelets constructed here may be used only in applications where either the reconstruction algorithm or the decomposition algorithm is not required. For example, it only needs the decomposition algorithm in some applications of signal analysis and classification. On the other hand, in other applications such as image compression, while the decomposition can be done offline, reconstruction has to be done online. The short supported reconstruction filter is essential, for example, in computer graphics and numerical algorithms. Furthermore, a fast reconstruction algorithm can be derived by adding a system $X(\psi(\cdot - 1/2))$ to $X(\psi)$ to generate a frame system and by using a compactly supported smooth dual wavelet system of the frame system $X(\psi) \cup X(\psi(\cdot - 1/2))$. The detailed discussion is given at the end of the next section.

4. Connections of Riesz wavelets to spline frame systems. The study here on Riesz wavelets also provides a better understanding of the structure of spline tight frame wavelet systems in [20] and [9, 12] by using the *unitary extension principle* of [20] and the *oblique extension principle* of [9] as we shall discuss in this section.

We first briefly describe the constructions of MRA-based tight frames by using the oblique extension principle in [9], where details can be found. For a given refinable function ϕ, with a refinement mask a, one first chooses a 2π-periodic trigonometric polynomial Θ with $\Theta(0) = 1$, called a fundamental function of the tight frame system, according to the approximation order of the refinable function ϕ and the required approximation order of the tight frame expansion (this is directly related to the order

of the vanishing moments of frame wavelets). Suppose that a fundamental function Θ can be chosen so that it satisfies (i) $\Theta(\xi) \geqslant 0$ for all $\xi \in \mathbb{R}$, and (ii)

$$H(\xi) := \Theta(\xi) - \Theta(2\xi)[|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2] \geqslant 0 \qquad \forall\, \xi \in \mathbb{R}. \tag{4.1}$$

The three wavelet masks are then

$$\widehat{b^1}(\xi) := e^{i\xi}\theta(2\xi)\overline{\hat{a}(\xi+\pi)}, \qquad \widehat{b^2}(\xi) := \frac{\sqrt{2}}{2}h(\xi), \quad \text{and} \quad \widehat{b^3}(\xi) := \frac{\sqrt{2}}{2}e^{i\xi}h(\xi),$$

where θ and h are the square roots of Θ and H, respectively; that is, $\theta(\xi)\overline{\theta(\xi)} = \Theta(\xi)$ and $h(\xi)\overline{h(\xi)} = H(\xi)$.

Define the frame wavelet set $\Psi := \{\psi^1, \psi^2, \psi^3\}$ by

$$\widehat{\psi^\ell}(\xi) := \widehat{b^\ell}(\xi/2)\hat{\phi}(\xi/2), \qquad \ell = 1, 2, 3. \tag{4.2}$$

Then it was proven in [9] that the system $X(\Psi)$ forms a tight frame system of $L_2(\mathbb{R})$ by using the oblique extension principle (see [9, Proposition 1.11]).

One can reduce the number of frame wavelets to two by defining

$$\widehat{b^1}(\xi) := e^{i\xi}\theta(2\xi)\overline{\hat{a}(\xi+\pi)}, \quad \widehat{b^2}(\xi) := \hat{a}(\xi)\, h(2\xi).$$

Then it was proven in [9] that the $X(\Psi)$ with $\Psi := \{\psi^1, \psi^2\}$ defined by

$$\widehat{\psi^\ell}(\xi) = \widehat{b^\ell}(\xi/2)\hat{\phi}(\xi/2), \qquad \ell = 1, 2 \tag{4.3}$$

is again a tight frame wavelet system in $L_2(\mathbb{R})$. Note that ψ^1 in (4.2) is the same function as ψ^1 in (4.3).

The construction of spline tight frame systems given in [9] starts with the MRA generated by a B-spline B_m with refinement mask $\hat{a}(\xi) = 2^{-m}(1 + e^{-i\xi})^m$. The fundamental function Θ can be chosen according to the needs of the approximation order of the truncated wavelet system; it satisfies $\Theta(\xi) > 0$ for all $\xi \in \mathbb{R}$, and (4.1). An explicit form of Θ is given in [9, Lemma 3.4]. With Θ, $\hat{a}$, and H, it is easy to obtain the functions defined in (4.2) and (4.3), whose corresponding wavelet systems form spline tight frame systems of $L_2(\mathbb{R})$. In this case, setting $\hat{c}(\xi) = e^{i\xi}\theta(\xi)$ in Corollary 3.4 and applying Theorem 2.2, one can show easily that $X(\psi^1)$ is a Riesz basis for $L_2(\mathbb{R})$, where ψ^1 is defined in either (4.2) or (4.3).

In both spline function sets defined in (4.2) and (4.3), the first function ψ^1 is already able to generate a Riesz basis of $L_2(\mathbb{R})$, i.e., $X(\psi^1)$ is a Riesz basis of $L_2(\mathbb{R})$, which is an even stronger statement than that $X(\psi^1)$ is a frame of $L_2(\mathbb{R})$. The first role of $X(\psi^2)$ and $X(\psi^3)$ in (4.2) and $X(\psi^2)$ in (4.3) is to reduce the condition number determined by the upper and lower frame bounds of $X(\Psi)$ to be one, so that the whole system $X(\Psi)$ becomes a tight frame. The second role of other functions in both (4.2) and (4.3) is to give better dual systems. As we remarked before, the (unique) function corresponding to the biorthogonal dual of the system $X(\psi^1)$ normally has low order of smoothness with infinite support. With the help of ψ^2 and ψ^3 in system (4.2) and ψ^2 in the system (4.3), the whole system $X(\Psi)$ in both cases becomes a tight frame that is a self-dual system. All these are obtained at the cost of changing a nonredundant system $X(\psi^1)$ to a redundant system $X(\Psi)$.

We further note that since the wavelet mask b^2 in (4.3) vanishes at both 0 and π, therefore $[\widehat{\psi^2}, \widehat{\psi^2}](0) = 0$. Since ψ^2 in (4.3) is compactly supported, ψ^2 and its shifts cannot form a Riesz system. This concludes that $X(\psi^2)$ is not a Riesz basis of $L_2(\mathbb{R})$.

The spline tight frame systems constructed above may not be symmetric since the square root of Θ may not be symmetric. This weak point was overcome in [12] by an elegant and careful choice of a symmetric θ such that $\theta^2 = \Theta$ is the required fundamental function satisfying (4.1). Let h be a square root of H and set

$$\begin{aligned}\widehat{b^1}(\xi) &:= e^{-i\xi}\theta(2\xi)\overline{\hat{a}(\xi+\pi)},\\ \widehat{b^2}(\xi) &:= \hat{a}(\xi)[h(2\xi) + \overline{h(2\xi)}]/2,\\ \widehat{b^3}(\xi) &:= \hat{a}(\xi)[h(2\xi) - \overline{h(2\xi)}]/2.\end{aligned}$$

Define

$$\widehat{\psi^\ell}(\xi) := \widehat{b^\ell}(\xi/2)\hat{\phi}(\xi/2), \qquad \ell = 1, 2, 3. \tag{4.4}$$

Then $X(\Psi)$ with $\Psi := \{\psi^1, \psi^2, \psi^3\}$ being defined in (4.4) forms a tight frame system in $L_2(\mathbb{R})$ and all the functions ψ^1, ψ^2, ψ^3 are either symmetric or antisymmetric (see [12] for details). Similarly, $X(\psi^1)$ is a Riesz basis for $L_2(\mathbb{R})$, but both $X(\psi^2)$ and $X(\psi^3)$ cannot be a Riesz basis of $L_2(\mathbb{R})$.

Next, we discuss the spline wavelet system given in [20] via the unitary extension principle of [20]. We discuss the construction from B-splines with an even order. The other case can be discussed similarly. Let ϕ be the centered B-spline of order $2m$. Then its refinement mask is $\hat{a}(\xi) := \cos^{2m}(\xi/2)$. We define $2m$ wavelet masks by

$$\widehat{b^\ell}(\xi) := i^{2m+\ell}\sqrt{\frac{(2m)!}{\ell!(2m-\ell)!}}\sin^\ell(\xi/2)\cos^{2m-\ell}(\xi/2), \quad \ell = 1, \dots, 2m.$$

Then it was shown in [20] that the $2m$ functions, $\Psi = \{\psi^1, \dots, \psi^{2m}\}$, defined by

$$\widehat{\psi^\ell}(\xi) := \widehat{b^\ell}(\xi/2)\hat{\phi}(\xi/2), \qquad \ell = 1, \dots, 2m, \tag{4.5}$$

form a tight frame system of $L_2(\mathbb{R})$ by using the unitary extension principle ([20, Corollary 6.7]).

Consider the function $\psi := \psi^{2m}(\cdot - 1/2)$. Then ψ is the function derived from an MRA generated by the centered B-spline ϕ with the refinement mask $\hat{a}(\xi) := \cos^{2m}(\xi/2)$. The corresponding wavelet mask of ψ is

$$\hat{b}(\xi) = e^{-i\xi}\sin^{2m}(\xi/2) = e^{-i\xi}\overline{\hat{a}(\xi+\pi)}.$$

Hence, $X(\psi)$ forms a Riesz basis of $L_2(\mathbb{R})$ by Theorem 2.2. Since all the other masks $\widehat{b^\ell}$, $1 \leqslant \ell < 2m$, vanish at both 0 and π, a similar discussion as above shows that $X(\psi^\ell)$, $1 \leqslant \ell < 2m$, cannot be a Riesz basis of $L_2(\mathbb{R})$.

Here are some remarks on the two extension principles mentioned above; the interested reader can find more details in [20, 9, 2]. Both the unitary extension principle of [20] and the oblique extension principle of [9] are derived from the characterization of MRA-based frames given in [20] (also see [11]). The unitary extension principle leads to the first set of the examples of compactly supported spline tight frames defined in (4.5). As pointed out in [9], the approximation order (see [9] for definitions) of the truncated tight frame expansion of $X(\Psi)$, where Ψ is defined in (4.5), cannot be over 2. The attempt to derive spline tight frame systems whose truncated expansions have a better approximation order leads to the oblique extension principle of [9], which is a generalization of the unitary extension principle. This leads to a systematic

construction of spline tight wavelet frame systems whose truncated expansion has an arbitrary preassigned approximation order. Functions defined in (4.2) and (4.3) are examples of constructions given in [9]. In a similar fashion, [2] obtained the oblique extension principle independently by an attempt to improve the vanishing moments of the functions obtained from the unitary extension principle of [20].

In the rest of this section, we discuss some relations between Riesz wavelet bases constructed in this paper and bi-frames constructed in [8, 9]. In order to do so, let us recall a result from [8, Corollary 3.4] (also see [9]). Let ϕ be a compactly supported refinable function in $L_2(\mathbb{R})$ with a finitely supported mask a and $\hat{\phi}(0) \neq 0$. For any finitely supported sequence b on $\mathbb{Z}$ such that

$$\hat{b}(0) = 0 \quad \text{and} \quad \lim_{\xi \to \pi} \frac{\hat{a}(\xi)}{\hat{b}(\xi)} = 0, \tag{4.6}$$

define $\hat{\psi}(2\xi) = \hat{b}(\xi)\hat{\phi}(\xi)$; then $X(\{\psi, \psi(\cdot - 1/2)\})$ forms a wavelet frame in $L_2(\mathbb{R})$. Moreover, there exists $\{\tilde{\psi}^1, \tilde{\psi}^2\}$ of compactly supported functions, which are derived via the mixed oblique extension principle (see [9]) from an MRA generated by an arbitrarily chosen compactly supported refinable function $\tilde{\phi} \in L_2(\mathbb{R})$ whose mask contains the factor $(1 + e^{-i\xi})^\ell$, where ℓ is the smallest integer that is greater than the multiplicity of zeros of $\hat{b}(\xi)$ at $\xi = 0$ such that $(X(\{\psi, \psi(\cdot - 1/2)\}), X(\{\tilde{\psi}^1, \tilde{\psi}^2\}))$ is a pair of bi-frames in $L_2(\mathbb{R})$. Since such a pair of bi-frames is MRA-based and all the wavelet masks are finitely supported, a fast frame transform (see [9]) associated with the bi-frames for both decomposition and reconstruction of functions is available.

Now let us discuss some relations between Riesz wavelet bases and bi-frames. Let $\phi \in L_2(\mathbb{R})$ be a compactly supported refinable function with $\hat{\phi}(0) \neq 0$ and a finitely supported mask a such that $\hat{a}(0) = 1$ and $\hat{a}(\pi) = 0$. Let b be a finitely supported sequence on $\mathbb{Z}$ such that $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$, where ψ is defined to be $\hat{\psi}(2\xi) = \hat{b}(\xi)\hat{\phi}(\xi)$. For example, such a wavelet mask b and a function ψ may be chosen as in Theorem 3.2. In particular, when ϕ is a B-spline of order m, the wavelet mask b can be chosen to be

$$e^{-i\xi}\Big(\frac{1 - e^{i\xi}}{2}\Big)^m.$$

Since $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$, the compactly supported function ψ must satisfy $\hat{\psi}(0) = 0$ and $[\hat{\psi}, \hat{\psi}](0) \neq 0$. Since $\hat{\phi}(0) \neq 0$ and $\hat{\psi}(0) = 0$, it follows from the definition $\hat{\psi}(\xi) = \hat{b}(\xi/2)\hat{\phi}(\xi/2)$ that $\hat{b}(0) = 0$. On the other hand, we must have $\hat{b}(\pi) \neq 0$ since otherwise $\hat{b}(0) = \hat{b}(\pi) = 0$ implies $[\hat{\psi}, \hat{\psi}](0) = 0$, which is a contradiction. Now by $\hat{a}(\pi) = 0$, it is evident that the wavelet mask b must satisfy the conditions in (4.6). It follows from our discussion above (see [8, Corollary 3.4]) that $X(\{\psi, \psi(\cdot - 1/2)\})$ is a wavelet frame. Note that $X(\{\psi, \psi(\cdot - 1/2)\}) = X(\psi) \cup X(\psi(\cdot - 1/2))$ with $X(\psi)$ being a Riesz basis. Moreover, there exists a set $\{\tilde{\psi}^1, \tilde{\psi}^2\}$ of compactly supported smooth L_2 functions which are derived via the mixed oblique extension principle (see [9]) from an MRA generated by an arbitrarily chosen smooth refinable function such that $(X(\{\psi, \psi(\cdot - 1/2)\}), X(\{\tilde{\psi}^1, \tilde{\psi}^2\}))$ forms a pair of bi-frames. For this MRA based bi-frame pair, one can derive fast decomposition and reconstruction algorithms as given in [9].

On the other hand, since $X(\psi)$ is a Riesz basis, it has a unique Riesz dual basis, say $X(\tilde{\psi})$. But the unique Riesz dual wavelet $\tilde{\psi}$, as indicated by Theorem 3.2, is not compactly supported and has a low order of smoothness. This indicates that $X(\psi)$,

being a Riesz basis, does not imply that the decomposition (analysis) operator must have an inverse in a function space other than the space $L_2(\mathbb{R})$ that is of interest to us in this paper, since its dual wavelet may have lower smoothness. Furthermore, though the decomposition operator may have an inverse in the function space, it may not have a fast reconstruction algorithm due to the slow decay of the dual masks with infinite support. However, the above discussions show that by introducing redundancy into a Riesz wavelet basis $X(\psi)$, one can obtain a compactly supported dual frame system which has the same smoothness (in some case, it can be extended to a tight frame wavelet system). Hence, the frame decomposition operator can be invertible in various function spaces, and a fast decomposition and reconstruction algorithm can be obtained based on the compactly supported dual frame.

The following example illustrates some of the above discussions. The reader is referred to [8, 9] for more details on bi-frames.

Example 4.1. Let B_3 be the B-spline function of order 3. The refinement mask for the refinable function B_3 is $\hat{a}(\xi) = 2^{-3}(1+e^{-i\xi})^3$. Define $\psi = \frac{1}{4}B_3(2\cdot -1) - \frac{3}{4}B_3(2\cdot) + \frac{3}{4}B_3(2\cdot +1) - \frac{1}{4}B_3(2\cdot +2)$, that is, $\hat{\psi}(2\xi) = 2^{-3}e^{-i\xi}(1-e^{i\xi})^3\widehat{B_3}(\xi)$. By Theorem 2.2, $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$. The function ψ has 3 vanishing moments and the regularity 5/2. By Theorem 2.2, its unique Riesz dual basis $X(\tilde{\psi})$ must have a noncompactly supported generator $\tilde{\psi}$. However, the frame system $X(\psi)\cup X(\psi(\cdot - 1/2))$ can have a compactly supported smooth dual frame as defined below. We choose $\tilde{\phi} = B_3$. Let $\tilde{b}^1$ and $\tilde{b}^2$ be two finitely supported sequences on $\mathbb{Z}$ given by

$$\widehat{b^1}(\xi) := \frac{(z-1)^3}{1920}\Big[13\,(z^{-6}+z^2) + 78\,(z^{-5}+z) + 356\,(z^{-4}+1) + 1226\,(z^{-3}+z^{-1}) + 2334\,z^{-2}\Big],$$

$$\widehat{b^2}(\xi) := \frac{(z-1)^3}{960}\Big[39\,(z^{-4}+z^2) + 234\,(z^{-3}+z) + 613\,(z^{-2}+1) + 948z^{-1}\Big],$$

where $z := e^{-i\xi}$. Define the dual wavelet functions $\tilde{\psi}^1$ and $\tilde{\psi}^2$ by

$$\widehat{\tilde{\psi}^1}(\xi) = \widehat{\tilde{b}^1}(\xi/2)\widehat{B_3}(\xi/2), \qquad \widehat{\tilde{\psi}^2}(\xi) = \widehat{\tilde{b}^2}(\xi/2)\widehat{B_3}(\xi/2).$$

Then it has been proved in [8, 9] that $(X(\{\psi, \psi(\cdot - 1/2)\}), X(\{\tilde{\psi}^1, \tilde{\psi}^2\}))$ is a pair of bi-frames in $L_2(\mathbb{R})$. Note that both $\tilde{\psi}^1$ and $\tilde{\psi}^2$ are compactly supported and antisymmetric. Moreover, they have the regularity 5/2 and 3 vanishing moments. See Figure 3 for the graphs of all the functions ψ, $\tilde{\psi}^1$ and $\tilde{\psi}^2$.

5. Other examples of Riesz wavelets with short support. In this section, we first show that one can further greatly shorten the support of the Riesz spline wavelets with the same order of smoothness given in section 2 by reducing the order of vanishing moments of the wavelets. The second part gives examples which are derived from interpolatory refinable functions.

THEOREM 5.1. *Let B_m be the B-spline function of order m. For any integer $\tilde{m}$ such that $\tilde{m} \geqslant m\log_6(8/3)$ (note that $\log_6(8/3) \approx 0.5474$) and $\tilde{m}+m$ is an even integer, define $\hat{\psi}(2\xi) := e^{i\xi(m-\tilde{m}-2)/2}(\frac{1-e^{i\xi}}{2})^{\tilde{m}}\widehat{B_m}(\xi)$. Then $X(\psi)$ is a Riesz wavelet basis in $L_2(\mathbb{R})$.*

Proof. We use Corollary 3.3. For this, we first note that the refinement mask of the spline B_m is $\hat{a}(\xi) = (\frac{1+e^{-i\xi}}{2})^m\hat{A}(\xi)$ with $\hat{A}(\xi) = 1$. Let $\hat{b}(\xi) := e^{i\xi(m-\tilde{m}-2)/2}(\frac{1-e^{i\xi}}{2})^{\tilde{m}}$

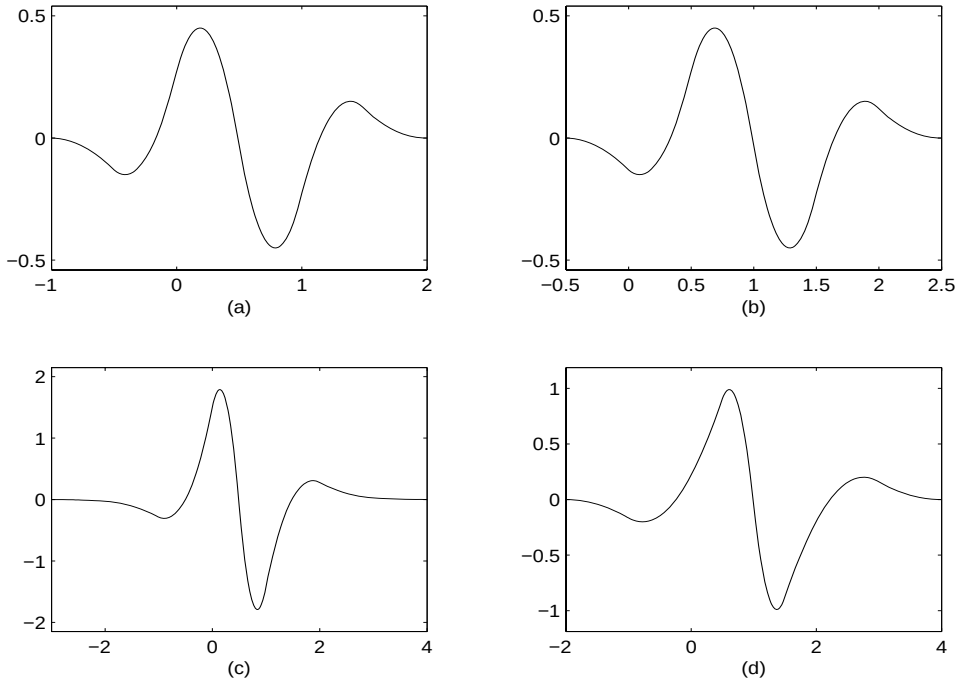


FIG. 3. *The graphs of the functions ψ and $\psi(\cdot - 1/2)$ (top row) and the graphs of the dual functions $\tilde{\psi}^1$ and $\tilde{\psi}^2$ (bottom row) in Example* 4.1. *All the functions ψ, $\psi(\cdot - 1/2)$, $\tilde{\psi}^1$ and $\tilde{\psi}^2$ have* 3 *vanishing moments and the regularity* 5/2. *The system $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$ and $(X(\{\psi, \psi(\cdot - 1/2)\}), X(\{\tilde{\psi}^1, \tilde{\psi}^2\}))$ is a pair of bi-frames derived from the B-spline function B_3 of order* 3.

and define $\hat{\psi}(2\xi) = \hat{b}(\xi)\widehat{B_m}(\xi)$. Then

$$\begin{aligned} \hat{d}(\xi) &:= \hat{a}(\xi)\hat{b}(\xi + \pi) - \hat{a}(\xi + \pi)\hat{b}(\xi) \\ &= e^{-i\xi}(-1)^{(m-\tilde{m}-2)/2}[\cos^{m+\tilde{m}}(\xi/2) + \sin^{m+\tilde{m}}(\xi/2)] \neq 0. \end{aligned} \tag{5.1}$$

The mask $\hat{\tilde{a}}$ in Theorem 3.2 is

$$\hat{\tilde{a}}(\xi) = \left(\frac{1 + e^{-i\xi}}{2}\right)^{\tilde{m}} \hat{\tilde{A}}(\xi) \quad \text{with} \quad \hat{\tilde{A}}(\xi) := \frac{e^{-i\xi(m-\tilde{m})/2}}{\cos^{m+\tilde{m}}(\xi/2) + \sin^{m+\tilde{m}}(\xi/2)}.$$

Consider $f_n(x) := x^n + (1-x)^n, 0 \leqslant x \leqslant 1$, and $n \in \mathbb{N}$. Since $f_n'(x) = n[x^{n-1} - (1-x)^{n-1}]$, f_n decreases on $[0, 1/2]$ and increases on $[1/2, 1]$. Therefore, $f_n(x) \geqslant f_n(1/2)$ for all $x \in [0, 1]$.

Since f_n decreases on $[0, 1/4]$, we have

$$f_n((2x-1)^2)f_n(x) \geqslant f_n(1/2)f_n(x) \geqslant f_n(1/2)f_n(1/4) \qquad \forall\, x \in [0, 1/4].$$

On the other hand, since the map $x \mapsto (2x-1)^2$ maps the interval $[1/4, 1/2]$ onto the interval $[0, 1/4]$, we have

$$f_n((2x-1)^2)f_n(x) \geqslant f_n(1/4)f_n(x) \geqslant f_n(1/2)f_n(1/4) \qquad \forall\, x \in [1/4, 1/2].$$

By the symmetry of f_n on $[0,1]$, we conclude that

$$f_n((2x-1)^2)f_n(x) \geqslant f_n(1/2)f_n(1/4) = 2^{1-n}(4^{-n}+3^n4^{-n}) > 2^{1-3n}3^n \qquad \forall\, x\in[0,1], n\in\mathbb{N}.$$

Now, since $m+\tilde{m}$ is an even integer, we observe that for all $\xi\in\mathbb{R}$,

$$|\hat{\tilde{A}}(2\xi)\hat{\tilde{A}}(\xi)| = \left[f_{(m+\tilde{m})/2}((2\cos^2(\xi)-1)^2)f_{(m+\tilde{m})/2}(\cos^2(\xi/2))\right]^{-1} < 2^{3(m+\tilde{m})/2-1}3^{-(m+\tilde{m})/2}.$$

Since $2^{3(m+\tilde{m})/2-1}3^{-(m+\tilde{m})/2} \leqslant 2^{2\tilde{m}-1}$ for all $\tilde{m} \geqslant m\log_6(8/3)$, we conclude that $\rho_{\tilde{A}} < 2^{\tilde{m}-1/2}$ for all $\tilde{m}\geqslant m\log_6(8/3)$. It is clear that $\rho_A = 1 < 2^{m-1/2}$. Hence, $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$ by Corollary 3.3. □

The assumption that $m+\tilde{m}$ is an even integer in Theorem 5.1 is used only to guarantee that (5.1) holds. A more refined analysis can be employed in Theorem 5.1 to show that $\rho_{\tilde{A}} \leqslant (4/3)^{(m+\tilde{m})/2}$ by estimating $\hat{\tilde{A}}_3$. Therefore, $\tilde{m} > m\log_3(4/3)+\log_3 2$ is enough and $\log_3(4/3)\approx 0.26186$.

Example 5.2. Let $m=4$ and $\tilde{m}=2$. Define $\hat{b}(\xi) = (1-e^{i\xi})^2/4$ and $\hat{\psi}(2\xi) = \hat{b}(\xi)\widehat{B_4}(\xi)$. Since $m+\tilde{m}=6$ is an even number and $\tilde{m} > m\log_3(4/3)+\log_3 2 \approx 1.6784$, $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$. The function ψ has 2 vanishing moments and the regularity 7/2. See Figure 4 for the graphs of the functions B_4 and ψ.

In the following, we consider a continuous refinable function ϕ that satisfies the condition $\phi(k)=\delta(k)$, $k\in\mathbb{Z}$. Such a refinable function is called an *interpolatory refinable function*. Clearly, the centered piecewise linear B-spline B_2 is interpolatory. However, the higher order B-splines are not interpolatory. Smooth interpolatory refinable functions can be obtained by a convolution of a B-spline with a distribution. More precisely, for any positive integer m, the mask a given by

$$\hat{a}(\xi) = \cos^{2m}(\xi/2)P_m(\sin^2(\xi/2)), \tag{5.2}$$

where

$$P_m(x) := \sum_{j=0}^{m-1}\frac{(m-1+j)!}{j!(m-1)!}x^j, \qquad x\in\mathbb{R}, \tag{5.3}$$

defines an interpolatory refinable function ϕ with the refinement mask a in (5.2).

This set of masks for interpolatory refinable functions was provided in [6]. Each of them is the autocorrelation of a refinement mask of some refinable function whose shifts form an orthonormal system. Interpolatory masks were constructed first, then the masks of the compactly supported orthonormal refinable functions were obtained as a square root of $\hat{a}$ in [6]. More details about this construction can be found in [6, 7].

THEOREM 5.3. *Let m be a positive integer. Let ϕ be the interpolatory refinable function with the mask a defined in* (5.2). *Define $\hat{\psi}(2\xi) = e^{-i\xi}\overline{\hat{a}(\xi+\pi)}\hat{\phi}(\xi)$. Then $X(\psi)$ is a Riesz basis in $L_2(\mathbb{R})$.*

Proof. To apply Lemma 2.1, we first note that

$$\hat{a}(\xi) = 2^{-2m}(1+e^{-i\xi})^{2m}\hat{A}(\xi) \quad \text{with} \quad \hat{A}(\xi) := e^{im\xi}P_m(\sin^2(\xi/2)),$$

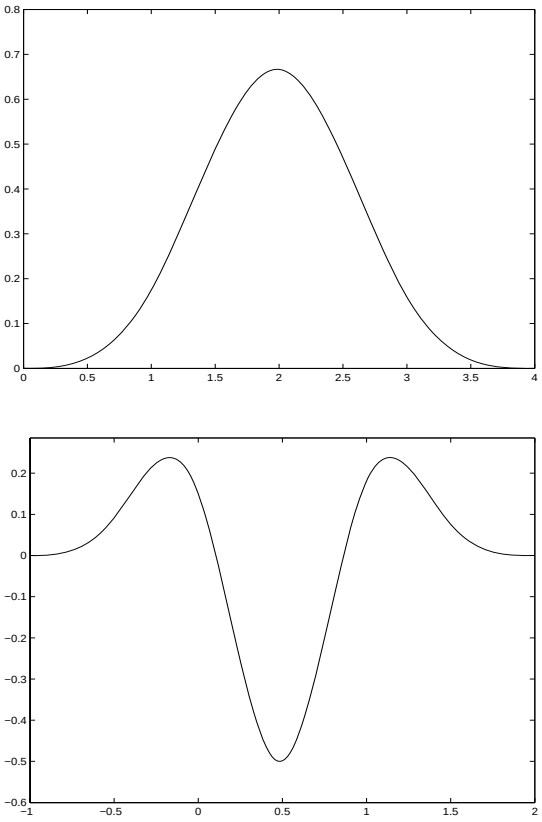


FIG. 4. *The graph of the spline B_4 (top) and the graph of the function ψ (bottom) in Example* 5.2. *The function ψ has* 2 *vanishing moments and the regularity* 7/2. *The system $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$.*

and

$$\hat{\tilde{A}}(\xi) = \frac{\hat{A}(\xi)}{|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2}.$$

Then [7, Lemmas 7.1.7 and 7.1.8] say that $\rho_A \leqslant P_m(3/4) \leqslant 3^{m-1}$ for all $m \in \mathbb{N}$ (see [7, p. 226]). Therefore, $\rho_A \leqslant 3^{m-1} < 2^{2m-1/2}$ for all $m \in \mathbb{N}$.

Since the mask of any interpolatory refinable function must satisfy

$$\hat{a}(\xi) + \hat{a}(\xi+\pi) = 1, \qquad \xi \in \mathbb{R},$$

we have $|\hat{a}(\xi)|^2 + |\hat{a}(\xi+\pi)|^2 \geqslant 1/2$ for all $\xi \in \mathbb{R}$. Thus, $|\hat{\tilde{A}}| \leqslant 2|\hat{A}|$. This leads to

$$\rho_{\tilde{A}} \leqslant 2\rho_A \leqslant 2 \times 3^{m-1} < 2^{2m-1/2} \qquad \forall\, m \in \mathbb{N}.$$

Hence, $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$ by Lemma 2.1. □

Example 5.4. Let ϕ be the interpolatory refinable function with the mask $\hat{a}(\xi) = \frac{1}{2} + \frac{9}{16}\cos(\xi) - \frac{1}{16}\cos(3\xi)$, that is, the mask a is given in (5.2) with $m = 2$. Define $\hat{b}(\xi) = e^{-i\xi}\overline{\hat{a}(\xi+\pi)}$ and $\hat{\psi}(2\xi) = \hat{b}(\xi)\hat{\phi}(\xi)$. By Theorem 5.3, $X(\psi)$ is a Riesz wavelet basis for $L_2(\mathbb{R})$. The function ψ has 4 vanishing moments and the regularity 2.44077. See Figure 5 for graphs of the interpolatory refinable function ϕ and the function ψ.

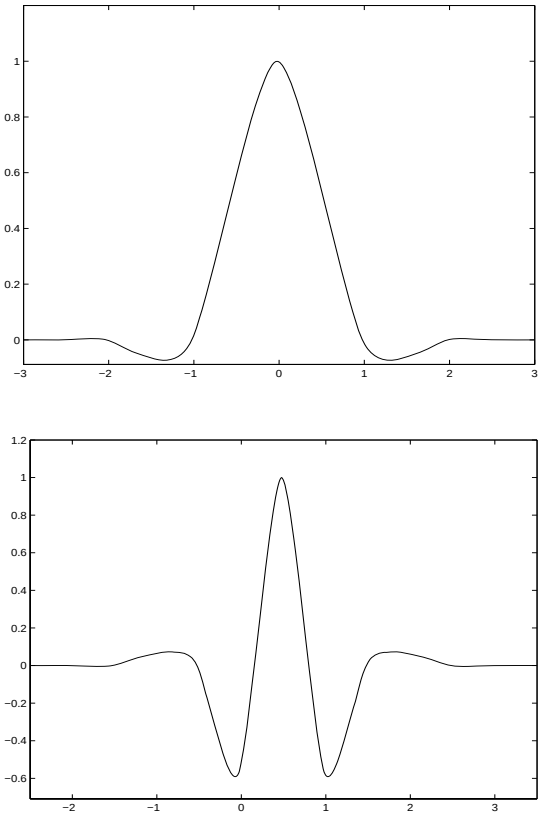


FIG. 5. *The graph of the interpolatory refinable function ϕ (top) and the graph of the function ψ (bottom) in Example* 5.4. *The function ψ has* 4 *vanishing moments and the regularity* 2.44077. *The system $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$.*

The support of the function in Theorem 5.3 can be further shortened by reducing the order of vanishing moments as shown in the next result.

THEOREM 5.5. *Let m be a positive integer. Let ϕ be the interpolatory refinable function with the mask a given in* (5.2). *Define $\hat{\psi}(2\xi) = (\frac{1-e^{i\xi}}{2})^2\hat{\phi}(\xi)$. Then $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$.*

Proof. We apply Corollary 3.3. For this, we note that

$$\hat{a}(\xi) = 2^{-2m}(1+e^{-i\xi})^{2m}\hat{A}(\xi) \quad \text{with} \quad \hat{A}(\xi) := e^{im\xi}P_m(\sin^2(\xi/2)),$$
$$\hat{b}(\xi) := (1-e^{i\xi})^2/4,$$

and

$$\hat{d}(\xi) = e^{i\xi}[\cos^{2m+2}(\xi/2)P_m(\sin^2(\xi/2)) + \sin^{2m+2}(\xi/2)P_m(\cos^2(\xi/2))] \neq 0 \qquad \forall\, \xi \in \mathbb{R}.$$

Then $\hat{\tilde{a}}(\xi) = 2^{-2}(1+e^{-i\xi})^2\hat{\tilde{A}}(\xi)$ with

$$\hat{\tilde{A}}(\xi) := \frac{e^{i\xi}}{\cos^{2m+2}(\xi/2)P_m(\sin^2(\xi/2)) + \sin^{2m+2}(\xi/2)P_m(\cos^2(\xi/2))}.$$

To apply Corollary 3.3, it remains to estimate $\hat{\tilde{A}}$. We note that for positive numbers a_1, a_2, a_3, a_4, it is easy to verify that $\frac{a_1}{a_2} \leqslant \frac{a_3}{a_4}$ implies $\frac{a_1}{a_2} \leqslant \frac{a_1+a_3}{a_2+a_4} \leqslant \frac{a_3}{a_4}$.

Since $P_m(x) = \sum_{j=0}^{m-1} c_j x^j$, where $c_j := \frac{(m-1+j)!}{j!(m-1)!} > 0$, it follows from

$$\frac{c_0}{c_0} \leqslant \frac{c_1(1-x)}{c_1 x} \leqslant \cdots \leqslant \frac{c_{m-1}(1-x)^{m-1}}{c_{m-1}x^{m-1}} = \frac{(1-x)^{m-1}}{x^{m-1}}, \qquad x \in (0, 1/2],$$

that

$$\frac{P_m(1-x)}{P_m(x)} \leqslant \frac{(1-x)^{m-1}}{x^{m-1}} \leqslant \frac{(1-x)^m}{x^m} \qquad \forall\, x \in (0, 1/2].$$

In other words, we have $x^m P_m(1-x) \leqslant (1-x)^m P_m(x)$ for all $x \in [0, 1/2]$. Thus, we deduce that

$$(1-x-x)x^m P_m(1-x) \leqslant (1-x-x)(1-x)^m P_m(x) \qquad \forall\, x \in [0, 1/2],$$

which is equivalent to

$$(1-x)x^m P_m(1-x) - x^{m+1} P_m(1-x) \leqslant (1-x)^{m+1} P_m(x) - x(1-x)^m P_m(x) \qquad \forall\, x \in [0, 1/2].$$

Hence,

$$(1-x)x^m P_m(1-x) + x(1-x)^m P_m(x) \leqslant x^{m+1} P_m(1-x) + (1-x)^{m+1} P_m(x) \qquad \forall\, x \in [0, 1/2].$$

Note that $x^m P_m(1-x) + (1-x)^m P_m(x) = 1$ for all $x \in [0, 1/2]$. For any $x \in [0, 1/2]$, we have

$$\begin{aligned} 1 &= (1-x+x)[x^m P_m(1-x) + (1-x)^m P_m(x)] \\ &= [(1-x)x^m P_m(1-x) + x(1-x)^m P_m(x)] + [x^{m+1} P_m(1-x) + (1-x)^{m+1} P_m(x)] \\ &\leqslant 2[x^{m+1} P_m(1-x) + (1-x)^{m+1} P_m(x)]. \end{aligned}$$

Consequently, by symmetry, we deduce that

$$x^{m+1} P_m(1-x) + (1-x)^{m+1} P_m(x) \geqslant 1/2 \qquad \forall\, x \in [0, 1].$$

Using the above inequality and taking $x = \cos^2(\xi/2)$, we have

$$|\hat{\tilde{A}}(\xi)| = [x^{m+1} P_m(1-x) + (1-x)^{m+1} P_m(x)]^{-1} \leqslant 2 \qquad \forall\, \xi \in \mathbb{R}.$$

It follows from the definition of $\rho_{\tilde{A}}$ that $\rho_{\tilde{A}} \leqslant 2 < 2^{2-1/2}$. Hence, $X(\psi)$ must be a Riesz basis for $L_2(\mathbb{R})$ by Corollary 3.3. □

Example 5.6. Let ϕ be the interpolatory refinable function with the mask $\hat{a}(\xi) = \frac{1}{2} + \frac{9}{16}\cos(\xi) - \frac{1}{16}\cos(3\xi)$, that is, the mask a is given in (5.2) with $m = 2$. Define $\hat{b}(\xi) = (1 - e^{i\xi})^2/4$ and $\hat{\psi}(2\xi) = \hat{b}(\xi)\hat{\phi}(\xi)$. By Theorem 5.5, $X(\psi)$ is a Riesz wavelet basis for $L_2(\mathbb{R})$. The function ψ has 2 vanishing moments and the regularity 2.44077. See Figure 6 for graphs of the interpolatory refinable function ϕ and the function ψ.

REFERENCES

[1] C. de Boor, R. DeVore, and A. Ron, *On the construction of multivariate (pre)wavelets*, Constr. Approx., 9 (1993), pp. 123–166.

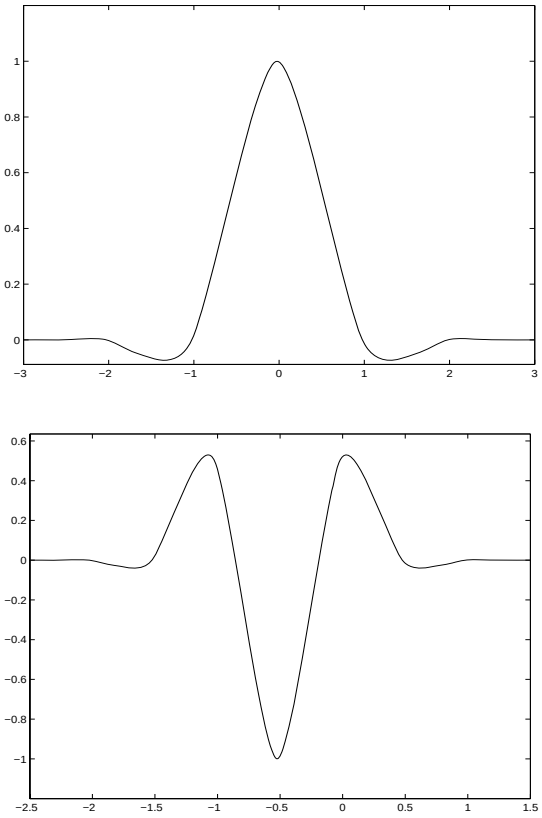


FIG. 6. *The graph of the interpolatory refinable function ϕ (top) and the graph of the function ψ (bottom) in Example* 5.6. *The function ψ has* 2 *vanishing moments and the regularity* 2.44077. *The system $X(\psi)$ is a Riesz basis for $L_2(\mathbb{R})$.*

[2] C. K. CHUI, W. HE, AND J. STÖCKLER, *Compactly supported tight and sibling frames with maximum vanishing moments*, Appl. Comput. Harmon. Anal., 13 (2002), pp. 224–262.
[3] C. K. CHUI AND J. WANG, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–915.
[4] A. COHEN AND I. DAUBECHIES, *A new technique to estimate the regularity of refinable functions*, Rev. Mat. Iberoamericana, 12 (1996), pp. 527–591.
[5] A. COHEN, I. DAUBECHIES, AND J.-C. FEAUVEAU, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.
[6] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.
[7] I. DAUBECHIES, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.
[8] I. DAUBECHIES AND B. HAN, *Pairs of dual wavelet frames from any two refinable functions*, Constr. Approx., 20 (2004), pp. 325–352.
[9] I. DAUBECHIES, B. HAN, A. RON, AND Z. SHEN, *Framelets: MRA-based constructions of wavelet frames*, Appl. Comput. Harmon. Anal., 14 (2003), pp. 1–46.
[10] B. HAN, *Vector cascade algorithms and refinable function vectors in Sobolev spaces*, J. Approx. Theory, 124 (2003), pp. 44–88.
[11] B. HAN, *On dual wavelet tight frames*, Appl. Comput. Harmon. Anal., 4 (1997), pp. 380–413.
[12] B. HAN AND Q. MO, *Tight wavelet frames generated by three symmetric B-spline functions with high vanishing moments*, Proc. Amer. Math. Soc., 132 (2004), pp. 77–86.
[13] R. Q. JIA AND C. A. MICCHELLI, *Using the refinement equations for the construction of pre-wavelets. II. Power of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and

L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.
[14] R.-Q. Jia, *Approximation properties of multivariate wavelets*, Comp. Math., 67 (1998), pp. 647–665.
[15] R. Q. Jia and Z. Shen, *Multiresolution and wavelets*, Proc. Edinburgh Math. Soc. (2), 37 (1994), pp. 271–300.
[16] R.-Q. Jia, J. Wang, and D.-X. Zhou, *Compactly supported wavelet bases for Sobolev spaces*, Appl. Comput. Harmon. Anal., 15 (2003), pp. 224–241.
[17] R. Lorentz and P. Oswald, *Criteria for hierarchical bases in Sobolev spaces*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 32–85.
[18] A. Ron, *Factorization theorems of univariate splines on regular grids*, Israel J. Math., 70 (1990), pp. 48–68.
[19] A. Ron, *Smooth refinable functions provide good approximation orders*, SIAM J. Math. Anal., 28 (1997), pp. 731–748.
[20] A. Ron and Z. Shen, *Affine systems in $L_2(\mathbb{R}^d)$: The analysis of the analysis operator*, J. Funct. Anal., 148 (1997), pp. 408–447.
[21] A. Ron and Z. Shen, *Affine systems in $L_2(\mathbb{R}^d)$* II: *Dual systems*, J. Fourier Anal. Appl., 3 (1997), pp. 617–637.
[22] A. Ron and Z. Shen, *Gramian analysis of affine basis and affine frames*, in Approximation Theory VIII, Vol. 2, Wavelets and Multilevel Approximation, C. K. Chui and L. L. Schumaker, eds., World Sci. Publishing, River Edge, New Jersey, 1995, pp. 375–382.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 557–573

NEUMANN PROBLEMS FOR QUASI-LINEAR PARABOLIC SYSTEMS MODELING POLYDISPERSE SUSPENSIONS*

STEFAN BERRES†, RAIMUND BÜRGER‡, AND HERMANO FRID§

Abstract. We discuss the well-posedness of a class of Neumann problems for $n \times n$ quasi-linear parabolic systems arising from models of sedimentation of polydisperse suspensions in engineering applications. This class of initial-boundary value problems includes the standard (zero-flux) Neumann condition in the limit as a positive perturbation parameter θ goes to 0. We call, in general, the problem associated with $\theta \geq 0$ the θ-flux Neumann problem. The Neumann boundary conditions, although natural and usually convenient for integration by parts, are nonlinear and couple the different components of the system. An important aspect of our analysis is a time stepping procedure that considers linear boundary conditions for each time step in order to circumvent the difficulties arising from the nonlinear coupling in the original boundary conditions. We prove the well-posedness of the θ-flux Neumann problems for $\theta > 0$ and obtain a solution of the standard (zero-flux) Neumann problem as the limit for $\theta \to 0$ of solutions of the θ-flux Neumann problems. Concerning applications, the analysis developed here supports a new model for the settling of polydisperse suspensions forming compressible sediments.

Key words. quasi-linear parabolic system, boundary value problem, polydisperse suspensions

AMS subject classifications. Primary, 35K50; Secondary, 76S05

DOI. 10.1137/050635195

1. Introduction.

We consider the initial-boundary value problem for certain quasi-linear parabolic systems of the (upper triangular) form

$$\partial_t u_i + \partial_x f_i(u) = \sum_{j=i}^{n} \partial_x \big(B_{ij}(u) \partial_x u_j \big), \quad i = 1, \ldots, n, \tag{1.1}$$

where $u = (u_1, \ldots, u_n)^{\mathrm{T}}$ and $(x,t) \in Q := (-1,1) \times (0,T)$. We consider the initial condition

$$u(x,0) = u_0(x), \quad x \in \Omega := (-1,1), \tag{1.2}$$

where $u_0 = (u_{01}, \ldots, u_{0n})^{\mathrm{T}}$ is a function on Ω for which regularity assumptions are made below. The so-called θ-flux Neumann boundary conditions are given by

$$f_i(u) - \sum_{j=i}^{n} B_{ij}(u) \big(\partial_x u_j \pm \theta (u_j - u_j^{\pm}(t)) \big) = 0, \quad x = \pm 1, \quad i = 1, \ldots, n. \tag{1.3}$$

*Received by the editors July 5, 2005; accepted for publication (in revised form) January 17, 2006; published electronically June 9, 2006. The first and second authors were supported by the Sonderforschungsbereich 404 at the University of Stuttgart and by the German Academic Exchange Service (DAAD) and CONICYT (Chile) through project Alechile/DAAD/CONICYT 2003154. The second author was supported by Fondecyt project 1050728 and Fondap in Applied Mathematics. The third author was partially supported by CNPq grants 352871/96-2, 46.5714/00-5, and 479416/2001-0 and by FAPERJ grant E-26/152.192/2002.

http://www.siam.org/journals/sima/38-2/63519.html

†Institut für Angewandte Analysis und Numerische Simulation, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany (berres@mathematik.uni-stuttgart.de).

‡Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile (rburger@ing-mat.udec.cl).

§Instituto de Matemática Pura e Aplicada (IMPA), Estrada Dona Castorina 110, Jardim Botânico, CEP 22460-320, Rio de Janeiro, RJ, Brazil (hermano@impa.br).

Here, $\theta > 0$ and $u^{\pm} = (u_1^{\pm}, \dots, u_n^{\pm})^{\mathrm{T}}$ are functions on $(0,T)$ for which regularity assumptions are also stated below. In matrix form, (1.1) and (1.3) read

$$\partial_t u + \partial_x f(u) = \partial_x\big(B(u)\partial_x u\big), \quad (x,t) \in Q, \tag{1.4}$$

$$f(u) - B(u)\big(\partial_x u \pm \theta(u - u^{\pm}(t))\big) = 0, \quad x = \pm 1, \quad 0 < t < T, \tag{1.5}$$

where we denote $f(u) = (f_1(u), \dots, f_n(u))^{\mathrm{T}}$ and $B(u) = (B_{ij}(u))_{i,j=1}^n$. We impose the parabolicity condition

$$B_{ii}(u) \geq \nu > 0, \quad u \in \Delta, \quad i = 1, \dots, n, \tag{1.6}$$

and assume that the matrix $B(u)$ is upper triangular, i.e.,

$$B_{ij}(u) \equiv 0 \quad \text{if } i > j. \tag{1.7}$$

When $\theta \to 0$, condition (1.3) reduces to the standard (zero-flux) Neumann boundary condition given by

$$f_i(u) - \sum_{j=i}^{n} B_{ij}(u)\partial_x u_j = 0, \quad x = \pm 1, \quad t \in (0,T), \quad i = 1, \dots, n. \tag{1.8}$$

We assume that the functions $f_i(u)$ and $B_{ij}(u)$ are smooth on the set

$$\Delta := \big\{u \in \mathbb{R}^n \,:\, u_1 \geq 0, \dots, u_n \geq 0,\ u_1 + \cdots + u_n \leq 1\big\} \tag{1.9}$$

and that

$$u_0(x) \in \Delta,\ u^{\pm}(t) \in \Delta \quad \text{for all } x \in \Omega,\ t \in (0,T). \tag{1.10}$$

More precisely, we assume

$$f_i,\ \frac{\partial f_i}{\partial u_j},\ B_{ij},\ \frac{\partial B_{ij}}{\partial u_k},\ \frac{\partial^2 B_{ij}}{\partial u_k \partial u_l} \in H^{\beta}(\Delta),\ \beta \in (0,1), \tag{1.11}$$

where $H^{\beta}(\Delta)$ is the space of Hölder continuous functions (with Hölder exponent β) defined on Δ.

Our interest in problem (1.1), (1.2), (1.8) comes from a mathematical model for the sedimentation of polydisperse suspensions in engineering applications. This model will be further analyzed in this paper. Guided by this model, we assume the following conditions, which are relevant for the invariance of Δ:

$$f_i|_{u_i=0} = 0, \quad i = 1, \dots, n; \quad (f_1 + \cdots + f_n)\big|_{u_1+\cdots+u_n=1} = 0, \tag{1.12}$$

$$\begin{cases} \{B_{ij}\big|_{u_i=0} = \mu_i\, \delta_{ij}, \quad i,j = 1, \dots, n, \\ B_{11}\big|_{u_1+\cdots+u_n=1} = (B_{12} + B_{22})\big|_{u_1+\cdots+u_n=1} \\ \quad = \cdots = \sum_{j=1}^{n} B_{jn}\big|_{u_1+\cdots+u_n=1}, \end{cases} \tag{1.13}$$

where $\delta_{ij} = 0$ for $i \neq j$, $\delta_{ii} = 1$, and the μ_i are positive functions defined over the hyperplanes $u_i = 0$.

We denote by $H^{k+\beta}([a,b])$ the space of the functions on the closed interval $[a,b]$ whose derivatives up to kth order are Hölder continuous with exponent β. By

$H^{\alpha,\alpha/2}(\bar{Q})$ we denote the space of Hölder continuous functions on $\bar{Q}$ associated with the norm

$$|u|_Q^{(\alpha)} := \sup_{\bar{Q}} |u(x,t)| + \sup_{x_1,x_2\in\overline{\Omega},\ t\in[0,T]} \frac{|u(x_1,t)-u(x_2,t)|}{|x_1-x_2|^\alpha} + \sup_{x\in\overline{\Omega},\ t_1,t_2\in[0,T]} \frac{|u(x,t_1)-u(x,t_2)|}{|t_1-t_2|^{\alpha/2}}.$$

Let $H^{2+\alpha,1+\alpha/2}(Q)$ be the space of the functions on Q whose first and second x-derivatives and the t-derivative belong to $H^{\alpha,\alpha/2}(K)$ for any compact set $K \subseteq Q$. For simplicity, we also write, e.g., $v \in H^{\alpha,\alpha/2}(K)$ for a vector function $v = (v_1,\dots,v_n)^{\mathrm{T}}$ if all components of v belong to $H^{\alpha,\alpha/2}$.

The following theorem states our main result concerning the θ-flux Neumann problem (1.1)–(1.3).

THEOREM 1.1. *Let $\theta > 0$ be fixed. Assume that*

$$u_0 \in H^{2+\beta}([-1,1]), \quad u^\pm \in H^{1+\beta}([0,T]), \tag{1.14}$$

and that the compatibility conditions

$$f_i\big(u_0(\pm1)\big) - \sum_{j=i}^n B_{ij}\big(u_0(\pm1)\big)\big(\partial_x u_{0j}(\pm1) \pm \theta(u_{0j}(\pm1) - u_j^\pm(0))\big) = 0, \quad i = 1,\dots,n, \tag{1.15}$$

are satisfied. Furthermore, suppose that $f(u), B(u), u_0, u^\pm$ satisfy (1.6), (1.7), *and* (1.10)–(1.13). *Then, problem* (1.1)–(1.3) *has a unique classical solution u such that $u \in H^{2+\beta,1+\beta/2}(Q)$ and* (1.1) *is satisfied on Q, $u, u_x \in H^{\alpha,\alpha/2}(\bar{Q})$ for some $0 < \alpha \le \beta$, and conditions* (1.2), (1.5) *are satisfied in the usual sense of continuous functions. Moreover, $u(x,t) \in \Delta$ for $(x,t) \in \bar{Q}$.*

We also state here our main result for the limit problem (1.1), (1.2), (1.8).

THEOREM 1.2. *Let $f(u)$ and $B(u)$ be as in Theorem* 1.1. *Assume $u_0 \in L^2(\Omega)$. Then problem* (1.1), (1.2), (1.8) *has a solution $u(x,t)$ in the sense that u satisfies* (1.1) *in the classical sense in Q, u satisfies* (1.2) *in the sense that $u(\cdot,t) \to u_0(\cdot)$ in $L^2(\Omega)$ as $t \to 0$, and the boundary condition* (1.8) *is satisfied in the sense that*

$$\lim_{h\to0} \frac{1}{h} \int_{\Omega_{h\pm}} \int_0^T \big(B(u)u_x - f(u)\big)\chi(t)\,dt\,dx = 0 \quad \text{for all } \chi \in L^2(0,T), \tag{1.16}$$

where $\Omega_{h+} := (1-h,1)$ and $\Omega_{h-} := (-1,-1+h)$. Moreover, $u(x,t) \in \Delta$ for $(x,t) \in \bar{Q}$.

1.1. Brief outline of the proofs of Theorems 1.1 and 1.2. We briefly outline the proofs of Theorems 1.1 and 1.2. Concerning Theorem 1.1, we first consider the approximate problems for $\varepsilon > 0$,

$$\partial_t u_i + \partial_x f_i(u) = \sum_{j=i}^n \partial_x\big(B_{ij}(u)\partial_x u_j\big) + \varepsilon\left(\frac{1}{n} - u_i\right), \quad i = 1,\dots,n, \tag{1.17}$$

$$u(x,0) = u_{0\varepsilon}(x) := (1-\varepsilon)u_0(x) + \frac{\varepsilon}{2n}e, \quad x \in \Omega, \quad e := (1,\dots,1)^{\mathrm{T}}, \tag{1.18}$$

$$f_i(u) - \sum_{j=i}^n B_{ij}(u)\big(\partial_x u_j \pm \theta(u_j - u_{\varepsilon j}^\pm(t))\big) = 0, \quad x = \pm1, \quad i = 1,\dots,n, \tag{1.19}$$

where we define

$$u_\varepsilon^\pm := (1-\varepsilon)u^\pm + \frac{\varepsilon}{2n}e.$$

The importance of introducing the perturbed problem (1.17)–(1.19) is related to the verification of the positive invariance of the region Δ. Indeed, the latter will follow from conditions (1.12) and (1.13) by proceeding as in [6] and [10]. The new point here is the argument on the boundary, which, as in [10], evaluates the sign of the space derivative given by the boundary condition (1.19) by using (1.12) and (1.13). Once we prove the invariance of Δ, for the subsequent analysis in the time stepping procedure we may assume $f(u)$ and $B(u)$ to be defined conveniently outside Δ in order to guarantee the a priori uniform boundedness of the solutions of the step problems; e.g., we may assume $f(u) = 0$ and $B(u) = \nu\,\mathrm{Id}$ outside a bounded open set containing Δ, where Id is the identity matrix.

Because of the nonlinear coupling in the boundary condition (1.19), the proof of Theorem 1.1 requires a *time stepping procedure*. This consists of linearizing the boundary condition in each step and thus obtaining the solution for each of the resulting problems. The initial-boundary value problem in the kth step with initial time $t = k\Delta t$, for $k \geq 1$, is given by (1.17) plus the value of the solution of the $(k-1)$th step as initial condition, i.e.,

$$u(x, k\Delta t) = u^{[k-1]}(x, k\Delta t - 0), \qquad x \in \Omega, \tag{1.20}$$

and the boundary condition

$$\partial_x u \pm \theta\big(u - u_\varepsilon^\pm(t)\big) = B\big(u^{[k-1]}\big)^{-1} f\big(u^{[k-1]}\big)\big|_{x=\pm 1,\ t=k\Delta t-0},\ t > k\Delta t, \tag{1.21}$$

where by $u^{[k]}$ we denote the solution of the kth step problem. Let us call the problem just described the kth step problem (1.17), (1.20), (1.21). For $k = 0$, we may define the 0th step problem (1.17), (1.20), (1.21) as above, agreeing to define $u^{[-1]}(x, 0) = u_0(x)$. Since the initial condition (1.20) and the boundary condition (1.21) may not satisfy the appropriate compatibility conditions, it is not possible to guarantee the well-posedness of (1.17), (1.20), (1.21). Nevertheless, a solution can always be obtained as the limit of solutions of well-posed problems obtained through a small regularization of those initial and boundary conditions enforcing the compatibility conditions. The well-posedness of the regularized kth step problem (1.17), (1.20), (1.21) is proved exactly in the same way as that of the similar problem in [10]. We recall that [10] follows the basic strategy of [13], whose main point is the application of the Leray–Schauder fixed point theorem, combined with bottom to top recursive a priori estimates similar to the procedure in [1]. Actually, the most important point in [10] is related to the regularity at the boundary. In general, in order to prove such regularity (see, e.g., [13]), we need to assume a certain smoothness of the coefficients of the second order terms with respect to one of the independent variables, x, t, which is not possible when dealing with system (1.17) since, although we assume that $B(u)$ is upper triangular, we *do not* assume that the B_{ij} depend only on u_k for $k \geq i$; rather, they can depend on all components of u. To surmount this difficulty, we use the idea introduced in [10], which is to use the good sign of the boundary term resulting from integration by parts in the crucial estimates.

We then define $u^\Delta(x, t)$ in Q by assuming $T = N\Delta t$ and setting $u^\Delta(x, t) := u^{[k]}(x, t)$ for $(x, t) \in \Omega \times [k\Delta t, (k+1)\Delta t)$, $k = 0, 1, \ldots, N$. Looking carefully at

the a priori estimates obtained for the kth step problem (1.17), (1.20), (1.21), $k = 0, 1, \ldots, N$, we show that it is possible to combine the first three, which imply that $u^\Delta \in H^{\beta,\beta/2}(\bar{Q})$ for some $\beta \in (0,1)$ with the Hölder norm bounded by a constant independent of ε and Δt, and that for some $c > 0$, also independent of ε and Δt,

$$\int_Q \left(\left| u^\Delta_{xx} \right|^2 + \left| u^\Delta_x \right|^4 + \left| u^\Delta_t \right|^2 \right) dx\, dt \leq c. \tag{1.22}$$

The uniform boundedness of the Hölder norm of u^Δ in $H^{\beta,\beta/2}(\bar{Q})$ provides the compactness of the sequence $\{u^\Delta\}$ in $H^{\alpha,\alpha/2}(\bar{Q})$ for any $0 < \alpha < \beta$, and so we may make $\Delta t \to 0$ and extract a subsequence $u^{\Delta'}$ converging to a certain u^ε in $H^{\alpha,\alpha/2}(\bar{Q})$ for each $\varepsilon > 0$. The limit u^ε will then satisfy an integral identity which is a weak formulation for problem (1.17)–(1.19). Using this integral identity and the estimate obtained from (1.22) we show that u^ε is in fact the unique smooth solution of (1.17)–(1.19).

Finally, we make $\varepsilon \to 0$ using the relative compactness guaranteed by the regularity of u^ε, which is uniform with respect to ε. It then follows that any vector function which is the limit of a converging subsequence of solutions of problem (1.17)–(1.19) is a classical solution of problem (1.1)–(1.3). Uniqueness follows in a standard way by taking the difference of the equations for any two solutions, multiplying each equation resulting from the difference of the corresponding components by the difference of these components, integrating over Ω, using integration by parts, and making estimates in a bottom-to-top iterative way.

As for Theorem 1.2, we obtain its proof by compactness, taking the limit as $\theta \to 0$ of a subsequence of solutions of (1.1)–(1.3) and proving that the limit function verifies the required properties. As for the uniqueness, no method for proving it is yet known for the class of regularity to which the solution obtained in this limit process belongs, in which we miss boundedness of the first space derivative up to the boundary.

Correlated works besides [9, 10] include the following. In the above mentioned book [13], the theory developed for linear and quasi-linear equations is applied to quasi-linear systems of type (1.1) (in the multidimensional case) where the nonlinear diffusion matrix is a scalar multiple of the identity matrix. We also mention the results of Amann [1] which require $f_i = f_i(u_i, \ldots, u_n)$ and $B_{ij} = B_{ij}(u_i, \ldots, u_n)$, $1 \leq i \leq j \leq n$, besides (1.6) and the upper-triangularity of $B(u)$, and which assume uniform boundedness of a certain Hölder norm of the local solution in order to extend the local solution to all times $t > 0$.

1.2. Brief description of the contents. This paper is organized as follows. In section 2, we prove positive invariance of Δ for problem (1.17)–(1.19). In section 3, we describe our time stepping procedure and obtain a priori estimates in the space of Hölder continuous functions for the solution of each time step problem. The Leray–Schauder fixed point theorem is applied in section 4 in order to prove existence of solutions for the time step problems. We also discuss the independence of some crucial estimates with respect to the time step Δt and the parameter enforcing the compatibility conditions at each time step. In section 5, we conclude the proof of Theorem 1.1 and also give the proof of Theorem 1.2. Finally, in section 6, we derive a model for polydisperse suspensions that improves an earlier model considered in [3], which can be cast, by a change of dependent variables, into a form which satisfies, after the addition of artificial viscosity, the assumptions of Theorem 1.1.

2. Invariance of the physical domain Δ. In this section we prove the positive invariance of Δ under the flow of the solution operator for problem (1.17)–(1.19). More precisely, we have the following result.

LEMMA 2.1. *Let the hypotheses of Theorem* 1.1 *hold. Then, if* $u(x,t)$ *is a classical solution of problem* (1.17)–(1.19) *in* $\bar{Q}$, *we have* $u(x,t) \in \Delta$ *for all* $(x,t) \in \bar{Q}$.

Proof. We argue by contradiction. Thus, assume that there is $(x,t) \in \bar{Q}$ such that $u(x,t) \notin \Delta$. Let

$$t_0 := \inf\{\, t \in [0,T] \,:\, u(x,t) \notin \Delta\}.$$

Since $u_0 \in \operatorname{int}\Delta$, we have $t_0 > 0$. Let $x_0 \in [-1,1]$ be such that $u(x_0,t_0) \in \partial\Delta$. Hence, we have $u_i(x_0,t_0) = 0$, for some $i = 1,\dots,n+1$, where we define $u_{n+1} := 1 - u_1 - \cdots - u_n$. We discuss separately two different cases: (i) $x_0 \in (-1,1)$; (ii) $x_0 = \pm 1$. Let us consider first case (i). In this case, for some $i \in \{1,\dots,n+1\}$, u_i assumes its minimum value in the region $[-1,1] \times [0,t_0]$ at (x_0,t_0), and so we must have

$$u_{it}(x_0,t_0) \le 0, \qquad u_{ix}(x_0,t_0) = 0, \qquad u_{ixx}(x_0,t_0) \ge 0. \tag{2.1}$$

Now, in light of (1.12) and (1.13) we may easily verify that at (x_0,t_0), the ith equation from (1.17) reduces to

$$u_{it} + \lambda_i u_{ix} = (\mu_i u_{ix})_x + \varepsilon l_i(u_i), \tag{2.2}$$

where λ_i and μ_i are scalar functions of u, $\mu_i > 0$ by (1.6), $l_i(u_i) := (\frac{1}{n} - u_i)$ if $i \in \{1,\dots,n\}$, $l_{n+1}(u_{n+1}) = u_{n+1}$, and the equation for u_{n+1} is obtained from the n equations of (1.17) in an obvious manner by summation. Hence, by (2.1) we arrive at a contradiction since the left-hand side of (2.2) is nonpositive, while the right-hand side is positive. We now consider case (ii) and assume without loss of generality that $x_0 = -1$. Let $i \in \{1,\dots,n+1\}$ be such that $u_i(-1,t_0) = 0$. Applying (1.12) and (1.13) to (1.19), we obtain that

$$u_{ix}(-1,t_0) < 0, \tag{2.3}$$

where, again, the equation for $i = n+1$ is obtained from (1.19) in an obvious manner. Now, (2.3) gives a contradiction since $u_i(x,t_0) \ge 0$ for all $x \in [-1,1]$. For $x_0 = 1$ we obtain a contradiction in a similar way. ∎

In view of the above result, we may assume $f_i(u)$ and $B_{ij}(u)$ to be defined as smooth functions for all $u \in \mathbb{R}^n$ in such a way that $f_i(u) \equiv 0$ and $B_{ij}(u) = \nu I$ for $u \notin U$, where $U \subseteq \mathbb{R}^n$ is a bounded open cube with faces parallel to the coordinate axes satisfying $\Delta \subseteq U$. This extension will be needed in what follows, since the invariance of Δ will not hold, in general, for the approximate solutions constructed by our time stepping procedure.

3. A time stepping approximation.

3.1. Initial-boundary value problem on short time intervals. In this section we construct approximate solutions of (1.17)–(1.19), defining them in a recursive way for time intervals $[k\Delta t,(k+1)\Delta t)$, $k = 0,\dots,N-1$, with $T = N\Delta t$. Our one-parameter approximate solution $u^{\Delta}(x,t)$ is a solution of the initial-boundary value problems, which are defined recursively for $(x,t) \in Q_k := [-1,1] \times [k\Delta t,(k+1)\Delta t)$

and consist of (1.17) with initial and boundary conditions given by

$$u^{\Delta}(x,k\Delta t)=u^{\Delta}(x,k\Delta t-0),\quad x\in\Omega, \tag{3.1}$$

$$\partial_x u\pm\theta\big(u-u_{\varepsilon}^{\pm}(t)\big)=\big[B(u^{\Delta}(k\Delta t-0))^{-1}f(u^{\Delta}(k\Delta t-0))\big]_{x=\pm1},\quad t>k\Delta t, \tag{3.2}$$

for $k=1,\dots,N-1$, while, for $k=0$, we set

$$u^{\Delta}(x,0)=u_{0\varepsilon}(x),\quad x\in\Omega, \tag{3.3}$$

$$\partial_x u\pm\theta\big(u-u_{\varepsilon}^{\pm}(t)\big)=B(u_{0\varepsilon}(\pm1))^{-1}f(u_{0\varepsilon}(\pm1)),\quad t>0. \tag{3.4}$$

We remark that the solutions of (1.17), (3.1), (3.2) will in general not be smooth at the corner points $(\pm1,k\Delta t)$, $k=1,\dots,N-1$, since the initial and boundary conditions at each time step, for $t=k\Delta t$, may not be compatible. Nevertheless, we will show that they can be defined in such a way that they are Hölder continuous in $\bar{Q}$, smooth in $\bar{Q}\setminus\{(\pm1,k\Delta t)\,:\,k=1,\dots,N-1\}$, and that (1.1) is satisfied in the classical sense in the interior of Q, while (3.1) and (3.2) are satisfied pointwise except at those corner points. The corresponding problem for $k=0$, that is, (1.17), (3.3), (3.4), is well-posed since here the compatibility conditions at the points $(\pm1,0)$ hold by assumption.

3.2. The kth step problem. Here we discuss the solution of the kth step problem (1.17), (3.1), (3.2), $k\geq1$. The case $k=0$ is treated similarly but does not require regularization of the initial data to enforce compatibility, since this already holds by assumption. As already mentioned, the solution of (1.17), (3.1), (3.2) is obtained as a limit of solutions of regularized problems given by (1.17) with the same boundary condition (3.2) and an initial condition obtained through a regularization of (3.1) in such a way that the new initial condition is compatible with (3.2) at the corners $(\pm1,k\Delta t)$. Thus, let $\zeta_\delta\in C_0(\Omega)$ be such that ζ_δ is even, $\zeta_\delta(x)=1$ for $-1+\delta\leq x\leq1-\delta$, $\zeta_\delta(x)=0$ for $1-\frac{\delta}{3}<|x|\leq1$, and ζ_δ is nonincreasing in $[0,1]$ with $|\zeta_\delta'(x)|\leq2\delta^{-1}$. We may define regularized initial data by

$$u_{\delta}^{\Delta}(x,k\Delta t)=\zeta_\delta(x)u^{\Delta}(x,k\Delta t-0)+\big(1-\zeta_\delta(x)\big)\big(\gamma_{k,\delta}^{-}\,\chi_{[-1,0)}(x)+\gamma_{k,\delta}^{+}\,\chi_{[0,1]}(x)\big), \tag{3.5}$$

where, as usual, χ_A denotes the indicator function of the set A and

$$\gamma_{k,\delta}^{\pm}=u_{\varepsilon}^{\pm}(k\Delta t)\pm\frac{1}{\theta}\big[B\big(u^{\Delta}(x,k\Delta t-0)\big)^{-1}f\big(u^{\Delta}(x,k\Delta t-0)\big)\big]_{x=\pm1}. \tag{3.6}$$

The well-posedness of the regularized kth problem (1.17), (3.2), (3.5) is proved in [10] for the case when $n=2$. The same result for general n is a straightforward consequence for the case when $n=2$. From the results in [10], we easily deduce that the solution of (1.17), (3.2), (3.5) is bounded in $H^{2+\beta,1+\beta/2}(K)$, where K is any compact set contained in $\bar{Q}_k\setminus\{(-1,k\Delta t),(1,k\Delta t)\}$ and $Q_k:=\Omega\times(k\Delta t,k\Delta t+T)$, with $\beta\in(0,1)$ and the corresponding bound depending, in general, on f,B,T, and K, but independent of δ, for δ sufficiently small. The latter is clear since the regularization in (3.5) is visible only for points in a small neighborhood of the corner points $(\pm1,k\Delta t)$ which will not intersect K if δ is small enough.

Using well-known compactness properties of spaces of Hölder continuous functions we easily obtain a solution of (1.17), (3.1), (3.2) as the limit of a subsequence of the solutions to (1.17), (3.2), (3.5). The solution of (1.17), (3.1), (3.2) obtained in this way is not known to belong to a class of well-posedness with respect to this problem

since we do not know whether its space derivative is bounded in the whole domain Q_k. On closer inspection we choose one such solution, and call it $u^{[k]}$.

The following estimate for the solution of the kth step problem (1.17), (3.1), (3.2) is an immediate consequence of the assumptions on f and B stated in section 2.

LEMMA 3.1. *The estimate*

$$\left\|u^{[k]}\right\|_{L^\infty(Q_k)} \le c \tag{3.7}$$

holds for some constant $c = c(\theta) > 0$, depending on θ but independent of ε, Δt, and k.

Proof. By assumption we have $f(u) = 0$ and $B(u) = \nu I$ if $u \notin U$, where U is a bounded open cube centered at the origin and containing Δ. Let $\tilde{U} \supset U$ be another bounded open cube with faces parallel to the coordinate and such that

$$\inf_{\substack{u\in\tilde{F}\\ t\in(0,T)}} |u - u^{\pm}(t)| > \frac{1}{\theta} \sup_{u\in U} |B(u)^{-1} f(u)| \tag{3.8}$$

for any face $\tilde{F}$ of $\tilde{U}$. The solution of the k-regularized problem (1.17), (3.2), (3.5) satisfies (3.7) by reasoning similar to that of the proof of Lemma 2.1, with a constant c determined only by $\tilde{U}$, and so depending on θ, because of (3.8), but independent of ε,Δt, k, and δ. Hence, (3.7) also holds for $u^{[k]}$ since it is a pointwise limit of solutions to (1.17), (3.2), (3.5) as $\delta \to 0$. ∎

We collect in the following lemma the a priori integral estimates for the solution of (1.17), (3.1), (3.2) whose bounds may be taken independent of ε, Δt, and k.

LEMMA 3.2. *The solution of* (1.17), (3.1), (3.2), *$u^{[k]}$ satisfies*

$$\int_\Omega \left|u^{[k]}(t)\right|^2 dx \le c, \qquad \int_{k\Delta t}^{k\Delta t+\tau} \int_\Omega \left|u_x^{[k]}\right|^2 dx\, dt \le c\tau, \quad 0 < \tau \le T. \tag{3.9}$$

Moreover, there exists $\beta \in (0,1)$, independent of ε, Δt, and k, such that $u^{[k]} \in H^{\beta,\beta/2}(\bar{Q}_k)$, with $Q_k = \Omega \times (k\Delta t, k\Delta t + T)$, and

$$\begin{aligned}\int_\Omega \left|u_x^{[k]}(k\Delta t+\tau)\right|^2 dx + c\int_{k\Delta t}^{k\Delta t+\tau} \int_\Omega \left(\left|u_{xx}^{[k]}\right|^2 + \left|u_x^{[k]}\right|^4 + \left|u_t^{[k]}\right|^2\right) dx\, dt \\ \le c'\left(\tau + \int_{k\Delta t}^{k\Delta t+\tau} \int_\Omega \left|u_x^{[k]}\right|^2 dx\, dt\right) + \int_\Omega \left|u_x^{[k]}(k\Delta t)\right|^2 dx.\end{aligned} \tag{3.10}$$

The constants c and c' above are independent of $\varepsilon, \Delta t, k$.

Proof. It suffices to consider the case when $n = 2$. In this case, the lemma follows from Lemmas 2.2, 2.3, 2.4, and 2.5 of [10]. ∎

As mentioned above, we define the approximate solution for the problem (1.17)–(1.19), $u^\Delta(x,t)$, by setting

$$\begin{aligned}&u^\Delta(x,t) = u^{[k]}(x,t)\\ &\text{for } (x,t) \in [-1,1] \times [k\Delta t, (k+1)\Delta t), \quad k = 0, 1, \ldots, N-1.\end{aligned} \tag{3.11}$$

As a corollary of Lemmas 3.1 and 3.2 we immediately obtain the following lemma.

LEMMA 3.3. *The following estimates hold for a constant $c > 0$ independent of ε and Δt:*

$$\|u^\Delta\|_{L^\infty(Q)} \le c, \tag{3.12}$$

$$\|u^\Delta\|_{H^{\beta,\beta/2}(\bar{Q})} \le c, \tag{3.13}$$

$$\int_Q \left|u_x^\Delta\right|^2 dx\, dt \le c, \tag{3.14}$$

$$\int_Q \left(\left|u_{xx}^\Delta\right|^2 + \left|u_x^\Delta\right|^4 + \left|u_t^\Delta\right|^2\right) dx\, dt \le c. \tag{3.15}$$

Moreover, for all $\varphi \in C_0(\mathbb{R} \times (-\infty, T))$ we have

$$\begin{aligned} &\int_Q \left(u^\Delta \varphi_t + f(u^\Delta)\varphi_x - B(u^\Delta)u_x^\Delta \varphi_x\right) dx\, dt + \int_\Omega u_{0\varepsilon}(x)\varphi(x,0)\, dx \\ &= \pm \int_{\{\pm 1\}\times(0,T)} \Big(\left(f(u^\Delta([t/\Delta t]\,\Delta t)) - f(u^\Delta)\right) \\ &\qquad\qquad - \left(B(u^\Delta([t/\Delta t]\,\Delta t)) - B(u^\Delta)\right) u_x^\Delta \Big) \varphi(\pm 1, t)\, dt \\ &\quad \pm \theta \int_{\{\pm 1\}\times(0,T)} B(u^\Delta)(u^\Delta - u_\varepsilon^\pm(t))\varphi(\pm 1, t)\, dt. \end{aligned} \tag{3.16}$$

Proof. The first two estimates follow directly from Lemmas 3.1 and 3.2, while the last two also follow from these lemmas by summing up from $k = 0$ to $k = N - 1$ the inequalities (3.9) and (3.10), respectively, with $\tau = \Delta t$. As for (3.16), we first observe that u^Δ satisfies (1.17) in Q, the boundary condition (3.2), for $t \in [k\Delta t, (k+1)\Delta t)$, $k = 0, 1, \ldots, N-1$, and the initial condition (3.4), for $t = 0$. Hence, (3.16) follows by multiplying (1.17) by φ, integrating over Q, and using integration by parts. ∎

4. Convergence of the approximate solutions. In this section we discuss the convergence of the approximate solutions $u^\Delta(x,t)$. From (3.13) in Lemma 3.3 we see that the family $\{u^\Delta\}$ is compact in $H^{\alpha,\alpha/2}(\bar{Q})$ for any $0 < \alpha < \beta$. Hence, we may extract a subsequence $u^{\Delta'}$ converging to some u^ε in $H^{\alpha,\alpha/2}(\bar{Q})$ as $\Delta t \to 0$. Also from Lemma 3.3 we deduce that u^ε satisfies

$$\|u^\varepsilon\|_{L^\infty(Q)} \le c, \tag{4.1}$$

$$\|u^\varepsilon\|_{H^{\beta,\beta/2}(\bar{Q})} \le c, \tag{4.2}$$

$$\int_Q |u_x^\varepsilon|^2\, dx\, dt \le c, \tag{4.3}$$

$$\int_Q \left(|u_{xx}^\varepsilon|^2 + |u_x^\varepsilon|^4 + |u_t^\varepsilon|^2\right) dx\, dt \le c, \tag{4.4}$$

and, for any $\varphi \in C_0(\mathbb{R} \times (-\infty, T))$,

$$\begin{aligned} &\int_Q \left(u^\varepsilon \varphi_t + f(u^\varepsilon)\varphi_x - B(u^\varepsilon)u_x^\varepsilon \varphi_x\right) dx\, dt + \int_\Omega u_{0\varepsilon}(x)\varphi(x,0)\, dx \\ &= \pm\theta \int_{\{\pm 1\}\times(0,T)} B(u^\varepsilon)(u^\varepsilon - u_\varepsilon^\pm(t))\varphi(\pm 1, t)\, dt. \end{aligned} \tag{4.5}$$

5. Conclusion of the proofs of Theorems 1.1 and 1.2.

Conclusion of the proof of Theorem 1.1. Estimates (4.1)–(4.5) allow for the application of Lemmas 2.7, 2.8, and 2.9 of [10] (these lemmas are stated for $n = 2$, but they are obviously also valid for general n) to conclude that

$$u_x^\varepsilon \in H^{\alpha,\alpha/2}(\bar{Q}) \quad \text{and} \quad |u_x^\varepsilon|^{(\alpha)} \le c \tag{5.1}$$

for some $\alpha \in (0,1)$ and $c > 0$ independent of ε. Hence, u^ε is a classical solution of (1.17)–(1.19), and by Lemma 2.1,

$$u^\varepsilon(x,t) \in \Delta \quad \text{for all } (x,t) \in \bar{Q}. \tag{5.2}$$

Making $\varepsilon \to 0$, using well-known compactness arguments in spaces of Hölder continuous functions, we may extract a subsequence $u^{\varepsilon'}$ converging in $H^{\gamma,\gamma/2}(\bar{Q})$, $0 < \gamma < \alpha$, to some u satisfying $u, u_x \in H^{\alpha,\alpha/2}(\bar{Q})$ and $u(x,t) \in \bar{Q}$ for all $(x,t) \in \bar{Q}$, which is a classical solution of (1.1), (1.2), (1.3). Clearly, we have

$$u(x,t) \in \Delta \quad \text{for all } (x,t) \in \bar{Q}. \tag{5.3}$$

Uniqueness of u in the class of such solutions is proved by the standard arguments already mentioned in section 1. □

Conclusion of the proof of Theorem 1.2. Let u^θ denote the solution of (1.1), (1.2), (1.3) obtained above. Given any compact subset $K \subseteq Q$ we have

$$u^\theta \in H^{2+\alpha,1+\alpha/2}(K) \quad \text{and} \quad \|u^\theta\|_{H^{2+\alpha,1+\alpha/2}(K)} \le c \tag{5.4}$$

for some $\alpha \in (0,1)$ and $c = c(K) > 0$ independent of θ. Moreover, we also have that u^θ satisfies (5.3) and

$$\|u_x^\theta\|_{L^2(Q)} \le c \tag{5.5}$$

for some $c > 0$ also independent of θ. Hence, we may make $\theta \to 0$ and extract a subsequence $u^{\theta'}$ converging in $L^1(Q)$ to a certain function $u \in H^{2+\alpha,1+\alpha/2}(Q)$, which satisfies (1.1) in Q. By (5.5) we may choose $u^{\theta'}$ so that $u_x^{\theta'} \rightharpoonup u_x$ in the weak topology of $L^2(Q)$. It satisfies

$$\int_Q \big(u\varphi_t + f(u)\varphi_x - B(u)u_x\varphi_x\big)\,dx\,dt + \int_\Omega u_0(x)\varphi(x,0)\,dx = 0 \tag{5.6}$$

for all $\varphi \in C_0^\infty(\mathbb{R} \times (-\infty, T))$, which follows from (4.5) by making first $\varepsilon \to 0$ and then $\theta \to 0$, and using the above mentioned convergences. Now, (1.16) follows easily from (5.6) by choosing $\varphi = \chi(t)\zeta_h(x)$, with $\chi \in C_0^1((0,T))$, $0 < h < 1/2$, and $\zeta_h(x) = 1$ for $x \in [-1+h, 1-h]$, $\zeta_h(x) = h^{-1}(1+x)$ for $x \in [-1, -1+h)$, and $\zeta_h(x) = h^{-1}(1-x)$ for $x \in (1-h, 1]$, and making $h \to 0$. Finally, it also satisfies (1.2) in the sense that

$$\lim_{t\to 0} \int_\Omega \big|u(x,t) - u_0(x)\big|^2\,dx = 0. \tag{5.7}$$

Indeed, from (5.6) we easily deduce

$$\lim_{t\to 0} \int_\Omega \big(u(x,t) - u_0(x)\big)\zeta(x)\,dx = 0 \tag{5.8}$$

for all $\zeta \in C_0^1(\mathbb{R})$. On the other hand, also from (5.6) we deduce the energy estimate in a standard way (see, e.g., [13, Chapter III, section 2]),

$$\int_\Omega |u(x,t)|^2\,dx + \frac{\nu}{2}\int_0^t\int_\Omega |u_x(x,s)|^2\,dx\,ds \le \int_\Omega |u_0(x)|^2 + c\int_0^t\int_\Omega |u(x,s)|^2\,dx\,ds,$$

from which it follows that

$$\limsup_{t\to 0}\int_\Omega |u(x,t)|^2\,dx \le \int_\Omega |u_0(x)|^2\,dx. \tag{5.9}$$

On the other hand, from (5.8) and convexity (see, e.g., [17]), we deduce

$$\liminf_{t\to 0}\int_\Omega |u(x,t)|^2\,dx \ge \int_\Omega |u_0(x)|^2\,dx. \tag{5.10}$$

Now, (5.8), (5.9), and (5.10) imply, as expected, (5.7). ∎

6. Sedimentation of polydisperse suspensions.

6.1. Balance equations and model assumptions. Our interest in the initial-boundary value problem (1.1), (1.2), (1.8) is motivated by a model of gravity sedimentation of polydisperse suspensions that form compressible sediments. After a change of variables, which ensures an upper triangular form, this model provides the coefficient functions $f_1(u),\dots,f_n(u)$ and $(B_{ij}(u))_{i,j=1,\dots,n}$. It is a variant of the model introduced in [3] and will be briefly derived here.

We consider small solid particles belonging to n different species having sizes $d_1 \ge d_2 \ge \cdots \ge d_n$ and densities $\varrho_1,\dots,\varrho_n$, where $d_i \ne d_j$ or $\varrho_i \ne \varrho_j$ for $i \ne j$, $1 \le i,j \le n$, which are dispersed in a viscous fluid of density $\varrho_{\rm f}$ and viscosity $\mu_{\rm f}$. We consider the solids and the fluid as $n+1$ superimposed continuous phases, and start from the multidimensional mass and linear momentum balances

$$\partial_t\phi_i + \nabla\cdot(\phi_i\mathbf{v}_i) = 0, \quad i=1,\dots,n; \quad \partial_t\phi - \nabla\cdot\big((1-\phi)\mathbf{v}_{\rm f}\big) = 0, \tag{6.1}$$

where $\Phi := (\phi_1,\dots,\phi_n)$ is the vector of the local volume fractions of the solid phases, $\phi := \phi_1 + \cdots + \phi_n$ is the total solids volume fraction, and $\mathbf{v}_1,\dots,\mathbf{v}_n$ and $\mathbf{v}_{\rm f}$ are the solids and fluid phase velocities, respectively. Summation over all mass balances implies $\nabla\cdot\mathbf{q} = 0$, where $\mathbf{q} := \phi_1\mathbf{v}_1 + \cdots + \phi_n\mathbf{v}_n + (1-\phi)\mathbf{v}_{\rm f}$ is the volume average mixture velocity. Introducing the relative velocities (or slip velocities) $\mathbf{u}_i := \mathbf{v}_i - \mathbf{v}_{\rm f}$ for $i=1,\dots,n$, we obtain $\mathbf{q} = \phi_1\mathbf{u}_1 + \cdots + \phi_n\mathbf{u}_n + \mathbf{v}_{\rm f}$ and

$$\mathbf{v}_i = \mathbf{u}_i + \mathbf{v}_{\rm f} = \mathbf{u}_i + \mathbf{q} - (\phi_1\mathbf{u}_1 + \cdots + \phi_n\mathbf{u}_n), \quad i=1,\dots,n.$$

The momentum balance equations for the n solid species and the fluid are

$$\varrho_i\phi_i(\partial_t\mathbf{v}_i + (\mathbf{v}_i\cdot\nabla)\mathbf{v}_i) = \nabla\cdot\mathbf{T}_i - \varrho_i\phi_i g\mathbf{k} + \mathbf{m}_i, \quad i=1,\dots,n, \tag{6.2}$$

$$\varrho_{\rm f}(1-\phi)(\partial_t\mathbf{v}_{\rm f} + (\mathbf{v}_{\rm f}\cdot\nabla)\mathbf{v}_{\rm f}) = \nabla\cdot\mathbf{T}_{\rm f} - \varrho_{\rm f}(1-\phi)g\mathbf{k} + \mathbf{m}_{\rm f}, \tag{6.3}$$

where $\mathbf{m}_i$ is the interaction force per unit volume between solid species i and the fluid and $\mathbf{m}_{\rm f} = -(\mathbf{m}_1+\cdots+\mathbf{m}_n)$. Interaction forces due to solid-solid contacts are assumed to be negligible compared with the solid-fluid transfer of momentum [5]. Moreover, $\mathbf{T}_i$ denotes the stress tensor of solids phase i, $i=1,\dots,n$, $\mathbf{T}_{\rm f}$ that of the fluid, g is the acceleration of gravity, and $\mathbf{k}$ is the upwards directed unit vector.

We assume that the stress tensors of the solid and fluid phases take the respective forms $\mathbf{T}_i = -p_i\mathbf{I} + \mathbf{T}_i^{\mathrm{E}}$, $i = 1, \ldots, n$, and $\mathbf{T}_{\mathrm{f}} = -p_{\mathrm{f}}\mathbf{I} + \mathbf{T}_{\mathrm{f}}^{\mathrm{E}}$, where p_i denotes the phase pressure of particle species i, p_{f} that of the fluid, $\mathbf{I}$ denotes the identity tensor, and $\mathbf{T}_i^{\mathrm{E}}$ and $\mathbf{T}_{\mathrm{f}}^{\mathrm{E}}$ are the extra (or viscous) stress tensors of particle species i and the fluid, respectively. Since viscous effects due to the motion of the mixture are not considered to be dominant for our analysis, we neglect the viscous stress tensors.

The theoretical phase pressures $p_1, \ldots, p_n$ and p_{f} are now expressed in terms of the pore pressure p and the effective solids stress σ_{e}, which can be measured. As in [3], we assume that σ_{e} is a known material-specific function of ϕ, which satisfies

$$\sigma_{\mathrm{e}}(\phi) \begin{cases} = 0 & \text{for } \phi \le \phi_{\mathrm{c}}, \\ > 0 & \text{for } \phi > \phi_{\mathrm{c}}, \end{cases} \qquad \sigma_{\mathrm{e}}'(\phi) := \frac{d\sigma_{\mathrm{e}}(\phi)}{d\phi} \begin{cases} = 0 & \text{for } \phi \le \phi_{\mathrm{c}}, \\ > 0 & \text{for } \phi > \phi_{\mathrm{c}}, \end{cases} \tag{6.4}$$

where ϕ_{c} is the critical concentration at which the solid particles touch each other.

In this paper, the relationship between the theoretical pressures $p_1, \ldots, p_n$ and p_{f} and the experimental variables p and $\sigma_{\mathrm{e}}(\phi)$ slightly differs from the one given in [3]. In that paper, it is argued that the portions of the total pressure $p_1 + \cdots + p_n + p_{\mathrm{f}} = \phi p + \sigma_{\mathrm{e}}(\phi)$ can be assigned to species i and to the fluid by

$$p_i = \phi_i\big(p + \sigma_{\mathrm{e}}(\phi)/\phi\big), \quad i = 1, \ldots, n; \quad p_{\mathrm{f}} = (1-\phi)p. \tag{6.5}$$

In this work, we do not fix $p_1, \ldots, p_n$ and p_{f} as volume quantities, as expressed by (6.5); rather, the gradient of the pressure of each phase is specified as a surface quantity. Thus, the gradients of the solid and fluid phase pressures introduced by the stress tensors in (6.2) and (6.3) are replaced by the respective expressions

$$\nabla p_i = \phi_i\big(\nabla p + \nabla\sigma_{\mathrm{e}}(\phi)/\phi\big) \quad \text{and} \quad \nabla p_{\mathrm{f}} = (1-\phi)\nabla p. \tag{6.6}$$

Our preference for (6.6) (instead of (6.5)) is in part motivated by the discussion of sediment diffusivity in section 7.5 of [3], where (6.5) is used. In fact, the numerical examples of [3] (see also [2]) show a strong differential relative movement of solid species within the sediment, which is driven by a diffusive term involving the gradient $\nabla(\phi_i/\phi)$. In particular, as is shown in [3], this term leads to a—probably unrealistic—equidistribution of the solid species within the sediment at steady state. The model analyzed herein is nearly the same as that of [3], but it does not include that particular term, and is expected to predict more realistic results.

We emphasize herein that a polydisperse sedimentation model based on (6.6) is supported by mathematical analysis, while a detailed quantitative (numerical) comparison with the predictions generated by using (6.5) is not within the scope of this paper. From an experimental point of view, both alternatives (6.5) and (6.6) are discussed in the literature. Both variants have been scrutinized [21] and considered in parallel [15], and (6.6) is postulated a priori, for example, in [16]. This ambiguity is still an unresolved issue, as is emphasized in [18].

Our analysis is valid only for the case of (6.6). In fact, based on (6.5), the final solid momentum balances would include the additional term $\nabla((\phi_i/\phi)\sigma_{\mathrm{e}}(\phi))$, which in turn later appears in the "Darcy-type law" expression for the slip velocity (6.10) as $-(\sigma_{\mathrm{e}}(\phi)/\phi_i)\nabla(\phi_i/\phi)$. Even in the diffusive flux term, the gradient of ϕ_i remains and precludes recasting the diffusion matrix into an upper-triangular matrix by a change of variables. This particular form is, however, essential for our treatment. The introduction of $\phi = \phi_1 + \cdots + \phi_n$ as a new variable requires an extra equation for ϕ, which can be generated by summing the equations for $\phi_1, \ldots, \phi_n$. If (6.5) is used,

then the resulting equation will involve all variables $\phi_1, \dots, \phi_n, \phi$, while starting from (6.6), the resulting equation depends on ϕ only.

Under the present assumptions, and in particular neglecting the viscous stress tensors, the momentum balances for the solid and fluid phases become

$$\begin{aligned} \phi_i \varrho_i \big(\partial_t \mathbf{v}_i + (\mathbf{v}_i \cdot \nabla)\mathbf{v}_i\big) &= \mathbf{m}_i - \varrho_i \phi_i g \mathbf{k} - \phi_i \nabla p - \frac{\phi_i}{\phi} \nabla \sigma_{\mathrm{e}}(\phi), \\ (1-\phi)\varrho_{\mathrm{f}} \big(\partial_t \mathbf{v}_{\mathrm{f}} + (\mathbf{v}_{\mathrm{f}} \cdot \nabla)\mathbf{v}_{\mathrm{f}}\big) &= \mathbf{m}_{\mathrm{f}} - \varrho_{\mathrm{f}}(1-\phi) g \mathbf{k} - (1-\phi)\nabla p. \end{aligned} \tag{6.7}$$

As a result of a dimensional analysis [3, 5] and due to the reduction to one space dimension, the left-hand sides in (6.7) can be neglected. After applying this simplification, summation of all equations in (6.7) yields the momentum balance of the suspension

$$0 = -\varrho(\Phi) g \mathbf{k} - \nabla p - \nabla \sigma_{\mathrm{e}}(\phi), \tag{6.8}$$

where $\varrho(\Phi) := \phi_1 \varrho_1 + \cdots + \phi_n \varrho_n + (1-\phi)\varrho_{\mathrm{f}}$ is the local density of the mixture. Next, we insert ∇p of the suspension momentum balance (6.8) into a solids momentum balance and assume the proportionality $\mathbf{m}_i = (\lambda_i(\Phi))^{-1}\mathbf{u}_i$ for $i = 1, \dots, n$, where the mobilities $\lambda_1(\Phi), \dots, \lambda_n(\Phi)$ are given by [3]

$$\lambda_i(\Phi) = \frac{d_i^2 \tilde{V}(\phi)}{18 \mu_{\mathrm{f}} \phi_i}, \quad i = 1, \dots, n, \tag{6.9}$$

where μ_{f} is the dynamic viscosity of the pure fluid and $\tilde{V}(\phi)$ is a hindered settling factor. This yields the Darcy-type law

$$\mathbf{u}_i = \lambda_i(\Phi)\phi_i \left(\big(\varrho_i - \varrho(\Phi)\big) g \mathbf{k} - \frac{1-\phi}{\phi} \sigma_{\mathrm{e}}'(\phi) \nabla \phi \right) \tag{6.10}$$

for $i = 1, \dots, n$. Wherever $\phi \le \phi_{\mathrm{c}}$, the solid effective stress $\sigma_{\mathrm{e}}'(\phi)$ vanishes, and (6.10) may be viewed as a generalization of Stokes' law, while for $\phi > \phi_{\mathrm{c}}$, it represents a version of Darcy's law. The hindered settling factor $\tilde{V}(\phi)$ is assumed to satisfy $\tilde{V}(0) = 1$ and $\tilde{V}(1) = 0$ and can, e.g., be chosen as [14] $\tilde{V}(\phi) = (1-\phi)^{N-2}$ for $0 < \phi \le 1$ with $N > 2$, and $\tilde{V}(\phi) = 0$ otherwise.

In one space dimension, the volume average velocity q vanishes for batch settling in a closed vessel. Then, the only equations that actually need to be solved are the continuity equations $\partial_t \phi_i + (\phi_i v_i)_x = 0$, $i = 1, \dots, n$, which now assume the form

$$\partial_t \phi_i + \partial_x \tilde{f}_i(\Phi) = \sum_{j=1}^{n} \partial_x \big(\tilde{B}_{ij}(\Phi) \partial_x \phi_j \big), \quad i = 1, \dots, n, \tag{6.11}$$

with the functions

$$\tilde{f}_i(\Phi) = \phi_i \big(d_i^2 \big(\varrho_i - \varrho(\Phi)\big) - \tilde{S}(\Phi) \big) V(\phi), \quad i = 1, \dots, n, \tag{6.12}$$

where, for convenience, we define

$$V(\phi) := \frac{\tilde{V}(\phi)}{18\mu_{\mathrm{f}}}, \quad \tilde{S}(\Phi) := \sum_{k=1}^{n} \phi_k d_k^2 \big(\varrho_k - \varrho(\Phi)\big)$$

and use the regularized diffusion matrix

$$\tilde{B}_{ij} := \delta_{ij}\nu + \hat{B}_i(\Phi), \quad i,j = 1,\dots,n, \tag{6.13}$$

where $\nu > 0$ is a viscous regularization parameter, and we define

$$\hat{B}_i(\Phi) := \phi_i \left(d_i^2 - \sum_{k=1}^{N} \phi_k d_k^2 \right) \left(\frac{1-\phi}{\phi} \sigma_{\mathrm{e}}'(\phi) \right) V(\phi), \quad i = 1,\dots,n. \tag{6.14}$$

Finally, we mention that a very similar model for polydisperse sedimentation, which equally gives rise to a parabolic system of type (6.11), was proposed and solved numerically in a recent paper by Watson, Barker, and Robins [20]. Summarizing the approaches of [3, 8, 20], one may view the initial-boundary value problem studied herein as a generic model for polydisperse sedimentation.

Remark. The diffusion constant $\nu > 0$ represents the hydrodynamic diffusion of the solids phases. This phenomenon is based on the observation that in reality, solid particles at a given concentration vector Φ do not settle at precisely the same velocity. A detailed account of this variability is given in section 7.4 of [3], but its main effect (as compared to nondiffusive models, e.g., those studied in [3, 4, 5]) is a blurring of otherwise sharp concentration discontinuities. This effect can be most easily modeled by the simple diffusion constant introduced here. In fact, Esipov [8] proposes a similar diffusive model that is approximated by constant diffusivities. The identification of nonlinear but positive diffusion functions accounting for hydrodynamic diffusion is, however, a topic of current research; see, e.g., [5, 7, 11, 12, 19] and the references cited in these papers.

6.2. Initial and boundary conditions. We consider batch settling of a suspension with a given initial composition $\Phi^0(x)$ in a closed cylindrical vessel, where the height of the suspension column has been normalized to 2 and $x \in [-1,1]$ is the corresponding downwards-decreasing depth variable. Thus, we prescribe the initial condition

$$\Phi(x,0) = \Phi^0(x) = \big(\phi_1^0(x),\dots,\phi_n^0(x)\big)^{\mathrm{T}} \in \tilde{\Delta}, \quad x \in [-1,1], \tag{6.15}$$

where

$$\tilde{\Delta} := \big\{ \Phi \in \mathbb{R}^n \,:\, \phi_1 \geq 0,\dots,\phi_n \geq 0,\ \phi_1 + \cdots + \phi_n \leq 1 \big\} \tag{6.16}$$

is the phase space of the original variables. At $x = -1$, the suspension surface is located and at $x = 1$ the vessel bottom is located. Both levels give rise to zero-flux boundary conditions, which means that all solids fluxes through $x = -1$ and $x = 1$ vanish:

$$(\phi_i v_i)|_{x=-1} = (\phi_i v_i)|_{x=1} = 0, \quad i = 1,\dots,n. \tag{6.17}$$

Equations (6.11) represent the solids continuity equations; therefore

$$\phi_i v_i = \tilde{f}_i(\Phi) - \sum_{m=1}^{n} \tilde{B}_{im}(\Phi)\partial_x \phi_m, \quad i = 1,\dots,n, \tag{6.18}$$

and the boundary conditions (6.17) can be written as the zero-flux or Neumann boundary conditions

$$\left(\tilde{f}_i(\Phi) - \sum_{m=1}^{n} \tilde{B}_{im}(\Phi)\partial_x \phi_m \right) \Big|_{x=\pm 1} = 0. \tag{6.19}$$

Thus, (6.11), (6.15), (6.19) is the relevant initial-boundary value problem in our application.

6.3. Transformation to upper-triangular form. The analysis of sections 1–5 applies to our model if we change variables such that the resulting diffusion matrix becomes upper triangular. To this end, we rewrite the model in terms of the new variables $u = (u_1, \dots, u_n)^{\mathrm{T}}$ defined by

$$u_1 := \phi_1, \dots, u_{n-1} := \phi_{n-1}, \quad u_n := 1 - \phi = 1 - (\phi_1 + \cdots + \phi_n). \tag{6.20}$$

The phase space Δ is then as defined in (1.9). With a slight abuse of notation, let us define

$$\varrho(u) = \varrho(\Phi) := u_1(\varrho_1 - \varrho_n) + \cdots + u_{n-1}(\varrho_{n-1} - \varrho_n) + (1 - u_n)\varrho_n + u_n\varrho_{\mathrm{f}}$$

and $S(u) = \tilde{S}(\Phi)$. Noting that the quantity $1 - \phi$ satisfies the equation

$$\begin{aligned} \partial_t(1-\phi) - \partial_x\left(\sum_{i=1}^n \tilde{f}_i(\Phi)\right) &= -\partial_x\left\{\nu\sum_{i=1}^n \partial_x\phi_i + \sum_{i=1}^n \hat{B}_i(\Phi)\left(\sum_{m=1}^n \partial_x\phi_m\right)\right\} \\ &= \nu\partial_x^2(1-\phi) + \partial_x\left(\left[\sum_{i=1}^n \hat{B}_i(\Phi)\right]\partial_x(1-\phi)\right), \end{aligned} \tag{6.21}$$

we then obtain, after the change of variables (6.20),

$$\begin{aligned} f_i(u) &= u_i\big(d_i^2(\varrho_i - \varrho(u)) - S(u)\big)V(1-u_n), \quad i = 1, \dots, n-1, \\ f_n(u) &= -u_n S(u) V(1-u_n), \end{aligned} \tag{6.22}$$

where

$$S(u) := \sum_{i=1}^{n-1} u_i d_i^2\big(\varrho_i - \varrho(u)\big) + \big(1 - (u_1 + \cdots + u_n)\big)d_n^2(\varrho_n - \varrho(u)\big).$$

We easily see that (1.12) is satisfied. Defining

$$d^2(u) := \sum_{i=1}^n \phi_i d_i^2 = \big(1 - (u_1 + \cdots + u_n)\big)d_n^2 + \sum_{i=1}^{n-1} u_i d_i^2,$$

in view of (6.21) and

$$\partial_t\phi_i + \partial_x\tilde{f}_i(\Phi) = \partial_x\left(\nu\partial_x\phi_i - \hat{B}_{in}(\Phi)\partial_x(1-\phi)\right), \quad i = 1, \dots, n-1, \tag{6.23}$$

we see that the new diffusion matrix $(B_{ij})_{1\le i,j\le n}$ is given by

$$\begin{aligned} B_{ii} &= \nu, \quad i = 1, \dots, n-1, \\ B_{in} &= -u_i(d_i^2 - d^2(u))\frac{u_n}{1-u_n}\sigma_{\mathrm{e}}'(1-u_n)V(1-u_n), \quad i = 1, \dots, n-1, \\ B_{nn} &= \nu + d^2(u)\frac{u_n^2}{1-u_n}\sigma_{\mathrm{e}}'(1-u_n)V(1-u_n), \end{aligned} \tag{6.24}$$

and $B_{ij} = 0$ otherwise. Note that the matrix has nonzero entries on its diagonal and in its last column only (as a special case of a triangular matrix). Satisfaction of conditions (1.13) can easily be verified.

The original problem is recast in the new variables, which means that we refer to system (1.1) with the flux vector f and the matrix B given by (6.22) and (6.24), respectively, and use the initial datum

$$u(x,0) = u^0(x) := \left(\phi_1^0(x), \dots, \phi_{n-1}^0(x), 1 - (\phi_1^0(x) + \cdots + \phi_n^0(x))\right)^{\mathrm{T}} \in \Delta, \quad x \in \overline{\Omega},$$

where $\phi_1^0(x), \dots, \phi_n^0(x)$ are the initial solids concentrations. Finally, it is easy to check from (6.19) and (6.21) that for the variables $u_1, \dots, u_n$, the boundary conditions are again the Neumann (zero-flux) conditions

$$\left(f_i(u) - \sum_{m=1}^{n} B_{im}(u) \partial_x u_m \right) \Bigg|_{x=\pm 1} = 0. \tag{6.25}$$

REFERENCES

[1] H. Amann, *Dynamic theory of quasi-linear parabolic systems* III. *Global existence*, Math. Z., 202 (1989), pp. 219–250.

[2] S. Berres and R. Bürger, *On gravity and centrifugal settling of polydisperse suspensions forming compressible sediments*, Internat. J. Solids Structures, 40 (2003), pp. 4965–4987.

[3] S. Berres, R. Bürger, K.H. Karlsen, and E.M. Tory, *Strongly degenerate parabolic-hyperbolic systems modeling polydisperse sedimentation with compression*, SIAM J. Appl. Math., 64 (2003), pp. 41–80.

[4] R. Bürger, K.-K. Fjelde, K. Höfler, and K.H. Karlsen, *Central difference solutions of the kinematic model of settling of polydisperse suspensions and three-dimensional particle-scale simulations*, J. Engrg. Math., 41 (2001), pp. 167–187.

[5] R. Bürger, K.H. Karlsen, E.M. Tory, and W.L. Wendland, *Model equations and instability regions for the sedimentation of polydisperse suspensions of spheres*, Z. Angew. Math. Mech., 82 (2002), pp. 699–722.

[6] K.N. Chueh, C.C. Conley, and J.A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[7] R.H. Davis, *Hydrodynamic diffusion of suspended particles: A symposium*, J. Fluid Mech., 310 (1996), pp. 325–335.

[8] S.E. Esipov, *Coupled Burgers equations: A model of polydispersive sedimentation*, Phys. Rev. E, 52 (1995), pp. 3711–3718.

[9] H. Frid and V. Shelukhin, *A quasi-linear parabolic system for three-phase capillary flow in porous media*, SIAM J. Math. Anal., 35 (2003), pp. 1029–1041.

[10] H. Frid and V. Shelukhin, *Initial boundary value problems for a quasi-linear parabolic system in three-phase capillary flow in porous media*, SIAM J. Math. Anal., 36 (2005), pp. 1407–1425.

[11] K. Höfler, *Simulation and Modeling of Mono- and Bidisperse Suspensions*, Doctoral Thesis, Institute of Computer Applications, University of Stuttgart, Stuttgart, Germany, 2000.

[12] A.J.C. Ladd, *Dynamical simulations of sedimenting spheres*, Phys. Fluids A, 5 (1993), pp. 299–310.

[13] O.A. Ladyzhenskaya, V.A. Solonnikov, and N.N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[14] J.F. Richardson and W.N. Zaki, *Sedimentation and fluidization: Part* I, Trans. Inst. Chem. Engrg. (London), 32 (1954), pp. 35–53.

[15] K. Rietema, *Science and technology of dispersed two-phase systems* I *and* II, Chem. Engrg. Sci., 37 (1982), pp. 1125–1150.

[16] Y.T. Shih, D. Gidaspow, and D.T. Wasan, *Hydrodynamics of sedimentation of multisized particles*, Powder Technol., 50 (1987), pp. 201–215.

[17] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Res. Notes Math. 4, R. J. Knops, ed., Pitman, New York, 1979, pp. 136–211.

[18] C. Tien, S.K. Teoh, and R.B.H. Tan, *Cake filtration analysis—the effect of the relationship between the pore liquid pressure and the cake compressive stress*, Chem. Engrg. Sci., 56 (2001), pp. 5361–5369.

[19] E.M. Tory, *Stochastic sedimentation and hydrodynamic diffusion*, Chem. Engrg. J., 80 (2000), pp. 81–89.
[20] A.D. Watson, G.C. Barker, and M.M. Robins, *Sedimentation in bidisperse and polydisperse colloids*, J. Colloid Interface Sci., 286 (2005), pp. 176–186.
[21] M.S. Willis, M. Shen, and K.J. Gray, *Investigations of the fundamental assumptions relating compression-permeability data with filtration*, Canad. J. Chem. Engrg., 52 (1974), pp. 331–337.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 574–594

LOCALIZED TIGHT FRAMES ON SPHERES*

F. J. NARCOWICH†, P. PETRUSHEV‡, AND J. D. WARD†

Abstract. In this paper we wish to present a new class of tight frames on the sphere. These frames have excellent pointwise localization and approximation properties. These properties are based on pointwise localization of kernels arising in the spectral calculus for certain self-adjoint operators, and on a positive-weight quadrature formula for the sphere that the authors have recently developed. Improved bounds on the weights in this formula are another by-product of our analysis.

Key words. tight frames, n-sphere, localization

AMS subject classifications. 42C10, 42C40, 65D32

DOI. 10.1137/040614359

1. Introduction.

Frames were introduced in the 1950s by Duffin and Schaeffer [4] to represent functions via over-complete sets. Let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$. In that case, a set $\{\psi_j\}_{j\in\mathcal{J}}$ is a *frame* if there are constants $c, C > 0$ such that for all $f \in \mathcal{H}$

$$c\|f\|^2 \le \sum_{j\in\mathcal{J}} |\langle f, \psi_j\rangle|^2 \le C\|f\|^2.$$

The smallest C and largest c are called upper and lower *frame bounds*. If $C = c$, we say the frame is *tight*. If $C = c = 1$, then the frame is *normalized*, and if in addition $\|\psi_j\| = 1$ for all j, then the frame is an orthonormal basis.

Frames, including tight ones, arise naturally in wavelet analysis on $\mathbb{R}^n$ when continuous wavelet transforms are discretized. They provide a redundancy that helps reduce the effect of noise in data, and they have been constructed, studied, and employed extensively in both theoretical and applied problems [1, 2, 6, 7, 10, 12].

Tight frames are similar in many respects to orthonormal wavelet bases; decomposing and synthesizing a signal or image from known data are tasks carried out with the same set of functions, the ones in the frame or in the basis. A feature that makes one frame preferable to another is simultaneous localization of the frame functions in both space and frequency. Frames with this feature have been successfully developed in $\mathbb{R}^n$ [1, 2].

On $\mathbb{S}^n$, the n-dimensional unit sphere in $\mathbb{R}^{n+1}$, various types of both wavelets and frames have been constructed and used; see [8, 13, 16, 21] for references and more discussion. Tight, well-localized frames are another matter.

The purpose of this paper is to construct and study a class of well-localized, computationally implementable, tight frames on $\mathbb{S}^n$. Central to this construction is a key result of this paper, Theorem 3.5. This result concerns pointwise localization for a

*Received by the editors September 2, 2004; accepted for publication (in revised form) January 24, 2006; published electronically June 9, 2006.
http://www.siam.org/journals/sima/38-2/61435.html

†Department of Mathematics, Texas A & M University, College Station, TX 77843-3368 (fnarc@math.tamu.edu, jward@math.tamu.edu). This research was supported by grants DMS-0204449 and DMS-0504353 from the National Science Foundation.

‡Department of Mathematics, University of South Carolina, Columbia, SC 29208 (pencho@math.sc.edu). This research was supported by grant DMS-0200665 from the National Science Foundation.

family of kernels for certain operators on $\mathbb{S}^n$; the family depends on a parameter and localization increases as the parameter becomes small. The frame functions, which are compactly supported in the frequency domain, are constructed from such kernels. This construction has an interesting connection to wavelet masks, which we will point out below. Another application of our localization result, one that is essential to turning the frame functions into a tight frame—that is, a hierarchical, multiresolution setting—is an improved positive-weight quadrature formula for $\mathbb{S}^n$, where the weights have known bounds. This quadrature formula is used for discretization purposes. In addition to Theorem 3.5, the main results of this paper are Proposition 5.1, Theorem 5.2, and Corollary 5.3. The first of these concerns the approximation power of the frames, the second shows that the frames are tight, and the third shows that the frame functions have excellent spatial localization.

The frame functions and quadrature formula are of interest in their own right. In particular, they can be used in the construction and characterization of many of the classical Banach spaces, including $L^p(\mathbb{S}^n)$, Besov spaces, and Triebel–Lizorkin spaces [18]. We mention also that the operator-theoretic approach that we use here may provide a foundation for extending our results to other Riemannian manifolds.

Strategy. The best way to view our method for constructing frames is to take an operator-theoretic approach. Let E_λ be the (right-continuous) spectral family for an unbounded, nonnegative, self-adjoint operator L defined on a Hilbert space $\mathcal{H}$. Thus, $\mathsf{L} = \int_{0^-}^{\infty} \lambda d\mathsf{E}_\lambda$. On the sphere $\mathbb{S}^n$, this will be related to the square root of the Laplace–Beltrami operator shifted by a constant. For now, that connection isn't required.

We wish to decompose the spectral family in a way reminiscent of the decomposition of frequency space used by Meyer [10, 12] in connection with the construction of his wavelets. For this, we need a function $a \in C(\mathbb{R})$, with support in $[\frac{1}{2}, 2]$, and satisfying $|a(t)|^2 + |a(2t)|^2 \equiv 1$ on $[\frac{1}{2}, 1]$. Such a function can be easily constructed out of an *orthogonal wavelet mask* m_0 [2, section 8.3]. In fact, if $m_0(\xi) \in C^{k+1}$, then $a(t) := m_0(\pi \log_2(t))$ on $[\frac{1}{2}, 2]$, and 0 otherwise, is a C^k function that satisfies the appropriate criteria.

Define $b \in C(\mathbb{R})$ by

$$b(t) := \begin{cases} 1, & t \le 1, \\ a(t)\overline{\tilde{a}(t)}, & t > 1. \end{cases} \tag{1}$$

Using the properties of a we see that $\sum_{j=-\infty}^{J} |a(t/2^j)|^2 = b(t/2^J)$ if $t > 0$ and is 0 if $t \le 0$. Integrating both sides above with respect to $d\mathsf{E}_\lambda$ and using the spectral calculus for L, we obtain $\sum_{j=-\infty}^{J} a(\mathsf{L}/2^j)a(\mathsf{L}/2^j)^* = b(\mathsf{L}/2^J) - \mathsf{E}_0$. Define the operators,

$$\mathsf{A}_j = a(\mathsf{L}/2^j), \tag{2}$$

$$\mathsf{B}_J := b(\mathsf{L}/2^J), \tag{3}$$

and note that the relationship derived above becomes $\sum_{j=-\infty}^{J} \mathsf{A}_j\mathsf{A}_j^* = \mathsf{B}_J - \mathsf{E}_0$. Finally, it is easy to show that the strong limit of B_J as $J \to \infty$ is I, the identity. Taking limits above then yields $\sum_{j=-\infty}^{\infty} \mathsf{A}_j\mathsf{A}_j^* = \mathsf{I} - \mathsf{E}_0$.

We now can use this identity to define *decomposition* and *reconstruction* operators for $f \in \mathcal{H}$, which are, respectively,

$$f \to w_j = \mathsf{A}_j^* f \text{ and } f = \mathsf{E}_0 f + \sum_{j=-\infty}^{\infty} \mathsf{A}_j w_j.$$

PROPOSITION 1.1. *For any $a \in C(\mathbb{R})$ satisfying the conditions above, the operator frame that we have constructed is tight in the sense that*

$$\|f\|^2 = \|\mathsf{E}_0 f\|^2 + \sum_{j=-\infty}^{\infty} \|\mathsf{A}_j^* f\|^2.$$

In addition, we have that $\langle w_j, w_{j'} \rangle = 0$ for $|j - j'| \geq 2$, where $w_j = \mathsf{A}_j^* f$.

Proof. This follows immediately from the decomposition and reconstruction formulas above, the properties of a and of the spectral family. ∎

Note that the decomposition arrived at above is *nearly* orthogonal. The level j decomposition w_j is *not* orthogonal to $w_{j\pm1}$, but it *is* orthogonal to the decomposition at all other levels.

As we pointed out above, when we deal with the sphere $\mathbb{S}^n$, we will take L proportional to $L_n := \sqrt{\lambda_n^2 - \Delta_{\mathbb{S}^n}}$, where $\lambda_n := \frac{n-1}{2}$. Notation and background pertinent to this operator, spherical harmonics, and related topics can be found in section 2.1.

In section 2.2, we show that with this choice of L the decomposition operator A_j^* is given in terms of a kernel $\overline{A_j}(\xi \cdot \eta)$, $\xi, \eta \in \mathbb{S}^n$, which is a polynomial in $\xi \cdot \eta$. Using the addition theorem for spherical harmonics, one can see that the level j decomposition $w_j(\eta) = \langle f(\xi), A_j(\xi \cdot \eta) \rangle_{\mathbb{S}^n}$ is a finite sum of spherical harmonics.

In the reconstruction phase, we need to find $\mathsf{A}_j w_j(\omega) = \langle w_j(\eta), \overline{A_j}(\omega \cdot \eta) \rangle_{\mathbb{S}^n}$. The integrand in this inner product is also a finite sum of spherical harmonics. At this point, the order of the spherical harmonics is such that we can compute the integral exactly using a quadrature formula introduced in [14, 15] and, in section 4, developed into the tool we need here. The point is that the frame functions have the form $\psi_{j,\xi}(\eta) = \sqrt{c_{j,\xi}} A_j(\eta \cdot \xi)$, where the $c_{j,\xi}$ and $\xi \in X_j$ are weights and nodes for the quadrature formula appropriate to level j. The details are given in section 5.

What makes these frame functions special is that they have excellent pointwise localization properties. These properties follow from the results on pointwise localization of certain kernels, given in section 3.

2. Near-orthogonal spectral decomposition for $\mathbb{S}^n$.

2.1. Background and notation for $\mathbb{S}^n$.

Centers and decompositions of $\mathbb{S}^n$. Let X be a finite set of distinct points in $\mathbb{S}^n$; we will call these the *centers*. There are several important quantities associated with this set: the *mesh norm*, $h_X = \sup_{y \in \mathbb{S}^n} \inf_{\xi \in X} d(\xi, y)$, where $d(\cdot, \cdot)$ is the geodesic distance between points on the sphere; the *separation radius*, $q_X = \frac{1}{2} \min_{\xi \neq \xi'} d(\xi, \xi')$; and the *mesh ratio*, $\rho_X := h_X / q_X \geq 1$.

For $\rho \geq 1$, let $\mathcal{F}_\rho = \mathcal{F}_\rho(\mathbb{S}^n)$ be the family of all sets of centers X with $\rho_X \leq \rho$; we will say that the family $\mathcal{F}_\rho$ is *ρ-uniform*. Unless confusion would arise, we will not indicate $\mathbb{S}^n$, and just use $\mathcal{F}_\rho$ to designate a family. The specific sphere $\mathbb{S}^n$ will be clear from the context. We will also say that a set of centers X is ρ-uniform if $X \in \mathcal{F}_\rho$. It is possible to show that for every $\rho \geq 2$ there exist nonempty ρ-uniform families for any $\mathbb{S}^n$ and that they contain sets of centers X for which h_X becomes arbitrarily small. The result is stated below. For a proof of the facts mentioned here as well as further discussion, see [19, section 2].

PROPOSITION 2.1 (see [19, Proposition 2.1]). *Let $\rho \geq 2$ and let $\mathcal{F}_\rho$ be the corresponding ρ-uniform family. Then, there exists a sequence of sets $X_k \in \mathcal{F}_\rho$, $k = 0, 1, \ldots$, such that the sequence is nested, $X_k \subset X_{k+1}$, and such that at each step the mesh norms satisfy $\frac{1}{4} h_{X_k} < h_{X_{k+1}} \leq \frac{1}{2} h_{X_k}$.*

We will need to consider a decomposition of $\mathbb{S}^n$ into a finite number of nonoverlapping, connected regions R_ξ, each containing an interior point ξ that will serve for function evaluations as well as labeling. For example, if $\mathcal{X}$ is the Voronoi tessellation for a set of centers X, then we may take R_ξ to be the region associated with $\xi \in X$. In any case, we will let X be the set of the ξ's used for labels and $\mathcal{X} = \{R_\xi \subset \mathbb{S}^n \,|\, \xi \in X\}$. In addition, let $\|\mathcal{X}\| = \max_{\xi\in X}\{\mathrm{diam}(R_\xi)\}$. Du, Gunzburger, and Ju [3] construct a very interesting Voronoi tessellation in which $\xi \in X$ is the centroid of $R_\xi \in \mathcal{X}$.

Spherical harmonics. We turn to the situation in which the underlying Hilbert space is $\mathcal{H} = L^2(\mathbb{S}^n)$, with $d\mu$ being the usual measure on the n-sphere. Throughout the paper, we will let $\lambda_n := \frac{n-1}{2}$ and $\{Y_{\ell,m} : \ell = 0, 1, \ldots, m = 1 \ldots d_\ell^n\}$ be the usual orthonormal set of spherical harmonics [17, 24] associated with $\mathbb{S}^n$, where for $n \geq 2$,

$$d_\ell^n = \frac{\ell+\lambda_n}{\lambda_n}\binom{\ell+n-2}{\ell} \overset{\ell\to\infty}{\sim} \frac{\ell^{n-1}}{\lambda_n(n-2)!}. \tag{4}$$

Denote by $\mathcal{H}_\ell$ the span of the spherical harmonics with fixed order ℓ, and let $\Pi_L = \bigoplus_{\ell=0}^L \mathcal{H}_\ell$ be the span of all spherical harmonics of order at most L. The orthogonal projection P_ℓ onto $\mathcal{H}_\ell$ is given by

$$\mathsf{P}_\ell f = \sum_{m=1}^{d_\ell^n} \langle f, Y_{\ell,m}\rangle Y_{\ell,m}. \tag{5}$$

Using the addition formula for spherical harmonics, one can write the kernel for this projection as

$$P_\ell(\xi,\eta) = \sum_{m=1}^{d_\ell^n} Y_{\ell,m}(\xi)\overline{Y_{\ell,m}(\eta)} = \frac{\ell+\lambda_n}{\lambda_n\omega_n} P_\ell^{(\lambda_n)}(\xi\cdot\eta), \tag{6}$$

where $\lambda_n = \frac{n-1}{2}$ and $P_\ell^{(\lambda_n)}$ is the ultraspherical polynomial of order λ_n and degree ℓ. We regard $\mathbb{S}^n$ as being the unit sphere in $\mathbb{R}^{n+1}$, and we let the quantity $\xi\cdot\eta$ denote the usual "dot" product for $\mathbb{R}^{n+1}$.

On the sphere, an operator K with a kernel of the form $K(\xi\cdot\eta)$ can be written as a convolution on $\mathbb{S}^n$; that is, $\mathsf{K}f = K * f$, where

$$K * f(\xi) = \int_{\mathbb{S}^n} K(\xi\cdot\eta) f(\eta) d\mu(\eta).$$

Because of the form of the convolution, these operators commute with rotations. Depending on the properties of the kernel, one may (and will!) apply these operators to spaces other than $L^2(\mathbb{S}^n)$.

The spherical harmonic $Y_{\ell,m}$ is an eigenfunction corresponding to the eigenvalue $-\ell(\ell+n-1) = \lambda_n^2 - (\ell+\lambda_n)^2$ for Laplace–Beltrami operator $\Delta_{\mathbb{S}^n}$ on $\mathbb{S}^n$. It follows that $\ell+\lambda_n$ is an eigenvalue corresponding to the eigenfunctions $Y_{\ell,m}\,, m = 1\ldots d_\ell^n$, of the pseudodifferential operator

$$\mathsf{L}_n := \sqrt{\lambda_n^2 - \Delta_{\mathbb{S}^n}} = \sum_{\ell=0}^\infty (\ell+\lambda_n)\mathsf{P}_\ell. \tag{7}$$

2.2. Operator frames and their kernels on $\mathbb{S}^n$. We now turn to the operators A_j defined in (2), when the underlying Hilbert space is $\mathcal{H} = L^2(\mathbb{S}^n)$ and L is proportional to the self-adjoint operator L_n given by (7). It is convenient to normalize

the L_n's when $n \geq 2$ so that the lowest eigenvalue in the spectrum is in the interval $[1, 2)$. To do that, let $j_n = \log_2\lfloor\lambda_n\rfloor$ for $n \geq 2$ and let $j_1 = 0$. We will work with $\mathsf{L} \to 2^{-j_n}\mathsf{L}_n$. Thus, $\mathsf{A}_j = a(2^{-j-j_n}\mathsf{L}_n)$, where the properties of $a \in C(\mathbb{R})$ are discussed in section 1. The spectral measure for $2^{-j_n}\mathsf{L}_n$ is $d\mathsf{E}_\lambda = \sum_{\ell=0}^\infty \mathsf{P}_\ell \delta\big(\lambda - 2^{-j_n}(\ell+\lambda_n)\big)$, where the P_ℓ's are the projections defined in (5) and have kernels given in (6). We can write the A_j's in kernel form:

$$A_j(\xi\cdot\eta) = \begin{cases} \frac{1}{\pi}\sum_{\ell=1}^\infty a(2^{-j}\ell)\cos(\ell\theta), & n = 1,\ \xi\cdot\eta = \cos\theta, \\ \sum_{\ell=0}^\infty a\left(\frac{\ell+\lambda_n}{2^{j+j_n}}\right)\frac{\ell+\lambda_n}{\lambda_n\omega_n}P_\ell^{(\lambda_n)}(\xi\cdot\eta), & n \geq 2,\ j_n = \lfloor\log_2(\lambda_n)\rfloor. \end{cases} \tag{8}$$

The operator $\mathsf{B}_J = b(2^{-J-j_n}\mathsf{L}_n)$, with b defined in (1), has the kernel

$$B_J(\xi\cdot\eta) = \begin{cases} \frac{1}{2\pi}b(0) + \frac{1}{\pi}\sum_{\ell=1}^\infty b(2^{-J}\ell)\cos(\ell\theta), & n = 1,\ \xi\cdot\eta = \cos\theta \\ \sum_{\ell=0}^\infty b\left(\frac{\ell+\lambda_n}{2^{J+j_n}}\right)\frac{\ell+\lambda_n}{\lambda_n\omega_n}P_\ell^{(\lambda_n)}(\xi\cdot\eta), & n \geq 2,\ j_n = \lfloor\log_2(\lambda_n)\rfloor. \end{cases} \tag{9}$$

Taking into account the support of a, when $n \geq 2$ in these operators it is easy to see that $\mathsf{B}_J = \sum_{j=0}^J \mathsf{A}_j\mathsf{A}_j^*$. For $n = 1$, the projection P_0 enters and $\mathsf{B}_J = \mathsf{P}_0 + \sum_{j=0}^J \mathsf{A}_j\mathsf{A}_j^*$.

We will study and establish various properties of operator kernels similar to these in section 3. In section 5 we will discuss how these give rise to tight frames on $\mathbb{S}^n$ and discuss approximation properties of these frames.

3. Localization of kernels on $\mathbb{S}^n$. We want to study the localization properties of operator kernels related to the Laplace–Beltrami operator $\Delta_{\mathbb{S}^n}$ on the sphere. As we did earlier, let $\mathsf{L}_n := \sqrt{\lambda_n^2 - \Delta_{\mathbb{S}^n}}$ and let $\kappa(t) \in C^k(\mathbb{R})$, with $k \geq \max\{2, n-1\}$, be even and satisfy

$$|\kappa^{(r)}(t)| \leq C_\kappa(1+|t|)^{r-\alpha} \text{ for all } t \in \mathbb{R},\ r = 0, \ldots, k, \tag{10}$$

where $\alpha > n + k$ and $C_\kappa > 0$ are fixed constants. We remark that all compactly supported, even C^k functions satisfy (10), as do even functions in the Schwartz class $\mathcal{S}$. Even functions in $\mathcal{S}$ satisfy (10) for arbitrarily large k and α. Define the family of operators

$$\mathsf{K}_{\varepsilon,n} := \kappa(\varepsilon\mathsf{L}_n) = \sum_{\ell=0}^\infty \kappa(\varepsilon(\ell+\lambda_n))\mathsf{P}_\ell, \quad 0 < \varepsilon \leq 1,$$

along with the associated family of kernels

$$K_{\varepsilon,n}(\underbrace{\xi\cdot\eta}_{\cos\theta}) := \begin{cases} \frac{1}{2\pi}\kappa(0) + \frac{1}{\pi}\sum_{\ell=1}^\infty \kappa(\varepsilon\ell)\cos\ell\theta, & n = 1, \\ \sum_{\ell=0}^\infty \kappa(\varepsilon(\ell+\lambda_n))\frac{\ell+\lambda_n}{\lambda_n\omega_n}P_\ell^{(\lambda_n)}(\cos\theta), & n \geq 2, \end{cases} \tag{11}$$

where $\cos\theta = \xi\cdot\eta$ and $0 < \varepsilon \leq 1$.

Our aim in this section is to obtain uniform bounds on the kernel $K_\varepsilon(\xi\cdot\eta)$ for small ε, with the bounds being explicitly dependent on ε.

The simple estimates given below in section 3.1 on the terms in the series used to define the kernels $K_{\varepsilon,n}$ confirm that, under mild conditions, these series are uniformly convergent. Let $n \geq 2$. Consider the ultraspherical identity [25, (4.7.14)] with $\lambda = \lambda_n$, $\frac{d}{dx}P_\ell^{(\lambda_n)}(x) = 2\lambda_n P_{\ell-1}^{(\lambda_n+1)}(x)$. Since $\lambda_n + 1 = \lambda_{n+2}$ and $\omega_n = \lambda_{n+2}\omega_{n+2}/\pi$, we have, for $\ell \geq 1$,

$$\frac{d}{dx}\left\{\left(\frac{\ell+\lambda_n}{\lambda_n\omega_n}\right)P_\ell^{(\lambda_n)}(x)\right\} = 2\pi\left(\frac{\ell-1+\lambda_{n+2}}{\lambda_{n+2}\omega_{n+2}}\right)P_{\ell-1}^{(\lambda_{n+2})}(x).$$

Multiply both sides by $\kappa(\varepsilon(\ell+\lambda_n))$ and sum on ℓ from 1 to ∞. Adjust the summation index on the right side and on the left use $\frac{d}{dx}P_0^{(\lambda_n)}(x) = 0$ to arrive at the identity below, which holds even when $n = 1$:

$$\frac{d}{dx}K_{\varepsilon,n}(x) = 2\pi K_{\varepsilon,n+2}(x). \tag{12}$$

3.1. Convergence issues and an L^∞ estimate on $K_{\varepsilon,n}$. The series defining the kernels are uniformly and absolutely convergent, by the M-test. This is easy to see for $n = 1$. For $n \geq 2$, start with the bound [25, (4.7.3) and (7.33.1)]

$$|P_\ell^{(\lambda_n)}(\cos\theta)| \leq \binom{\ell+n-2}{\ell} = P_\ell^{(\lambda_n)}(1), \tag{13}$$

and note that

$$\frac{\ell+\lambda_n}{\lambda_n}\binom{\ell+n-2}{\ell} \leq 2\binom{\ell+n-1}{\ell} \leq 2(1+\ell)^{n-1}.$$

From this and the assumptions on $\kappa(t)$, the terms in the series satisfy the bound

$$|\kappa(\varepsilon(\ell+\lambda_n))|\frac{\ell+\lambda_n}{\lambda_n\omega_n}\big|P_\ell^{(\lambda_n)}(\cos\theta)\big| \leq \frac{2\,C_\kappa(1+\ell)^{n-1}}{\omega_n(1+\varepsilon(\ell+\lambda_n))^\alpha} \leq \frac{2\,C_\kappa\varepsilon^{-(n-1)}}{\omega_n(1+\varepsilon\ell)^{\alpha-n+1}},$$

which suffices for the M-test, since $\alpha > n+k \geq n+2$ implies the series on the right above is convergent. Note that the estimate holds even when $n = 1$, provided the terms on the right are properly adjusted.

It is easy to take this a step further and obtain an estimate on $\|K_{\varepsilon,n}\|_\infty$, which we will need later on anyway.

PROPOSITION 3.1. *If κ satisfies* (10)*, then*

$$\|K_{\varepsilon,n}\|_\infty \leq \frac{3\,C_\kappa}{\omega_n}\varepsilon^{-n}. \tag{14}$$

Proof. From the series definition of the kernel and the estimate on each term, we get this chain of inequalities:

$$\begin{aligned}\|K_{\varepsilon,n}\|_\infty &\leq \sum_{\ell=0}^{\infty}\frac{2\,C_\kappa\varepsilon^{-(n-1)}}{\omega_n(1+\varepsilon\ell)^{\alpha-n+1}}\\ &\leq \frac{2\,C_\kappa\varepsilon^{-(n-1)}}{\omega_n} + \int_0^\infty \frac{2\,C_\kappa\varepsilon^{-(n-1)}du}{\omega_n(1+\varepsilon u)^{\alpha-n+1}}\\ &\leq \frac{2\,C_\kappa\varepsilon^{-n}}{\omega_n}\left(\varepsilon + \frac{1}{\alpha-n}\right)\end{aligned}$$

Using $\varepsilon \leq 1$ and $\alpha - n > k \geq 2$ in the previous inequality and simplifying, we obtain (14). ∎

3.2. Integral representations. We now wish to obtain integral representations for the kernels $K_\varepsilon(\cos\theta)$. We begin with the Dirichlet–Mehler integral representation for the Gegenbauer polynomials [5, p. 177],

$$P_\ell^{(\lambda)}(\cos\theta) = \frac{2^\lambda\Gamma(\lambda+\frac{1}{2})\Gamma(\ell+2\lambda)}{\sqrt{\pi}\ell!\Gamma(\lambda)\Gamma(2\lambda)(\sin\theta)^{2\lambda-1}}\int_\theta^\pi \frac{\cos\big((\ell+\lambda)\varphi - \lambda\pi\big)}{(\cos\theta-\cos\varphi)^{1-\lambda}}d\varphi,$$

which holds for any real $\lambda > 0$. We will take $\lambda = \lambda_n = \frac{n-1}{2}$, with $n \geq 2$ throughout this section. Multiply both sides of the previous equation by $\frac{\ell+\lambda_n}{\lambda_n \omega_n}$ and then simplify to get

$$\frac{\ell+\lambda_n}{\lambda_n\omega_n} P_\ell^{(\lambda_n)}(\cos\theta) = \frac{\gamma_n(\ell+\lambda_n)(\ell+n-2)!}{\ell!(\sin\theta)^{n-2}} \int_\theta^\pi \frac{\cos\big((\ell+\lambda_n)\varphi - \lambda_n\pi\big)}{(\cos\theta-\cos\varphi)^{1-\lambda_n}} d\varphi, \tag{15}$$

where

$$\gamma_n := \frac{2^{\lambda_n}\Gamma(\lambda_n+\frac{1}{2})}{\sqrt{\pi}\lambda_n\omega_n\Gamma(\lambda_n)\Gamma(2\lambda_n)}. \tag{16}$$

Using the expression on the right in (15) in the series definition of $K_{\varepsilon,n}$, we get this representation:

$$K_{\varepsilon,n}(\cos\theta) = \frac{\gamma_n}{(\sin\theta)^{n-2}} \int_\theta^\pi \frac{C_{\varepsilon,n}(\varphi)}{(\cos\theta-\cos\varphi)^{1-\lambda_n}} d\varphi, \tag{17}$$

where $C_{\varepsilon,n}$ is given by the series

$$\begin{aligned} &C_{\varepsilon,n}(\varphi) \\ &:= \sum_{\ell=0}^\infty \kappa(\varepsilon(\ell+\lambda_n)) \frac{(\ell+\lambda_n)(\ell+n-2)!}{\ell!} \begin{cases} \sin(\lambda_n\pi)\sin(\ell+\lambda_n)\varphi, & n \text{ even}, \\ \cos(\lambda_n\pi)\cos(\ell+\lambda_n)\varphi, & n \text{ odd}. \end{cases} \end{aligned} \tag{18}$$

We want to put this series in a more convenient form. To begin, the factor $\frac{(\ell+\lambda_n)(\ell+n-2)!}{\ell!}$ is the product $(\ell+\lambda_n)(\ell+n-2)(\ell+n-3)\cdots(\ell+1)$, which can be rewritten as

$$\frac{(\ell+\lambda_n)(\ell+n-2)!}{\ell!} = \prod_{r=1}^{\lfloor\frac{n-1}{2}\rfloor} \big((\ell+\lambda_n)^2 - (\lambda_n-r)^2\big) \times \begin{cases} \ell+\lambda_n, & \text{even}, \\ 1, & \text{odd}. \end{cases}$$

From this, we see that if we define the degree $n-1$ polynomial

$$Q_{n-1}(z) := \prod_{r=1}^{\lfloor\frac{n-1}{2}\rfloor} \big(z^2 - (\lambda_n-r)^2\big) \times \begin{cases} z\sin(\lambda_n\pi), & n \text{ even}, \\ \cos(\lambda_n\pi), & n \text{ odd}, \end{cases} \tag{19}$$

then we have that

$$C_{\varepsilon,n}(\varphi) = \sum_{\ell=0}^\infty \kappa(\varepsilon(\ell+\lambda_n)) Q_{n-1}(\ell+\lambda_n) \begin{cases} \sin(\ell+\lambda_n)\varphi, & n \text{ even}, \\ \cos(\ell+\lambda_n)\varphi, & n \text{ odd}. \end{cases} \tag{20}$$

We want to make a few observations about the polynomial Q_{n-1}. First, by direct calculation we have that $Q_{n-1}(-z) = (-1)^{n-1}Q_{n-1}(z)$, so that Q_{n-1} is an even function for odd n and an odd function for even n. Second, the zeros of Q_{n-1} are located at $\pm(\lambda_n - r)$ for $r = 1, \ldots, \lfloor\frac{n-1}{2}\rfloor$. This means that the function

$$g(t) := \kappa(\varepsilon t) Q_{n-1}(t) \begin{cases} \sin(t\varphi), & n \text{ even}, \\ \cos(t\varphi), & n \text{ odd}, \end{cases}$$

is even in t and has its zeros at $t = \pm(\lambda_n - r)$ for $r = 1\ldots, \lfloor\lambda_n\rfloor$. In addition, we have defined g above so that from (20) we have $C_{\varepsilon,n}(\varphi) = \sum_{\ell=0}^\infty g(\ell+\lambda_n)$.

We want to apply the Poisson summation formula (PSF),

$$\sum_{\mu\in\mathbb{Z}} f(\mu) = \sum_{\nu\in\mathbb{Z}} \hat{f}(2\pi\nu), \quad \hat{f}(\omega) = \int_{\mathbb{R}} f(t)e^{-i\omega t}dt,$$

which holds for "nice" f, to $f(t) = g(t+\lambda_n)$. Using the evenness of g and what we said about its zeros, we see that the left side of the PSF becomes

$$\sum_{\mu\in\mathbb{Z}} g(\mu+\lambda_n) = 2\sum_{\ell=0}^{\infty} g(\ell+\lambda_n) = 2C_{\varepsilon,n}(\varphi).$$

Employing elementary properties of the Fourier transform, we can show that

$$\hat{f}(\omega) = e^{i\lambda_n\omega}\hat{g}(\omega) = \varepsilon^{-1}e^{i\lambda_n\omega}Q_{n-1}(i\tfrac{d}{d\omega})\hat{\kappa}(\tfrac{\varphi+\omega}{\varepsilon}),$$

and so the right side of the PSF is

$$\begin{aligned}\sum_{\nu\in\mathbb{Z}} \hat{f}(2\pi\nu) &= \varepsilon^{-1}\sum_{\nu\in\mathbb{Z}} e^{2\pi\nu i\lambda_n}Q_{n-1}(i\tfrac{d}{d\omega})\hat{\kappa}(\tfrac{\varphi+\omega}{\varepsilon})|_{\omega=2\pi\nu}\\ &= \varepsilon^{-1}\sum_{\nu\in\mathbb{Z}}(-1)^{(n-1)\nu}Q_{n-1}(i\tfrac{d}{d\varphi})\hat{\kappa}(\tfrac{\varphi+2\pi\nu}{\varepsilon}).\end{aligned}$$

Equating the two sides of the PSF and dividing by 2, we arrive at the following result.

PROPOSITION 3.2. *If κ satisfies* (10)*, then for $n \geq 2$* (17) *holds with $C_{\varepsilon,n}$ given by*

$$C_{\varepsilon,n}(\varphi) = (2\varepsilon)^{-1}\sum_{\nu\in\mathbb{Z}}(-1)^{(n-1)\nu}Q_{n-1}(i\tfrac{d}{d\varphi})\hat{\kappa}(\tfrac{\varphi+2\pi\nu}{\varepsilon}). \tag{21}$$

In addition, for the $n = 1$ case we have

$$K_{\varepsilon,1}(\cos\theta) = (2\pi\varepsilon)^{-1}\sum_{\nu\in\mathbb{Z}}\hat{\kappa}(\tfrac{\theta+2\pi\nu}{\varepsilon}). \tag{22}$$

3.3. Estimates on $C_{\varepsilon,n}$. We need to obtain bounds on the kernels $C_{\varepsilon,n}$ from the previous section. The key to obtaining these bounds is this result.

LEMMA 3.3. *Let κ satisfy* (10)*. If $0 \leq j \leq n-1$ and $0 \leq r \leq k$ are integers, then $\frac{d^r}{dt^r}\{t^j\kappa\} \in L^1$ and $|\omega|^r|\hat{\kappa}^{(j)}(\omega)| \leq \|\frac{d^r}{dt^r}\{t^j\kappa\}\|_{L^1}$.*

Proof. Since $\kappa \in C^k$, the derivative $\frac{d^r}{dt^r}\{t^j\kappa\}$ is a linear combination of terms of the form $t^p\kappa^{(q)}$, each of which is bounded by a multiple of $(1+|t|)^{p+q-\alpha}$. This is in L^1 because $\alpha - p - q > \alpha - (n-1) - k > 1$. This allows us to apply standard properties of the Fourier transform to obtain the formula $(-i)^{r+j}\omega^r\hat{\kappa}^{(j)}(\omega) = \widehat{\frac{d^r}{dt^r}\{t^j\kappa\}}$, which immediately implies the inequality. ∎

Consider the function $\left(\frac{\varphi+\omega}{\varepsilon}\right)^r Q_{n-1}(i\frac{d}{d\varphi})\hat{\kappa}(\frac{\varphi+\omega}{\varepsilon}) = \sum_{j=0}^{n-1}\varepsilon^{-j}q_{j,n}\left(\frac{\varphi+\omega}{\varepsilon}\right)^r\hat{\kappa}^{(j)}(\frac{\varphi+\omega}{\varepsilon})$, where $Q_{n-1}(z) = \sum_{j=0}^{n-1} q_{j,n}z^j$ is defined in (19). From Lemma 3.3, we have that

$$\begin{aligned}\left|\left(\tfrac{\varphi+\omega}{\varepsilon}\right)^r Q_{n-1}(i\tfrac{d}{d\varphi})\hat{\kappa}(\tfrac{\varphi+\omega}{\varepsilon})\right| &\leq \textstyle\sum_{j=0}^{n-1}\varepsilon^{-j}|q_{j,n}|\,\|\frac{d^r}{dt^r}\{t^j\kappa\}\|_{L^1}\\ &\leq B_{n,k,\kappa}\varepsilon^{-(n-1)},\end{aligned}$$

where

$$B_{n,k,\kappa} := \left(\sum_{j=0}^{n-1} |q_{n,j}|\right) \max_{j<n,\, r\le k} \| \tfrac{d^r}{dt^r} \left\{ t^j \kappa \right\} \|_{L^1}. \tag{23}$$

Adding the inequalities for $r = 0$ and $r = k$ and manipulating the result, we get that

$$\left| Q_{n-1}(i\tfrac{d}{d\varphi}) \hat{\kappa}(\tfrac{\varphi+\omega}{\varepsilon}) \right| \le \frac{2B_{n,k,\kappa}\varepsilon^{-(n-1)}}{1 + \left|\frac{\varphi+\omega}{\varepsilon}\right|^k}.$$

We can use this inequality in conjunction with the series for $C_{\varepsilon,n}$ in (21) to arrive at the bound

$$|C_{\varepsilon,n}(\varphi)| \le (2\varepsilon)^{-1} \sum_{\nu\in\mathbb{Z}} \frac{2B_{n,k,\kappa}\varepsilon^{-(n-1)}}{1 + \left|\frac{\varphi+2\pi\nu}{\varepsilon}\right|^k} = \sum_{\nu\in\mathbb{Z}} \frac{B_{n,k,\kappa}\varepsilon^{-n}}{1 + \left|\frac{\varphi+2\pi\nu}{\varepsilon}\right|^k}, \tag{24}$$

which holds for all $\varphi \in \mathbb{R}$ and $0 < \varepsilon \le 1$. If we restrict φ to be in the interval $[0, \pi]$, then the dominant term in the series on the right comes from $\nu = 0$. The other terms are each bounded above by $B_{n,k,\kappa}\varepsilon^{k-n}((2|\nu| - 1)\pi)^{-k}$. Summing them and then estimating the resulting series by an integral gives us

$$\sum_{\nu\in\mathbb{Z},\nu\neq 0} \frac{B_{n,k,\kappa}\varepsilon^{-n}}{1 + \left|\frac{\varphi+2\pi\nu}{\varepsilon}\right|^k} \le B_{n,k,\kappa}\varepsilon^{k-n}\pi^{-k}\frac{2k-1}{k-1}.$$

Multiply top and bottom on the left above by $1 + \left(\frac{\varphi}{\varepsilon}\right)^k$ and use $0 \le \varphi \le \pi$ and $k \ge 2$ to get

$$\sum_{\nu\in\mathbb{Z},\nu\neq 0} \frac{B_{n,k,\kappa}\varepsilon^{-n}}{1 + \left|\frac{\varphi+2\pi\nu}{\varepsilon}\right|^k} \le \frac{6B_{n,k,\kappa}\varepsilon^{-n}}{1 + \left(\frac{\varphi}{\varepsilon}\right)^k}.$$

Combining this bound with that from (24) yields the result below.

PROPOSITION 3.4. *Let κ satisfy* (10)*, with $k \ge 2$ and $n \ge 2$. If $0 \le \varphi \le \pi$, then the kernel $C_{\varepsilon,n}$ defined in* (18) *satisfies the bound*

$$|C_{\varepsilon,n}(\varphi)| \le \frac{7B_{n,k,\kappa}\varepsilon^{-n}}{1 + \left(\frac{\varphi}{\varepsilon}\right)^k}. \tag{25}$$

In addition, for the case $n = 1$, we have

$$|K_{\varepsilon,1}(\cos\theta)| \le \frac{7B_{1,k,\kappa}\varepsilon^{-1}}{1 + \left(\frac{\theta}{\varepsilon}\right)^k}. \tag{26}$$

Proof. Only the second inequality requires comment. The proof we gave works for the $n = 1$ case because it has the form given in (22), which is essentially the same as that for the $C_{\varepsilon,n}$'s. ∎

3.4. Estimates on $K_{\varepsilon,n}$. We now turn to obtaining explicit bounds on the ΨDO kernels $K_{\varepsilon,n}$ similar to the bound on $K_{\varepsilon,1}$ in (26). From the integral representation in (17) and the bound on $C_{\varepsilon,n}$, we have that

$$|K_{\varepsilon,n}(\cos\theta)| \le \frac{7B_{n,k,\kappa}\gamma_n\varepsilon^{-n}}{(\sin\theta)^{n-2}} \int_\theta^\pi \frac{(\cos\theta - \cos\varphi)^{\frac{n-3}{2}} d\varphi}{1 + \left(\frac{\varphi}{\varepsilon}\right)^k}. \tag{27}$$

The two values of θ that present difficulties are $\theta = 0$ and $\theta = \pi$. The form of the inequality above is adequate for the $\theta = 0$ case, but needs to be reformulated for the $\theta = \pi$ case. To do that, we begin by denoting the angle supplementary to an angle α by $\tilde{\alpha}$, so throughout this section we will let $\tilde{\theta} = \pi - \theta$ and $\tilde{\varphi} = \pi - \varphi$. Changing variables in the integral on the right above and using $\sin\tilde{\alpha} = \sin\alpha$ and $\cos\tilde{\alpha} = -\cos\alpha$, we have the following reformulation of (27):

$$|K_{\varepsilon,n}(\cos\theta)| \le \frac{7B_{n,k,\kappa}\gamma_n\varepsilon^{-n}}{(\sin\tilde{\theta})^{n-2}} \int_0^{\tilde{\theta}} \frac{(\cos\tilde{\varphi} - \cos\tilde{\theta})^{\frac{n-3}{2}} d\tilde{\varphi}}{1 + \left(\frac{\pi-\tilde{\varphi}}{\varepsilon}\right)^k}. \tag{28}$$

The next step is to bound both of these integrals. Recall the sum-to-product identity, $\cos\alpha - \cos\beta \equiv 2\sin\frac{\alpha+\beta}{2}\sin\frac{\beta-\alpha}{2}$, which holds for all α and β. Assuming that $\pi \ge \beta > \alpha \ge \pi/2$ and using the fact that $\frac{\sin t}{t}$ is decreasing for $0 \le t \le \pi$, we have that

$$6 < 8\frac{\sin(3\pi/4)}{3\pi/4}\frac{\sin(\pi/4)}{\pi/4} \le \frac{\cos\alpha - \cos\beta}{\beta^2 - \alpha^2} = 8\frac{\sin\frac{\alpha+\beta}{2}}{\frac{\alpha+\beta}{2}}\frac{\sin\frac{\beta-\alpha}{2}}{\frac{\beta-\alpha}{2}} \le 8,$$

and so

$$\left(\frac{\cos\alpha - \cos\beta}{\beta^2 - \alpha^2}\right)^{\frac{n-3}{2}} \le 2^{\frac{3(n-3)}{2}} \times \begin{cases} \frac{2}{\sqrt{3}}, & n = 2, \\ 1, & n \ge 3 \end{cases} \le 2 \cdot 2^{\frac{3(n-3)}{2}}. \tag{29}$$

Assume that $\varepsilon \le \theta \le \pi/2$, and apply (29) to (27) to get the following chain of inequalities:

$$\begin{aligned} |K_{\varepsilon,n}(\cos\theta)| &\le \frac{14 \cdot 2^{\frac{3(n-3)}{2}} B_{n,k,\kappa}\gamma_n\varepsilon^{-n}}{(\sin\theta)^{n-2}} \int_\theta^\pi \frac{(\theta^2 - \varphi^2)^{\frac{n-3}{2}} d\varphi}{1 + \left(\frac{\varphi}{\varepsilon}\right)^k} \\ &\le 14 \cdot 2^{\frac{3(n-3)}{2}} B_{n,k,\kappa}\gamma_n\varepsilon^{-n} \left(\frac{\theta}{\sin\theta}\right)^{n-2} \int_1^{\pi/\theta} \frac{(t^2-1)^{\frac{n-3}{2}} dt}{1 + (\theta/\varepsilon)^k t^k} \\ &\le \frac{14 \cdot 2^{\frac{3(n-3)}{2}} B_{n,k,\kappa}\gamma_n\varepsilon^{-n}(\pi/2)^{n-2}}{(\frac{\theta}{\varepsilon})^k} \int_1^\infty \frac{(t^2-1)^{\frac{n-3}{2}} dt}{t^k}. \end{aligned}$$

Use $2(\theta/\varepsilon)^k \ge 1 + (\theta/\varepsilon)^k$, change variables of integration from $t \to 1/t$, and note that because $k \ge \max\{2, n-1\} \ge n-1$, the resulting integral on the right is bounded above by $\int_0^1 (1-t^2)^{\frac{n-3}{2}} dt = 2^{n-3}\Gamma(\lambda_n)^2/\Gamma(2\lambda_n)$ [26, p. 255]. After simplifying, we arrive at this estimate:

$$|K_{\varepsilon,n}(\cos\theta)| \le \frac{14 \cdot 2^{\frac{3(n-3)}{2}} \pi^{n-2} B_{n,k,\kappa}\gamma_n\Gamma(\lambda_n)^2/\Gamma(2\lambda_n)}{1 + (\frac{\theta}{\varepsilon})^k} \varepsilon^{-n}.$$

The messy quantity in the numerator can be simplified considerably. This requires employing the definition of γ_n in (16), the formula for ω_n, the familiar properties of the Γ-function, along with the less familiar duplication formula [26, p. 240], $\sqrt{\pi}\Gamma(2z) = 2^{2z-1}\Gamma(z)\Gamma(z+\frac{1}{2})$, and manipulating the expressions involved. The result is that

$$2^{\frac{3(n-3)}{2}} \pi^{n-2}\gamma_n\Gamma(\lambda_n)^2/\Gamma(2\lambda_n) = \frac{\omega_{n-1}}{4\sqrt{\pi}}, \ \omega_{n-1} = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}.$$

Thus we can rewrite the previous inequality, which holds for $\varepsilon \le \theta \le \pi/2$, as

$$|K_{\varepsilon,n}(\cos\theta)| \le \frac{7\omega_{n-1}B_{n,k,\kappa}}{2\sqrt{\pi}(1+(\frac{\theta}{\varepsilon})^k)}\varepsilon^{-n}.$$

If we now apply (29) to (28), with $0 \le \tilde{\theta} \le \pi/2$ (or, equivalently, $\pi/2 \le \theta \le \pi$), then

$$\begin{aligned}|K_{\varepsilon,n}(\cos\theta)| &\le \frac{7B_{n,k,\kappa}\gamma_n\varepsilon^{-n}}{(\sin\tilde{\theta})^{n-2}}\int_0^{\tilde{\theta}}\frac{(\tilde{\theta}^2-\tilde{\varphi}^2)^{\frac{n-3}{2}}d\tilde{\varphi}}{1+\left(\frac{\pi-\tilde{\varphi}}{\varepsilon}\right)^k}\\ &\le \frac{14\cdot 2^{\frac{3(n-3)}{2}}B_{n,k,\kappa}\gamma_n\varepsilon^{-n}}{(1+\left(\frac{\theta}{\varepsilon}\right)^k)(\sin\tilde{\theta})^{n-2}}\int_0^{\tilde{\theta}}(\tilde{\theta}^2-\tilde{\varphi}^2)^{\frac{n-3}{2}}d\tilde{\varphi}.\end{aligned}$$

Carrying out manipulations analogous to those for the previous case, we obtain

$$|K_{\varepsilon,n}(\cos\theta)| \le \frac{7\omega_{n-1}B_{n,k,\kappa}}{4\sqrt{\pi}(1+(\frac{\theta}{\varepsilon})^k)}\varepsilon^{-n}.$$

The final case concerns $0 \le \theta \le \varepsilon$. For such θ, we have, from the L^∞ bound in (14), that

$$|K_{\varepsilon,n}(\cos\theta)| \le \frac{3\,C_\kappa}{\omega_n}\varepsilon^{-n} \le \frac{3\,C_\kappa}{\omega_n}\left(\frac{1+(\frac{\theta}{\varepsilon})^k}{1+(\frac{\theta}{\varepsilon})^k}\right)\varepsilon^{-n} \le \frac{6\,C_\kappa}{\omega_n(1+(\frac{\theta}{\varepsilon})^k)}\varepsilon^{-n},$$

which, when combined with (22) for $n=1$, gives us the main result of this section.

THEOREM 3.5. *Let κ satisfy* (10)*, with $k \ge \max\{2, n-1\}$. If $0 \le \theta \le \pi$, then the kernel $K_{\varepsilon,n}$ satisfies the bound*

$$|K_{\varepsilon,n}(\cos\theta)| \le \frac{\beta_{n,k,\kappa}}{1+(\frac{\theta}{\varepsilon})^k}\varepsilon^{-n}, \tag{30}$$

where

$$\beta_{n,k,\kappa} := \begin{cases} 7B_{1,k,\kappa} & \text{if } n=1,\\ \max\left\{\frac{6\,C_\kappa}{\omega_n}, \frac{7\omega_{n-1}B_{n,k,\kappa}}{2\sqrt{\pi}}\right\} & \text{if } n \ge 2.\end{cases} \tag{31}$$

We conclude this section with an application of this theorem to obtaining a bound on the L^1 norm of $K_{\varepsilon,n}(\xi\cdot\eta)$, with η fixed. By the Funk–Hecke formula [17, Theorem 6], this norm is given by

$$\int_{\mathbb{S}^n}|K_{\varepsilon,n}(\xi\cdot\eta)|d\mu(\xi) = \omega_{n-1}\int_0^\pi |K_{\varepsilon,n}(\cos\theta)|\sin^{n-1}\theta\,d\theta,$$

which is of course independent of η. For that reason we will drop any reference to η and denote the norm by $\|K_{\varepsilon,n}\|_1$. Here is the bound we want.

COROLLARY 3.6. *Let $n \ge 1$. If κ satisfies* (10)*, with $k > \max\{2, n\}$, then*

$$\|K_{\varepsilon,n}\|_1 \le 2\omega_{n-1}\beta_{n,k,\kappa}.$$

Proof. By Theorem 3.5 and the remarks above, we have

$$\|K_{\varepsilon,n}\|_1 \le \omega_{n-1}\int_0^\pi |K_{\varepsilon,n}(\cos\theta)|\sin^{n-1}\theta\,d\theta \le \beta_{n,k,\kappa}\omega_{n-1}\varepsilon^{-n}\int_0^\pi \frac{\sin^{n-1}\theta\,d\theta}{1+(\frac{\theta}{\varepsilon})^k}.$$

The integral on the right side above can be estimated this way:

$$\begin{aligned}\int_0^\pi \frac{\sin^{n-1}\theta\,d\theta}{1+(\frac{\theta}{\varepsilon})^k} &< \varepsilon^n\int_0^{\pi/\varepsilon}\frac{t^{n-1}dt}{1+t^k}\\ &< \varepsilon^n\left\{\int_0^1 t^{n-1}dt + \int_1^\infty \frac{dt}{t^{k+1-n}}\right\} < 2\varepsilon^n.\end{aligned}$$

The corollary then follows immediately from the estimate. ∎

3.5. Operator properties of $\mathsf{K}_{\varepsilon,n}$. We now turn to the operator properties of $\mathsf{K}_{\varepsilon,n}$. Our first result is calculating the norm of the map of $\mathsf{K}_{\varepsilon,n} : L^p \to L^q$. After that we will prove a lemma showing that for certain κ the operator $\mathsf{K}_{\varepsilon,n}$ will be a reproducing kernel on Π_L. We will close the section with a result showing that for such κ and $\varepsilon \le (L+\lambda_n)^{-1}$ the norm of $f - \mathsf{K}_{\varepsilon,n}f$ is comparable to the distance from f to Π_L in appropriate norms.

THEOREM 3.7. *If κ satisfies* (10)*, with $k > \max\{2,n\}$, then, for all $1 \le p \le \infty$ and $1 \le q \le \infty$, the operator $\mathsf{K}_{\varepsilon,n}\colon L^p(\mathbb{S}^n) \to L^q(\mathbb{S}^n)$ is bounded and its norm satisfies*

$$\|\mathsf{K}_{\varepsilon,n}\|_{p,q} \le 2\omega_{n-1}\beta_{n,k,\kappa}(4\omega_{n-1}\varepsilon^n)^{-(\frac{1}{p}-\frac{1}{q})_+},$$

where $\beta_{n,k,\kappa}$ is defined in (31) *and $(x)_+ = x$ for $x > 0$ and $(x)_+ = 0$ otherwise.*

Proof. The operators are all of the form $K_{\varepsilon,n} * f$ and so, for the (p,q) pairs $(1,1)$, (∞,∞), $(\infty,1)$, all satisfy $\|K_{\varepsilon,n} * f\|_q \le \|K_{\varepsilon,n}\|_1\|f\|_p$. By the Riesz–Thorin theorem [28, p. 95] and Corollary 3.6, we then have for $1 \le q \le p \le \infty$

$$\|\mathsf{K}_{\varepsilon,n}\|_{p,q} \le \|K_{\varepsilon,n}\|_1 \le 2\omega_{n-1}\beta_{n,k,\kappa}.$$

For the pair $(1,\infty)$, we have $\|K_{\varepsilon,n} * f\|_\infty \le \|K_{\varepsilon,n}\|_\infty\|f\|_1$. By (14) and (31), we have $\|K_{\varepsilon,n}\|_\infty \le \frac{1}{2}\beta_{n,k,\kappa}\varepsilon^{-n}$, and so $\|K_{\varepsilon,n} * f\|_\infty \le \frac{1}{2}\beta_{n,k,\kappa}\varepsilon^{-n}\|f\|_1$. Apply the Riesz–Thorin theorem to the pairs (p,q), where $\frac{1}{p} = (1-t)\alpha + t$ and $\frac{1}{q} = (1-t)\alpha$, where $0 < t < 1$ and $0 < \alpha < 1$, $(\frac{1}{\alpha},\frac{1}{\alpha})$ and $(1,\infty)$ to get

$$\|\mathsf{K}_{\varepsilon,n}\|_{p,q} \le (2\beta_{n,k,\kappa}\omega_{n-1})^{1-t}\left(\frac{1}{2}\beta_{n,k,\kappa}\varepsilon^{-n}\right)^t = 2\omega_{n-1}\beta_{n,k,\kappa}(4\omega_{n-1}\varepsilon^n)^{-t}.$$

Since $\frac{1}{p} = (1-t)\alpha + t = \frac{1}{q} + t$, $t = \frac{1}{p} - \frac{1}{q}$. Thus, for $q > p$, we have

$$\|\mathsf{K}_{\varepsilon,n}\|_{p,q} \le 2\omega_{n-1}\beta_{n,k,\kappa}(4\omega_{n-1}\varepsilon^n)^{-(\frac{1}{p}-\frac{1}{q})}.$$

Putting the last inequality together with that for $q \le p$ yields the result. ∎

The following lemma is obvious.

LEMMA 3.8. *Let $L > 0$ be an integer and let $0 < \varepsilon \le (L+\lambda_n)^{-1}$. If κ satisfies* (10)*, with $k \ge \max\{2,n-1\}$, and if $\kappa(t) \equiv 1$ on $[0,1]$, then $K_{\varepsilon,n}(\xi\cdot\eta)$ is a reproducing kernel on Π_L, the space of spherical harmonics having degree at most L.*

Remark 3.9. Let $L > 0$ be an integer. If we choose ε so that $L = \lfloor \varepsilon^{-1} - \lambda_n \rfloor$, then by combining the previous theorem and lemma we get a familiar result about harmonic polynomials: *If $S \in \Pi_L$, then $\|S\|_q \le C_n L^{n(\frac{1}{p}-\frac{1}{q})_+}\|S\|_p$.*

We let $E_L(f)_p$ denote the distance of $f \in L^p(\mathbb{S}^n)$ to Π_L, i.e.,

$$E_L(f)_p := \inf_{S \in \Pi_L} \|f - S\|_p. \tag{32}$$

COROLLARY 3.10. *Let κ satisfy (10), with $k > \max\{2, n\}$, and in addition suppose $\kappa(t) \equiv 1$ on $[0, 1]$. If $f \in L^p(\mathbb{S}^n)$, $1 \le p \le \infty$, and $\varepsilon \le (L + \lambda_n)^{-1}$, then*

$$\|f - K_{\varepsilon,n} * f\|_p \le (1 + 2\omega_{n-1}\beta_{n,k,\kappa}) E_L(f)_p. \tag{33}$$

Also, for $1 \le p < \infty$ or, if $p = \infty$, for $f \in C(\mathbb{S}^n)$, we have $\lim_{\varepsilon \downarrow 0} K_{\varepsilon,n} * f = f$.

Proof. By Lemma 3.8, then $K_{\varepsilon,n} * S = S$ if $S \in \Pi_L$. It follows that $f - K_{\varepsilon,n} * f = (I + \mathsf{K}_\varepsilon)(f - S)$. From this and Theorem 3.7, we have that $\|f - K_{\varepsilon,n} * f\|_p \le (1 + 2\omega_{n-1}\beta_{n,k,\kappa})\|f - S\|_p$. Taking the infimum over all $S \in \Pi_L$ yields (33). That $\lim_{\varepsilon \downarrow 0} K_{\varepsilon,n} * f = f$ follows from (33) together with the fact that the spherical harmonics are dense in L^p for $1 \le p < \infty$ and in $C(\mathbb{S}^n)$ in the usual L^∞ norm [24, section IV.2]. ∎

The estimate in (33) is useful for obtaining rates of approximation, simply because rates of approximation by spherical harmonics are well known for many classes of functions; see, for example, Rustamov [23]. For further discussion, see the remarks following Proposition 5.1.

4. Quadrature on $\mathbb{S}^n$. To do the discretizations required to construct tight spherical frames in section 5, we need a strengthened version of the quadrature formula given in [14, 15]. There are two reasons for this. First, the earlier quadrature formula applies to a partition of $\mathbb{S}^n$ that is restricted. Second, it utilizes a set of centers that is not a general set of scattered points, but rather a set that has been "culled" from one. Our aim is to use the results obtained in section 3 to produce an improved positive-weight quadrature formula that avoids these restrictions. Indeed, out of this will also come strengthened versions of the inequalities derived in [14].

4.1. Marcinkiewicz–Zygmund inequalities. In this section we wish to give Marcinkiewicz–Zygmund type inequalities. These inequalities provide equivalences between norms defined through integrals and discrete norms stemming from sampled points and certain weights. Here, instead of polynomials, we will work with functions of the form $K_{\varepsilon,n} * f$ for $f \in L^1(\mathbb{S}^n)$.

The place to start is with a decomposition of the sphere into a finite number of nonoverlapping, connected regions R_ξ, each containing an interior point ξ that will serve for function evaluations as well as labeling. For example, given a set of centers X, one can form the corresponding Voronoi tessellation, and then take R_ξ to be the region associated with $\xi \in X$. In any case, we will let X be the set of the ξ's used for labels and $\mathcal{X} = \{R_\xi \subset \mathbb{S}^n \,|\, \xi \in X\}$. In addition, let $\|\mathcal{X}\| = \max_{\xi \in X}\{\mathrm{diam}(R_\xi)\}$.

The quantity that we wish to estimate first is the magnitude of the difference between the continuous and discrete norms for $g = K_{\varepsilon,n} * f$,

$$E_{\mathcal{X}} := \big|\|g\|_1 - \sum_{\xi \in X} |g(\xi)|\mu(R_\xi)\big|,$$

where we assume that $f \in L^1(\mathbb{S}^n)$. It is straightforward to show that

$$E_{\mathcal{X}} \le \sum_{\xi \in X} \int_{R_\xi} |g(\eta) - g(\xi)| d\mu(\eta) \le \sup_{\zeta \in \mathbb{S}^n} F_{\varepsilon,\mathcal{X}}(\zeta) \|f\|_1 ,$$

where $F_{\varepsilon,\mathcal{X}}(\zeta) := \sum_{\xi\in X}\int_{R_\xi}\big|K_{\varepsilon,n}(\eta\cdot\zeta)-K_{\varepsilon,n}(\xi\cdot\zeta)\big|d\mu(\eta)$, which is the quantity we need to estimate.

Choose ζ to be the north pole of $\mathbb{S}^n$ and let θ be the colatitude in spherical coordinates; set $\theta_\eta = \cos^{-1}(\eta\cdot\zeta)$ and $\theta_\xi = \cos^{-1}(\xi\cdot\zeta)$. Denote by θ_ξ^+ and θ_ξ^-, respectively, the high and low values for θ over R_ξ. Using (12) for the derivative of $K_{\varepsilon,n}$, we can write $F_{\varepsilon,\mathcal{X}}(\zeta)$ as

$$\begin{aligned}F_{\varepsilon,\mathcal{X}}(\zeta) &= 2\pi\sum_{\xi\in X}\int_{R_\xi}\left|\int_{\theta_\zeta}^{\theta_\eta}K_{\varepsilon,n+2}(\cos t)\sin t dt\right|d\mu(\eta)\\ &\le 2\pi\sum_{\xi\in X}\mu(R_\xi)\int_{\theta_\xi^-}^{\theta_\xi^+}|K_{\varepsilon,n+2}(\cos t)|\sin t\,dt.\end{aligned}$$

Divide $\mathbb{S}^n$ into $M = \lfloor\pi/\|\mathcal{X}\|\rfloor$ equal bands in which $(m-1)\pi/M\le\theta\le m\pi/M$, $m=1,\dots,M$. To avoid trivial situations and simplify later inequalities, we will assume that $M\ge 3$. Call these bands $B_1,\dots,B_M$. Each R_ξ can have nontrivial intersection with at most two adjacent bands, because $\operatorname{diam}(R_\xi)\le\|\mathcal{X}\|\le\pi/M$. So if $R_\xi\subset B_m\cup B_{m+1}$, then $(m-1)\pi/M\le\theta_\xi^-\le\theta_\xi^+\le(m+1)\pi/M$. In addition, the sum of the contributions from all $R_\xi\subset B_m\cup B_{m+1}$ is bounded above by the quantity

$$I_m := 2\pi\mu(B_m\cup B_{m+1})\int_{\frac{m-1}{M}\pi}^{\frac{m+1}{M}\pi}|K_{\varepsilon,n+2}(\cos t)|\sin t\,dt,$$

where $\mu(B_m\cup B_{m+1}) = \omega_{n-1}\int_{\frac{m-1}{M}\pi}^{\frac{m+1}{M}\pi}\sin^{n-1}t\,dt$. It follows that $F_{\varepsilon,\mathcal{X}}(\zeta)\le\sum_{m=1}^{M-1}I_m$. From Theorem 3.5, if we assume $k\ge n+2>\max\{2,n+1\}$ and if we use various linear approximations to the sine, we have

$$I_m \le 2\pi\omega_{n-1}\beta_{n+2,k,\kappa}\varepsilon^{-n-2}\int_{\frac{m-1}{M}\pi}^{\frac{m+1}{M}\pi}t^{n-1}dt\int_{\frac{m-1}{M}\pi}^{\frac{m+1}{M}\pi}\frac{t}{1+(\frac{t}{\varepsilon})^k}dt. \tag{34}$$

For $2\le m\le M-1$, we can bound the first integral by $\frac{2\pi}{M}\left(\frac{m+1}{M}\pi\right)^{n-1}$. In the second integral, we divide and multiply the integrand by t^{n-1}, and replace the t^{n-1} in the denominator by its lowest value. The result is that

$$\int_{\frac{m-1}{M}\pi}^{\frac{m+1}{M}\pi}\frac{t}{1+(\frac{t}{\varepsilon})^k}dt\le\left(\frac{M}{(m-1)\pi}\right)^{n-1}\int_{\frac{m-1}{M}\pi}^{\frac{m+1}{M}\pi}\frac{t^n}{1+(\frac{t}{\varepsilon})^k}dt.$$

Putting these two bounds together yields

$$I_m\le\frac{4\pi^2}{M}\omega_{n-1}\beta_{n+2,k,\kappa}\varepsilon^{-n-2}\underbrace{\left(\frac{m+1}{m-1}\right)^{n-1}}_{\le 3^{n-1}}\int_{\frac{m-1}{M}\pi}^{\frac{m+1}{M}\pi}\frac{t^n}{1+(\frac{t}{\varepsilon})^k}dt.$$

Summing both sides from $m=2$ to $M-1$, taking account of intervals appearing twice in the sum, and doing some obvious manipulations, we obtain

$$\begin{aligned}\sum_{m=2}^{M-1}I_m &< \frac{8\pi^2 3^{n-1}\omega_{n-1}}{M\varepsilon}\beta_{n+2,k,\kappa}\int_{\frac{\pi}{M\varepsilon}}^{\frac{\pi}{\varepsilon}}\frac{t^n}{1+t^k}dt\\ &< \frac{8\pi^2 3^{n-1}\omega_{n-1}}{M\varepsilon}\beta_{n+2,k,\kappa}\underbrace{\int_0^\infty\frac{t^n}{1+t^k}dt}_{\le 3/2} < \frac{4\pi^2 3^n\omega_{n-1}}{M\varepsilon}\beta_{n+2,k,\kappa}.\end{aligned}$$

We now need to estimate I_1. From (34) we have

$$I_1 \le n^{-1}\omega_{n-1}(2\pi/M)^n \int_0^{\frac{2\pi}{M}} \frac{\beta_{n+2,k,\kappa}\varepsilon^{-n-2}}{1+(\frac{t}{\varepsilon})^k} t dt < \frac{\omega_{n-1}}{2n}\left(\frac{2\pi}{M\varepsilon}\right)^{n+2}\beta_{n+2,k,\kappa}.$$

We arrive at the estimate

$$F_{\varepsilon,\mathcal{X}}(\zeta) \le 2\pi\omega_{n-1}\beta_{n+2,k,\kappa}\frac{2\pi}{M\varepsilon}\left\{\frac{1}{2n}\left(\frac{2\pi}{M\varepsilon}\right)^{n+1} + 3^n\right\}.$$

To finish up, we want to put our inequalities in terms of the ratio $\|\mathcal{X}\|/\varepsilon$. Since we have assumed that $M \ge 3$, we have that $\pi/M \le \frac{4}{3}\|\mathcal{X}\|$. Using this in the previous inequality and simplifying, we arrive at

$$F_{\varepsilon,\mathcal{X}}(\zeta) < 16\pi\cdot 3^{n-1}\omega_{n-1}\beta_{n+2,k,\kappa}\frac{\|\mathcal{X}\|}{\varepsilon}\left\{1+\frac{3}{2n}\left(\frac{8\|\mathcal{X}\|}{9\varepsilon}\right)^{n+1}\right\}.$$

We remark that if $\|\mathcal{X}\| \le \varepsilon \le 1$, then the assumption that $M \ge 3$ is automatically fulfilled. In addition, the right side of the inequality above is independent of ζ, so it holds for the left replaced by $\sup_{\zeta\in\mathbb{S}^n} F_{\varepsilon,\mathcal{X}}(\zeta)$. Finally, the inequality itself simplifies considerably. We collect all these observations in the result below.

PROPOSITION 4.1. *Let κ satisfy (10) with $k \ge n+2$, and for $f \in L^1(\mathbb{S}^n)$ let $g = K_{\varepsilon,n} * f$. If $\mathcal{X}$ is the decomposition of $\mathbb{S}^n$ described above and if $\|\mathcal{X}\| \le \varepsilon \le 1$, then*

$$\left|\|g\|_1 - \sum_{\xi\in X}|g(\xi)|\mu(R_\xi)\right| \le 16\pi\cdot 3^n\omega_{n-1}\beta_{n+2,k,\kappa}\frac{\|\mathcal{X}\|}{\varepsilon}\|f\|_1. \tag{35}$$

This result leads immediately to a version of the Marcinkiewicz–Zygmund inequalities for $\mathbb{S}^n$. This result extends an earlier result proved in [14, Theorem 3.1]. As we noted at the start of the section, the earlier result held only for restricted classes of decompositions.

THEOREM 4.2. *Let $L > 0$ be an integer and let $\delta \in (0,1)$. If $\mathcal{X}$ is the decomposition of $\mathbb{S}^n$ described above and $S \in \Pi_L$, then there exists a constant $s_n \ge 1$, which depends only on n, such that*

$$(1-\delta)\|S\|_1 \le \sum_{\xi\in X}|S(\xi)|\mu(R_\xi) \le (1+\delta)\|S\|_1 \tag{36}$$

holds whenever $\|\mathcal{X}\| \le \delta s_n^{-1}(L+\lambda_n)^{-1}$.

Proof. Let κ satisfy (10), with $k \ge n+2$. In addition, require $\kappa(t) \equiv 1$ for $t \in [0,1]$. Choose $\varepsilon = (L+\lambda_n)^{-1}$. By Lemma 3.8, $S = K_{\varepsilon,n} * S$, and so if we take $f = S$ and $\|\mathcal{X}\| \le \varepsilon = (L+\lambda_n)^{-1} \le 1$ in Proposition 4.1, then $g = K_{\varepsilon,n} * S = S$ there. Manipulating the resulting expression in (35) then gives us

$$\tilde{s}_n := \sup \frac{\left|\|S\|_1 - \sum_{\xi\in X}|S(\xi)|\mu(R_\xi)\right|}{(L+\lambda_n)\|\mathcal{X}\|\|S\|_1} \le 16\pi\cdot 3^n\omega_{n-1}\beta_{n+2,k,\kappa}\,,$$

where the supremum is over all $\mathcal{X}$ and $L > 0$ such that $\|\mathcal{X}\| \le (L+\lambda_n)^{-1}$ and clearly depends *only* on n. Now, let

$$s_n := \max\{1, \tilde{s}_n\} \le \max\{1, 16\pi\cdot 3^n\omega_{n-1}\beta_{n+2,k,\kappa}\}. \tag{37}$$

If we further restrict $\|\mathcal{X}\|$ so that $\|\mathcal{X}\| \le \delta s_n^{-1}(L+\lambda_n)^{-1}$, then (36) follows. ∎

4.2. Positive-weight quadrature for $\mathbb{S}^n$. Our aim is to extend the quadrature formula in [14, Theorem 4.1] to more general sets of centers and decompositions than the restricted class covered there. Even more important for us here is obtaining upper and *lower* bounds on the positive weights. For the restricted case covered in [14], upper bounds were given in [15], but nothing was said about lower bounds, which we need for constructing tight-frames on $\mathbb{S}^n$.

There is an important map associated with Π_L and the decomposition $\mathcal{X}$ and the corresponding finite set X. Let $|X|$ be the cardinality of X. We define the sampling map, $T_X : \Pi_L \to \mathbb{R}^{|X|}$, by $T_X S := (S(\xi))_{\xi\in X}$. From Theorem 4.2, it follows that if $\|\mathcal{X}\| \le \delta s_n^{-1}(L+\lambda_n)^{-1}$ holds and if $T_X S = 0$, we have that $\|S\|_1 = 0$ and, hence, $S \equiv 0$. The sampling map, which is linear, is therefore injective. Moreover, if we let the subspace $V_L = T_X\Pi_L \subset \mathbb{R}^{|X|}$, then the inverse map $T_X^{-1} : V_L \to \Pi_L$ is of course linear. Also, we will let $S_X = (S(\xi))_{\xi\in X}$.

Since our interest here is in weights for quadrature, we start with the linear functional $\Phi : \Pi_L \to \mathbb{R}$ given by

$$\Phi(S) := \int_{\mathbb{S}^n} S(\eta)d\mu(\eta),\ S \in \Pi_L.$$

Let $\Phi_X(S_X) = \Phi(T_X^{-1}(S_X)) = \Phi(S)$. If $S_X \ge 0$, then $|S(\xi)| = S(\xi)$ for $\xi \in X$, and so from (36) we have that

$$\big|\Phi(S) - \sum_{\xi\in X} S(\xi)\mu(R_\xi)\big| \le \big|\|S\|_1 - \sum_{\xi\in X} S(\xi)\mu(R_\xi)\big| \le \frac{\delta}{1-\delta}\sum_{\xi\in X} S(\xi)\mu(R_\xi),$$

provided only that $\|\mathcal{X}\| \le \delta s_n^{-1}(L+\lambda_n)^{-1}$. For any $\delta < \frac{1}{2}$, this implies that

$$\frac{1-2\delta}{1-\delta}\sum_{\xi\in X} S(\xi)\mu(R_\xi) \le \Phi(S) \le \frac{1}{1-\delta}\sum_{\xi\in X} S(\xi)\mu(R_\xi).$$

From this, we see that the linear functional

$$\Psi_X(S_X) := \Phi_X(S_X) - \frac{1-2\delta}{1-\delta}\sum_{\xi\in X} S(\xi)\mu(R_\xi) \tag{38}$$

is positive on the cone $0 \le S_X \in V_L$, which itself is contained in the positive cone of $\mathbb{R}^{|X|}$.

There are two facts we will take account of. The first is that the positive cone of V_L is contained in the positive cone of $\mathbb{R}^{|X|}$. The second is that the vector $(1)_{\xi\in X}$, which is in both cones, is an interior point of the positive cone of $\mathbb{R}^{|X|}$. By the Krein–Rutman theorem [9], there exists a positive linear functional $\widetilde{\Psi}_X$ that extends Ψ_X to all $\mathbb{R}^{|X|}$. Consequently, there exist weights $\alpha_\xi \ge 0$ such that $\widetilde{\Psi}_X(x) = \sum_{\xi\in X}\alpha_\xi x_\xi$. Using this and $\Phi_X(S_X) = \Phi(S)$ in (38), we obtain

$$\Phi(S) = \sum_{\xi\in X} c_\xi S(\xi), \quad c_\xi := a_\xi + \frac{1-2\delta}{1-\delta}\mu(R_\xi),\ a_\xi \ge 0. \tag{39}$$

This is of course a positive-weight quadrature formula on $\mathbb{S}^n$, with weights bounded below by $\frac{1-2\delta}{1-\delta}\mu(R_\xi)$.

We want to get upper bounds as well. To do that, we let $L' = \lfloor \frac{L}{2} \rfloor$ and fix $\xi_0 \in X$. If $S \in \Pi_{L'}$, then S^2 is in Π_L. The quadrature formula (39) then implies that

$$\|S\|_2^2 = \Phi(S^2) = \sum_{\xi \in X} c_\xi (S(\xi))^2 \geq c_{\xi_0} (S(\xi_0))^2.$$

Choose $S(\eta) = \sum_{\ell=0}^{L'} \sum_{m=1}^{d_\ell^n} Y_{\ell,m}(\eta) \overline{Y_{\ell,m}(\xi_0)} = \sum_{\ell=0}^{L'} \frac{\ell+\lambda_n}{\omega_n \lambda_n} P_\ell^{(\lambda_n)}(\xi_0 \cdot \eta)$, which is real valued. Using the orthogonality of the $Y_{\ell,m}$'s, one can show that $\|S\|_2^2 = S(\xi_0) = \sum_{\ell=0}^{L'} \frac{\ell+\lambda_n}{\omega_n \lambda_n} \binom{\ell+n-2}{\ell}$. From the previous inequality, (13) and (4), and the fact that $\dim \Pi_{L'} = d_{L'}^{n+1}$ [17, p. 4], we get $c_{\xi_0} \leq \omega_n / d_{L'}^{n+1}$, where $L' := \lfloor L/2 \rfloor$. We summarize these results below.

THEOREM 4.3. *Adopt the notation of Theorem* 4.2. *In particular,* s_n *is given by* (37) *and depends only on* n. *For any* $0 < \delta < \frac{1}{2}$ *and any integer* $L > 0$, *if* $\|\mathcal{X}\| \leq \delta s_n^{-1} (L + \lambda_n)^{-1}$, *then there exist positive weights* c_ξ, $\xi \in X$, *such that the quadrature formula*

$$\int_{\mathbb{S}^n} f(\eta) d\mu(\eta) \doteq \sum_{\xi \in X} c_\xi f(\xi) \tag{40}$$

is exact for spherical harmonics in Π_L. *Also, the weights satisfy the bounds*

$$\frac{1-2\delta}{1-\delta} \mu(R_\xi) \leq c_\xi \leq \frac{\omega_n}{d_{L'}^{n+1}}, \ L' = \lfloor L/2 \rfloor. \tag{41}$$

The theorem just proved starts with L and puts conditions on the decomposition $\mathcal{X}$. The centers in X play a secondary role, serving as labels for regions in $\mathcal{X}$ and as evaluation points in the quadrature formula.

It's useful to turn this around and have the centers X play the primary role. To do that, we need to make the assumption that we are considering only ρ-uniform X; that is, for some fixed ρ we assume that the mesh ratio $h_X/q_X = \rho_X \leq \rho$. We will take the $\mathcal{X} = \mathcal{X}_V$ to be the Voronoi decomposition associated with X. For this decomposition, we have $h_X \leq \|\mathcal{X}_V\|$. Also, since the smallest distance between two points in X is $2q_X$, every $R_\xi \in \mathcal{X}_V$ contains a spherical cap with center ξ and radius $q_X \geq h_X/\rho$; hence, $\mu(R_\xi) \geq \omega_{n-1}(2/\pi)^{n-1} \rho^{-n} h_X^n / n$. Applying Theorem 4.3, we arrive at this result.

COROLLARY 4.4. *Adopt the notation of Theorem* 4.3 *and let* X *be a* ρ*-uniform set of centers. If* $h_X \leq \delta s_n^{-1} (L + \lambda_n)^{-1}$, *then the quadrature formula* (40) *holds with weights satisfying*

$$\omega_{n-1}(2/\pi)^{n-1} \left(\frac{1-2\delta}{1-\delta} \right) \rho^{-n} h_X^n \leq c_\xi \leq \frac{\omega_n}{d_{L'}^{n+1}}, \qquad L' = \lfloor L/2 \rfloor. \tag{42}$$

Set $\delta = 1/4$. To get a better idea of how the weights are bounded in terms of $h = h_X$ or L, note that by (4) we have $d_{L'}^{n+1} \sim \frac{(L/2)^n}{\lambda_{n+1}(n-1)!}$. In addition, if we take L as large as possible, but still consistent with the condition that $h_X \leq \delta s_n^{-1} (L + \lambda_n)^{-1}$, then $L \sim h^{-1}$. In that case, we see that

$$c_\xi = \mathcal{O}\{h^n\} = \mathcal{O}\{L^{-n}\}, \tag{43}$$

where the constants hidden by $\mathcal{O}$ are dependent only on the dimension n.

So far we have only addressed the *existence* of positive weights, along with bounds on them. In fact, the existence of such weights implies the *feasibility* of solving a quadratic programming problem that produces weights minimizing $\sum_{\xi\in X} c_\xi^2$, subject to constraints. Thus it is possible to numerically compute the weights. For more details, see [14, section 4.3].

5. Tight frames on $\mathbb{S}^n$. In this section, we discuss three important features of the operator frames on $\mathbb{S}^n$ introduced earlier in section 2.2. The first is the approximation power of these frames in various spaces. The second is how to turn them into *tight* frames for $\mathbb{S}^n$. This requires discretizing them using the quadrature results from the previous section. The third and final feature is their excellent localization properties.

We will turn to discussing the approximation power of these operator frames, after a brief word about notation. Throughout this section, the operators A_j and B_J are their kernels A_j and B_J, which are defined in section 2.2. The function $b(t)$ is defined in (1). We assume that the function $a(t)$, whose properties are discussed in section 1, is in $C^k(\mathbb{R})$.

PROPOSITION 5.1. *Let $k > \max\{n, 2\}$, and let b be defined by* (1)*, with $a \in C^k(\mathbb{R})$. If $f \in L^p(\mathbb{S}^n)$, $1 \le p \le \infty$, and if $L > 0$ is an integer such that $2^{-J-j_n} \le (L+\lambda_n)^{-1}$, then*

$$\|f - \mathsf{B}_J f\|_p \le C_{b,k,n} E_L(f)_p, \qquad E_L(f)_p := \operatorname{dist}_{L^p}(f, \Pi_L). \tag{44}$$

Also, for $1 \le p < \infty$ or, if $p = \infty$, for $f \in C(\mathbb{S}^n)$, we have $\lim_{J\to\infty} \mathsf{B}_J f = f$.

Proof. Apply Corollary 3.10 with $\kappa = b$, k as above, and $\varepsilon = 2^{-J-j_n}$. ∎

The proposition implies that $\mathsf{B}_J f$ approximates f to within an error comparable to $E_L(f)_p$, which is that for the best approximation to f from Π_L in L^p. Much work [11, 20, 22, 23, 27] has been done on estimating this error for various smoothness classes and spaces. This work allows us to obtain rates of approximation when f has additional smoothness requirements. A typical result [11] is this: If $f \in L^p(\mathbb{S}^n)$, with $\|f\|_p = 1$, belongs to a smoothness class $W_p^\alpha(\mathbb{S}^n)$, which is analogous to a Sobolev space, then $E_L(f)_p \sim L^{-\alpha}$. Choosing f similarly and taking $L \sim 2^J$, we get a corresponding result for our case: $\|f - \mathsf{B}_J f\|_p \sim 2^{-\alpha J}$.

We now turn to constructing tight frames on $\mathbb{S}^n$. The quadrature formulas from section 4.2 will play a pivotal role in their construction; we will also require a sequence of sets of centers to use in conjunction with them. Let $\rho \ge 2$ be fixed. By Proposition 2.1, we can find a sequence of sets of centers $\{X_j \in \mathcal{F}_\rho\}_{j=0}^\infty$ such that X_j is nested and such that the mesh norm $h_j := h_{X_j}$ halves going from j to $j+1$; that is, $h_{j+1} \le h_j/2$. In what follows, assume that the X_j's form such a sequence.

Recall that on $\mathbb{S}^n$, the frame transform $f \to w_j = \widetilde{\mathsf{A}}_j^* f$ takes the form $w_j(\eta) = \mathsf{A}_j^* f(\eta) = \langle f(\zeta), A_j(\zeta\cdot\eta)\rangle_{L^2(\mathbb{S}^n)}$. Because $A_j(\zeta\cdot\eta)$ is a spherical polynomial with degree less than 2^{j+j_n+1}, the function $w_j(\eta)$ is a spherical polynomial of degree less than 2^{j+j_n+1}. In the reconstruction formula this then contributes the term

$$\mathsf{A}_j w_j(\omega) = \int_{\mathbb{S}^n} A_j(\omega\cdot\eta) w_j(\eta)\, d\mu(\eta).$$

The product $A_j(\omega\cdot\eta)w_j(\eta)$ is a spherical polynomial having degree less than $2^{j+j_n+1} + 2^{j+j_n+1} = 2^{j+j_n+2}$.

We can integrate this *exactly* using the quadrature formula (40), with $L = 2^{j+j_n+2}$. First of all, the condition on the mesh norm h in both Theorem 4.3

and Corollary 4.4 is that $h \le \delta s_n^{-1}(L+\lambda_n)^{-1}$, where $\delta \in (0,1/2)$ is arbitrary. Choose $\delta = 1/4$ to be definite. For $n = 1$ (the circle), we have $\lambda_1 = 0$ and $j_1 = 0$, and the condition is $h \le \delta s_1^{-1} 2^{-j-2} = s_1^{-1} 2^{-j-4}$. For $n \ge 2$, note that $2^{j+j_n+2} + \lambda_n \le 2^{j+2}\lfloor \lambda_n \rfloor + \lambda_n < 2^{j+3}\lambda_n$. The condition for $n \ge 2$ is then fulfilled if $h \le \delta(\lambda_n s_n)^{-1} 2^{-j-3} = (\lambda_n s_n)^{-1} 2^{-j-5}$. It is clear that these conditions can be met by using the sets X_j.

Let the quadrature weight corresponding to the center $\xi \in X_j$ be denoted by $c_{j,\xi}$, so that

$$\mathsf{A}_j w_j(\omega) = \sum_{\xi \in X_j} c_{j,\xi} A_j(\xi \cdot \omega) w_j(\omega) = \sum_{\xi \in X_j} \langle f, \psi_{j,\xi} \rangle \psi_{j,\xi}, \tag{45}$$

where

$$\psi_{j,\xi}(\eta) := \sqrt{c_{j,\xi}}\, A_j(\eta \cdot \xi),\ \xi \in X_j, \tag{46}$$

is the analysis frame function at level j. The frame function $\psi_{j,\xi}$ is computable: A_j is known and, as we noted at the end of section 4.2, the weights can be found numerically. We can now prove this result.

THEOREM 5.2. *Let $k > \max\{n, 2\}$, and let A_j be the kernel in (8), with $a \in C^k(\mathbb{R})$. If $f \in C(\mathbb{S}^n)$ or, for $1 \le p < \infty$, if $f \in L^p(\mathbb{S}^n)$, then*

$$f = \sum_{j=0}^{\infty} \sum_{\xi \in X_j} \langle f, \psi_{j,\xi} \rangle \psi_{j,\xi},$$

with convergence being in the appropriate space. In addition, if $f \in L^2(\mathbb{S}^n)$, the frame $\{\psi_{j,\xi}\}_{j \in \mathbb{Z}_+, \xi \in X_j}$ is tight:

$$\|f\|^2 = \begin{cases} \frac{1}{2\pi} |\langle f, 1 \rangle|^2 + \sum_{j=0}^{\infty} \sum_{\xi \in X_j} |\langle f, \psi_{j,\xi} \rangle|^2, & n = 1, \\ \sum_{j=0}^{\infty} \sum_{\xi \in X_j} |\langle f, \psi_{j,\xi} \rangle|^2, & n \ge 2. \end{cases}$$

Finally, the frame functions have vanishing moments that increase with j, and are orthogonal on nonadjacent levels.

Proof. From (9) and (45), for $n \ge 2$ we get $\mathsf{B}_J f = \sum_{j=0}^{J} \sum_{\xi \in X_j} \langle f, \psi_{j,\xi} \rangle \psi_{j,\xi}$. By Proposition 5.1 this converges to f in all of the spaces mentioned. To prove that the frame is tight, just observe that for $f \in L^2(\mathbb{S}^n)$, we have $\langle \mathsf{B}_J f, f \rangle = \sum_{j=0}^{J} \sum_{\xi \in X_j} \left| \langle f, \psi_{j,\xi} \rangle \right|^2$. Taking the limit as $J \to \infty$ then yields the equation for $\|f\|^2$. The statement concerning vanishing moments follows from the structure of the A_j's, and the orthogonality between nonadjacent levels is proved in Proposition 1.1. The $n = 1$ case has a projection P_0 in B_J, where P_0 projects onto the constants. The effect of this is to add a term to the series for $\|f\|^2$. ∎

Our last result concerns the localization properties of the frame function defined by (46).

COROLLARY 5.3. *Let $k > \max\{n, 2\}$ and let $\psi_{j,\xi}$ be given by (46). If $\theta := \cos^{-1}(\eta \cdot \xi)$, then for all $\theta \in [0, \pi]$ there are constants C and C', which depend on k, n, and a, such that these hold:*

$$|\psi_{j,\xi}(\eta)| \le \frac{2^{n(j+j_n)/2}\, C}{1 + (2^{j+j_n}\theta)^k} \text{ and } |B_J(\eta \cdot \xi)| \le \frac{2^{n(J+j_n)}\, C'}{1 + (2^{J+j_n}\theta)^k}.$$

Proof. Use Theorem 3.5, with $\kappa = b$ and $\varepsilon = 2^{-J-j_n}$, to bound $B_J(\xi \cdot \eta)$, and again, with $\kappa = a$ and $\varepsilon = 2^{-j-j_n}$, to bound $A_j(\eta \cdot \xi)$. Next, use $L = 2^{j+j_n+2}$ in (43) to see that $c_\xi = \mathcal{O}\{2^{-(j+j_n)n}\}$, where the constants depend only on n. To bound $\psi_{j,\xi}$, use the bounds on A_j and c_ξ in (46). □

REFERENCES

[1] O. Christensen, *An Introduction to Frames and Riesz Bases*, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, 2003.

[2] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[3] Q. Du, M. D. Gunzburger, and L. Ju, *Constrained centroidal tessellations for surfaces*, SIAM J. Sci. Comput., 24 (2003), pp. 1488–1506.

[4] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341–366.

[5] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.

[6] M. Frazier and B. Jawerth, *A discrete transform and decomposition of distribution spaces*, J. Funct. Anal., 93 (1990), pp. 34–170.

[7] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces*, CBMS Reg. Conf. Ser. Math. 79, AMS, Providence, RI, 1991.

[8] W. Freeden, T. Gervens, and M. Schreiner, *Constructive Approximation on the Sphere: With Applications to Geomathematics*, Numer. Math. Sci. Comput., The Clarendon Press, Oxford University Press, New York, 1998.

[9] R. B. Holmes, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, New York, 1975.

[10] S. Jaffard, Y. Meyer, and R. D. Ryan, *Wavelets. Tools for Science and Technology*, revised ed., SIAM, Philadelphia, 2001.

[11] A. I. Kamzolov, *The best approximation of classes of functions $W_p^\alpha(S^n)$ by polynomials in spherical harmonics*, Mat. Zametki, 32 (1982), pp. 285–293 (in Russian); 32 (1983), pp. 622–626 (in English).

[12] Y. Meyer, *Ondeletes et Opérateurs* I : *Ondeletes*, Hermann, Paris, 1990.

[13] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, *Representing and analyzing scattered data on spheres*, in Multivariate Approximation and Applications, N. Dyn, D. Leviaton, D. Levin, and A. Pinkus, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 44–72.

[14] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, *Spherical Marcinkiewicz–Zygmund inequalities and positive quadrature*, Math. Comp., 70 (2001), pp. 1113–1130.

[15] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, *Corrigendum to "Spherical Marcinkiewicz–Zygmund inequalities and positive quadrature*," Math. Comp., 71 (2001), pp. 453–454.

[16] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, *Zonal function network frames on the sphere*, Neural Networks, 16 (2003), pp. 183–203.

[17] C. Müller, *Spherical Harmonics*, Lecture Notes in Math. 17, Springer-Verlag, Berlin, 1966.

[18] F. J. Narcowich, P. Petrushev, and J. D. Ward, *Decomposition of Besov and Triebel–Lizorkin spaces on the sphere*, J. Funct. Anal., to appear, 2006.

[19] F. J. Narcowich, X. Sun, J. D. Ward, and H. Wendland, *Direct and inverse Sobolev error estimates for scattered-data interpolation via spherical basis functions*, Found. Comput. Math., to appear, 2006.

[20] S. M. Nikol'skiĭ and P. I. Lizorkin, *Approximation of functions on a sphere*, Izv. Akad. Nauk SSSR Ser. Mat., 51 (1987), pp. 635–651 (in Russian); translation in Math. USSR-Izv., 30 (1988), pp. 599–614 (in English).

[21] H.-S. Oh and T.-H. Li, *Estimation of global temperature fields from scattered observations by a spherical-wavelet-based spatially adaptive method*, J. R. Stat. Soc. Ser. B Stat. Methodol., 66 (2004), pp. 221–238.

[22] S. Pawelke, *Über die Approximationsordnung bei Kugelfunktionen und algebraischen Polynomen*, Tôhoku Math. J., 24 (1972), pp. 473–486.

[23] Kh. P. Rustamov, *On the approximation of functions on a sphere*, Russian Acad. Sci. Izv. Math., 43 (1994), pp. 311–329.

[24] E. M. Stein and G. Weiss, *Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[25] G. Szegö, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. 23, AMS, Provi-

dence, RI, 1975.
[26] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, UK, 1965.
[27] Y. Xu, *Weighted approximation of functions on the unit sphere*, Constr. Approx., 21 (2005), pp. 1–28.
[28] A. Zygmund, *Trigonometric Series*, Vol. II, Cambridge University Press, Cambridge, UK, 1988.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 595–616

COUPLING CONDITIONS FOR A CLASS OF SECOND-ORDER MODELS FOR TRAFFIC FLOW*

M. HERTY† AND M. RASCLE‡

Abstract. This paper deals with a model for traffic flow based on a system of conservation laws [A. Aw and M. Rascle, *SIAM J. Appl. Math.*, 60 (2000), pp. 916–938]. We construct a solution of the Riemann problem at an arbitrary junction of a road network. Our construction provides a solution of the full system. In particular, all moments are conserved.

Key words. hyperbolic systems of conservation laws, traffic flow, road networks, homogenization

AMS subject classifications. 35LXX, 35L6

DOI. 10.1137/05062617X

1. Introduction.

Macroscopic modeling of vehicular traffic started with the work of Lighthill and Whitham (LWR) [25]. Since then there has been intense discussion and research; see [26, 10, 2, 19, 21, 22, 6, 24] and the references therein. Today, fluid dynamic models for traffic flow are appropriate for describing traffic phenomena such as congestion and stop-and-go waves [18, 14, 20]. The case of road networks based on the LWR model has been considered in particular in [17, 5, 16]. Recently in [12], Garavello and Piccoli considered a road network based on the Aw–Rascle (AR) model [2] of traffic flow. Here, in contrast to [12], we propose a modeling of the junctions conserving the mass *and* the pseudo-"momentum" $\rho v\, w$. We will discuss below further differences between the two modelings.

We consider a finite directed graph as a model for a road network with unidirectional flow. Each road $i = 1, \dots, \mathcal{I}$ is modeled by an interval $I_i := [a_i, b_i] \subset \mathbb{R}$ possibly with $a_i = -\infty$ or $b_i = \infty$. Each vertex of the graph corresponds to a junction. For a fixed junction k the set δ_k^- contains all road indices i which are incoming roads, so that $\forall i \in \delta_k^- : b_i = k$. Similarly, δ_k^+ denotes the indices of outgoing roads: $\forall j \in \delta_k^+ : a_j = k$. We skip the index k whenever the situation is clear.

The evolution of $\rho_i(x,t)$ and $v_i(x,t)$ on each road i is given by the AR model [2]

$$\partial_t \rho_i + \partial_x(\rho_i v_i) = 0, \tag{1.1a}$$

$$\partial_t(\rho_i w_i) + \partial_x(\rho_i v_i w_i) = 0, \tag{1.1b}$$

$$w_i = v_i + p_i(\rho_i), \tag{1.1c}$$

where for each i, $\rho \mapsto p_i(\rho)$ is a known function ("traffic pressure") with the following properties:

$$\forall \rho : \rho p_i''(\rho) + 2p_i'(\rho) > 0 \text{ and } p_i(\rho) \sim \rho^\gamma \text{ near } \rho = 0 \tag{1.2}$$

*Received by the editors March 7, 2005; accepted for publication (in revised form) December 12, 2005; published electronically June 23, 2006. This research was partially supported by European Union financed network grant HPRN-CT-2002-00282, by the Kaiserslautern Excellence Cluster "Dependable Adaptive Systems and Mathematical Modelling," and by French ACI-NIM (Nouvelles Interactions des Mathématiques) grant 193 (2004).

http://www.siam.org/journals/sima/38-2/62617.html

†Fachbereich Mathematik, TU Kaiserslautern, D-67653 Kaiserslautern, Germany (herty@rhrk.uni-kl.de).

‡Laboratoire J. A. Dieudonné, Université de Nice, F-06108 Nice Cedex 2, France (rascle@math.unice.fr).

and where $\gamma > 0$. The conservative form of (1.1) is

$$\partial_t \begin{pmatrix} \rho_i \\ y_i \end{pmatrix} + \partial_x \begin{pmatrix} y_i - \rho_i p_i(\rho_i) \\ (y_i - \rho_i p_i(\rho_i)) y_i/\rho_i \end{pmatrix} = 0,$$

where $y_i = \rho_i w_i = \rho(v_i + p_i(\rho_i))$. Since w_i and v_i are related by (1.1), we choose to describe solutions in terms of ρ_i and $\rho_i v_i$. For a motivation and a complete discussion of these equations we refer to section 2 and [2], respectively.

We consider weak solutions of the network problem as in [17]: Let a set $i = 1, \dots, \mathcal{I}$ of smooth functions $\phi_i : [0, +\infty] \times I_i \to \mathbb{R}^2$ having compact support in $I_i = [a_i, b_i]$, which are "smooth" across each junction k, be given, i.e.,

$$\phi_i(b_i) = \phi_j(a_j) \quad \forall i \in \delta_k^-, \quad \forall j \in \delta_k^+. \tag{1.3}$$

Then a set of functions

$$U_i = (\rho_i, \rho_i v_i), \quad i = 1, \dots, \mathcal{I}, \tag{1.4}$$

is called a weak solution of (1.1) if and only if equations (1.5) hold for all families of test functions $\{\phi_i\}_{i \in \mathcal{I}}$ with the property (1.3).

$$\sum_{i=1}^{\mathcal{I}} \int_0^\infty \int_{a_i}^{b_i} \begin{pmatrix} \rho_i \\ \rho_i w_i \end{pmatrix} \cdot \partial_t \phi_i + \begin{pmatrix} \rho_i v_i \\ \rho_i v_i w_i \end{pmatrix} \cdot \partial_x \phi_i dx dt$$

$$- \int_{a_i}^{b_i} \begin{pmatrix} \rho_{i,0} \\ \rho_{i,0} w_{i,0} \end{pmatrix} \cdot \phi_i(x, 0) dx = 0, \tag{1.5a}$$

$$w_i(x,t) = v_i(x,t) + p_i^\dagger(\rho_i(x,t)). \tag{1.5b}$$

Here, $U_{i,0}(x) = (\rho_{i,0}(x), (\rho_{i,0} v_{i,0})(x))$ are the initial data. *The functions $p_i^\dagger(\cdot)$ are initially unknown.* The explicit form of each $p_i^\dagger$ depends on the initial data and the type of junction. Near any junction k the function $p_i^\dagger$ is equal to p_i for all incoming roads. The same is true for all outgoing roads of the junction if there is only *one* incoming road. This is discussed in sections 3 and 4. In section 6 we discuss the case where $p_i^\dagger \neq p_i$ and give arguments for the necessity of introducing $p_i^\dagger$. At this point let us just note that in the general case $p_i^\dagger$ depends on a mixture of the incoming flows.

In the case of a single junction we derive from (1.5a), (1.5b) the Rankine–Hugoniot conditions for piecewise smooth solutions:

$$\sum_{i \in \delta^-} (\rho_i v_i)(b_i^-, t) = \sum_{i \in \delta^+} (\rho_i v_i)(a_i^+, t), \tag{1.6a}$$

$$\sum_{i \in \delta^-} (\rho_i v_i w_i)(b_i^-, t) = \sum_{i \in \delta^+} (\rho_i v_i w_i)(a_i^+, t). \tag{1.6b}$$

Properties (1.6a) and (1.6b) correspond to conservation of mass and of (pseudo-) "momentum." We remark that the solution constructed in [12] does *not* conserve the (pseudo-)"momentum" (see Proposition 2.3 in [12]) and therefore is *not* a weak solution in the sense of (1.5a), (1.6a), and (1.6b).

In the next section we discuss the construction of weak solutions in the sense of (1.5) for initial data constant on each road:

$$(\rho_{i,0}, \rho_{i,0} v_{i,0}) = U_{i,0} = \text{const}_i. \tag{1.7}$$

We consider a single junction. We look for solutions to Riemann problems on each road i as if the road were extended to $]-\infty, \infty[$:

$$\partial_t \begin{pmatrix} \rho_i \\ \rho_i w_i \end{pmatrix} + \partial_x \begin{pmatrix} \rho_i v_i \\ \rho_i v_i w_i \end{pmatrix} = 0, \qquad U_i(x,0) = \begin{pmatrix} U^- & x < x_0 \\ U^+ & x > x_0 \end{pmatrix}. \tag{1.8}$$

Depending on the road, only one of the Riemann data is defined for $t = 0$:

$$\text{If } i \in \delta^- : U^- = U_{i,0}\,,\ x_0 = b_i \ \text{ and if } i \in \delta^+ : U^+ = U_{i,0}\,,\ x_0 = a_i. \tag{1.9}$$

We construct an (entropy) solution to (1.5) such that all generated waves have nonpositive ($i \in \delta^-$) or nonnegative ($i \in \delta^+$) speed. Moreover, the solutions satisfy conditions (1.6a) and (1.6b).

We have to impose additional conditions [12] to obtain a unique solution. First, the flux ρv is nonnegative. Next, it has to be distributed according to a priori given ratios; see sections 3 to 7 for further details. Finally, we require that the *total flux be maximized* subject to the other conditions.

The paper is organized as follows. In section 2 we discuss the general properties of the Riemann problem for (1.1). First, we construct the demand and supply functions, which are necessary to determine the flux at the junction. Refer to [23, 8, 9] for the presentation of supply and demand functions for first-order models. Next, we define admissible states on each road at the junction and finally we construct all intermediate states in the solution of (1.1).

In section 3 we consider the easiest possible situation, namely, two roads connected by a junction. In section 4 we extend the results to a junction with one incoming and two outgoing roads. For the results on two incoming and one outgoing road we need a description of the mixture of flows on the outgoing road. Therefore we briefly revisit the main results of [1] and [3] in section 5. In section 6 we solve the case of two incoming and two outgoing roads and define homogenized flow. In section 7 we consider the general case of an intersection with an arbitrary number of incoming and outgoing roads.

2. Preliminary discussion. The conservative variables are ρ_i and $y_i := \rho_i w_i$. We assume $\forall i : 0 \le \rho_i \le \rho_{\max} = 1$ and $\forall i : 0 \le v_i \le v_{\max} = 1$. Furthermore, we set

$$U_i := (\rho_i, \rho_i v_i),\ U := (\rho, \rho v) \tag{2.1}$$

and we skip the subindex i at ρ_i and v_i whenever the intention is clear. The system (1.1) is strictly hyperbolic if $\rho_i > 0$ for all i. The eigenvalues are

$$\lambda_{1,i}(U) = v - \rho p_i'(\rho) \quad \text{ and } \quad \lambda_{2,i}(U) = v. \tag{2.2}$$

The right eigenvectors corresponding to $\lambda_{1,i}$ and $\lambda_{2,i}$ are

$$\mathbf{r}_{1,i} = \begin{pmatrix} 1 \\ -p_i'(\rho) \end{pmatrix} \quad \text{ and } \quad \mathbf{r}_{2,i} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Let ∇ denote the gradient with respect to (ρ, v). We recall that k is called a genuinely nonlinear characteristic family if $\nabla \lambda_{k,i}(\rho, v) \cdot \mathbf{r}_{k,i}(\rho, v) \neq 0\ \forall (\rho, v)$. Depending on

the initial data, the associated waves are rarefaction or shock waves. If $\nabla\lambda_{k,i}(\rho, v) \cdot \mathbf{r}_{k,i}(\rho, v) = 0\ \forall(\rho, v)$, then k is called a linearly degenerated characteristic family and the associated waves are contact discontinuities. We refer to Definitions 7.2.1 and 7.5.1 in [7] for more details.

Here, $k = 1$ is a genuinely nonlinear and $k = 2$ is a linearly degenerated characteristic family for all roads i. Moreover, the 1-shock and 1-rarefaction curves coincide and we have a 2-contact discontinuity; see [2]. For each road i the Riemann invariants are

$$w_i(U) = v + p_i(\rho) \quad \text{and} \quad v_i(U) = v. \tag{2.3}$$

Let us be more specific about the physical interpretation of w and $p(\cdot)$. Other descriptions than (2.3) could be envisioned. In particular, the *additive* role of $p_i(\cdot)$ in w_i (as in the Payne–Whitham model [26]) is not essential. It was introduced in [2] for "historical" reasons, but it has a drawback: The associated individual fundamental diagram (see Figure 1 below) implies a zero speed at a maximal (jam) traffic density which is *different* for each category of car-driver pairs, i.e., each pairing (w_i, p_i). We keep the above expression (2.3) throughout the paper for the sake of simplicity. As noted in [3], the only crucial property of w_i is that it is a *Lagrangian* marker. As an example assume that on each road i, the (pseudo-)pressure is $p_i(\rho) := v_{max} - V_i(\rho)$, where, e.g., v_{max} is the maximal speed on all roads and $V_i(\rho)$ is an *equilibrium* speed on road i. Therefore, the *function* $U := (\rho, v) \to w_i(U) = v + p_i(\rho)$ describes the *distance to equilibrium*. The "momentum" equation tells us that each value w is a Lagrangian property, such as a label or a color. Hence, when passing from road i to another road j, each driver will preserve its "color." In other words, he will keep the *same value* w, which will now satisfy

$$w_j(U) = w = w_i(U).$$

This simple observation will be essential in what follows. In particular, it will lead to a very natural homogenization problem in section 6.

The classical description by *first*-order models is just a *particular case* of our second-order model. It corresponds to setting all the w's equal to the same constant. So our description can be drastically simplified when no sophisticated information is needed.

We return to the mathematical description. Usually, we draw the level curves of the Riemann invariants (in short the Riemann invariants) in the $(\rho, \rho v)$ plane. An example of the curves is depicted in Figure 1. There is a one-to-one correspondence to the (ρ, y) plane; see [2].

For an arbitrary fixed i we discuss the shape of the Riemann invariants in the $(\rho, \rho v)$ plane and characterize important points.

The Riemann invariant $\{v_i(U) = c\}$ is a straight line with slope c passing through the origin. Consider the curve $\{w(U) := w_i(U) = c\}$, where $c \in \mathbb{R}$ denotes a constant. By assumption (1.2) on $p := p_i$ this curve is strictly concave and passes through the origin. Furthermore, if $c > 0$, then the curve $\{w(U) = c\}$ lies in the first quadrant of the $(\rho, \rho v)$ plane for ρ between 0 and a maximal value $\bar{\rho} \in]0, 1]$. The maximal value $\bar{\rho}$ depends on c and $p(\cdot)$. Due to the strict concavity there exists a unique point (i.e., the "sonic point") $\sigma(w, c)$ with $0 < \sigma(w, c) \leq 1$, depending on c and the function $p(\cdot)$. The point $\sigma(w, c)$ maximizes the flux ρv on $\{w(U) = c\}$.

The total flux has to be conserved through an intersection. Therefore, we introduce the functions $r(\rho; w, c)$ and $u(\rho; w, c)$ below. Assume $c > 0$. Then for all

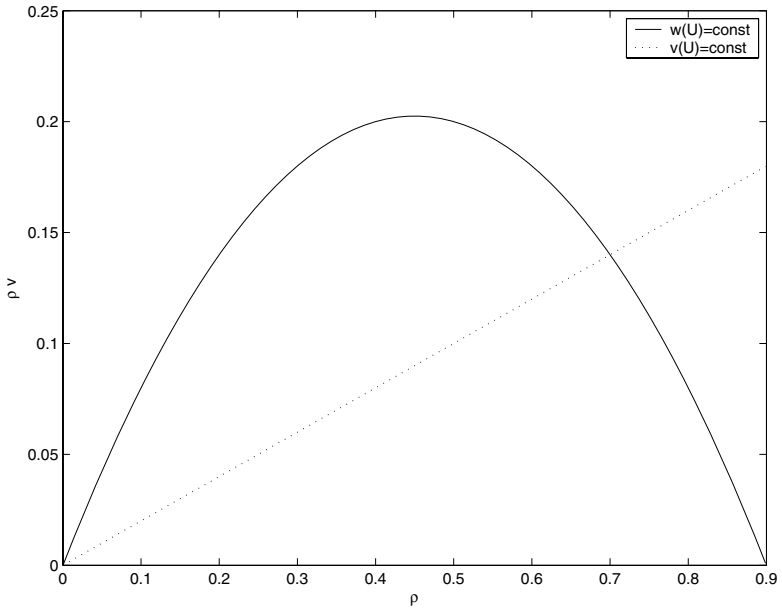


FIG. 1. *Riemann invariants in the* $(\rho, \rho v)$ *plane.*

$\rho \in [0, \bar{\rho}]$ there exists a unique v such that $w((\rho, \rho v)) = c$. Moreover, there exists a unique pair (r, u) such that

$$w(r, r\ u) = w(\rho, \rho\ v), \tag{2.4a}$$

$$r\ u = \rho\ v, \tag{2.4b}$$

$$r \neq \rho \text{ except for } \rho = \sigma(w, c). \tag{2.4c}$$

In other words (ρ, v) and (r, u) correspond to the same flux and the same level curve of w; see Figure 2 for an example. Hence, for each curve $\{w(U) = c\}$ with $c > 0$ there exist two unique functions $\rho \to r(\rho; w, c)$ and $\rho \to u(\rho; w, c)$ satisfying (2.4) for all $\rho \in [0, \bar{\rho}]$.

Next, we describe the construction of the demand and supply functions for a given curve $\{w(U) = c\}$, $c \geq 0$. As in the case of first-order models, e.g., [23], in the $(\rho, \rho v)$ plane, the demand function $d(\rho; w, c)$ is an extension of the *nondecreasing* part of the curve $\{w(U) = c\}$ for $\rho \geq 0$, whereas the supply function $s(\rho; w, c)$ is an extension of the *nonincreasing* part of the curve $\{w(U) = c\}$ and $\rho \geq 0$; see Figures 2 and 3 for examples.

Now we consider the Riemann problem (1.8) for a given *incoming* road $i \in \delta^-$. Hence, only the initial datum $U^- = U_{i,0}$ is given. We want to determine all "admissible" states U^+: A state U^+ is called "admissible" if and only if either the waves of the solution to (1.8) with initial data (U^-, U^+) have negative speed or the solution is constant $(U^+ = U^-)$. As in [17] we neglect waves of zero speed (stationary waves). Later on U^+ will be an intermediate state in the solution $U_i(\cdot, \cdot)$ on the incoming road i for the full Riemann problem at the junction, i.e., $U_i(x_0-, t) = U^+$.

PROPOSITION 2.1. *Let* $U^- = (\rho^-, \rho^- v^-) \neq (0, 0)$ *be the initial value on an incoming road* i. *Let the* 1*-curve through* U^- *be* $w_i(U) = v + p_i(\rho) = w^-$ *with* $w^- := w_i(U^-)$. *Then the "admissible" states* $U^+ = (\rho^+, \rho^+ v^+)$ *for the Riemann problem*

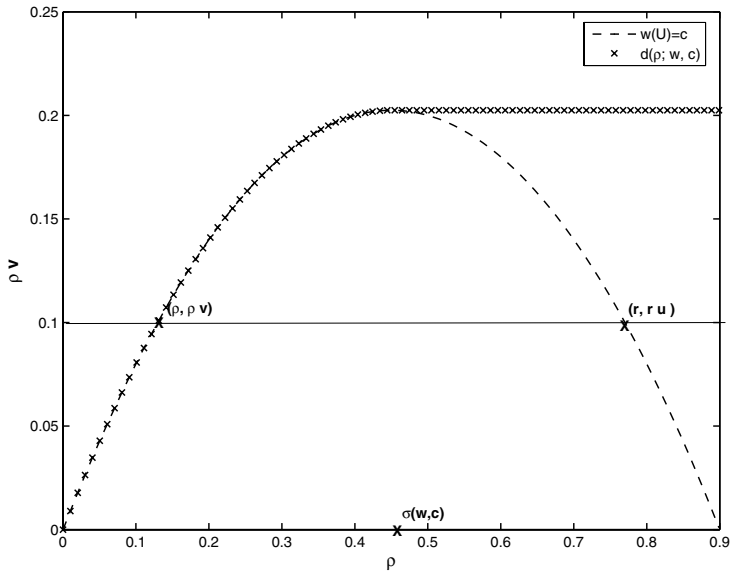


FIG. 2. *Demand function for given* $w(U) = v + p(\rho) = \text{const}$. *Additionally,* $\sigma(w, c)$ *and the position of a sample point* $(\rho, \rho\, v)$ *and the corresponding* $(r, r\, u)$ *are shown.*

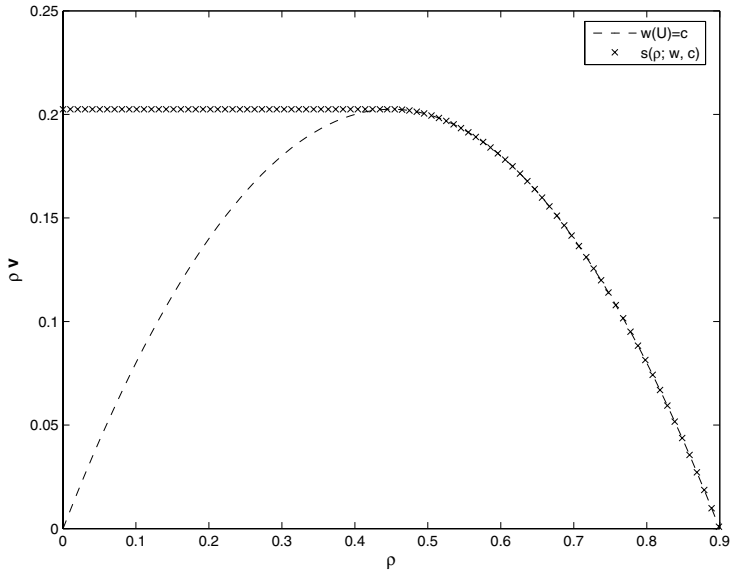


FIG. 3. *Supply function for a given* $w(U) = v + p(\rho) = \text{const}$.

must belong to that curve; i.e., $w_i(U^+) = w^-$ *and* $\rho^+ v^+ \geq 0$. *Depending on* U^- *we distinguish two cases:*

1. $\rho^- < \sigma(w_i, w^-)$: U^+ *is admissible if and only if* $\rho^+ > r(\rho^-; w_i, w^-)$ *or if* $U^+ \equiv U^-$.
2. $\rho^- \geq \sigma(w_i, w^-)$: U^+ *is admissible if and only if* $\sigma(w_i, w^-) \leq \rho^+ \leq 1$.

If $U^- = (0, 0)$, *then the admissible state is* $U^+ \equiv U^-$.

In all cases the maximal possible flux associated with any admissible state U^+ is $d(\rho^-; w, w^-)$ with $w = w_i$.

Proof. For $U^- \neq (0,0)$ the 2-contact discontinuities are waves with speed $v^- > 0$. Hence, we only have to discuss 1-shock or 1-rarefaction waves. Following [2], a left state U^- can be connected to a right state U^+ by a 1-shock if and only if $\rho^+ > \rho^-$. The shock speed is then given by the slope of the chord U^-U^+. A left state U^- can be connected to a right state U^+ by a 1-rarefaction wave if and only if $\rho^+ < \rho^-$. Note that in the $(\rho, \rho v)$ plane the slope of the tangent to the curve $\{w_i(U) = c\}$ at a point U is the characteristic speed $\lambda_1(U)$.

By the discussion in the previous section there exists a state U^* with $\rho^* = r(\rho^-; w_i, w^-)$ and $v^* = u(\rho^-; w_i, w^-)$, such that $w_i(U^*) = w^-$. Furthermore, the chord U^-U^* has a zero slope. Hence, we have a 1-rarefaction wave for all states U^+ with $\sigma(w_i, w^-) \leq \rho^+ \leq \rho^-$, and a 1-shock for $\rho^+ \geq \rho^-$.

In both cases the associated flux is not greater than the demand $d(\rho^-; w, w^-)$.

Finally, if $\rho^+ > 0$, then $U^- = (0,0)$ can be connected to U^+ by a 2-contact discontinuity, which has either positive speed or zero speed; cf. Case 5 in [2]. Hence, only $U^+ \equiv (0,0)$ is admissible. □

Next, we consider the Riemann problem (1.8) for a given *outgoing* road $i \in \delta^+$, a function $w(U) := v + p_i(\rho)$, and a nonnegative constant c. Later on, c will of course depend on the initial states on the incoming roads (!); see sections 3 to 7. We look for "admissible" states U^-, i.e., all the states such that the waves of the solution have a positive speed or such that the solution is a constant. Again, we exclude the case of stationary waves. As in the previous case, U^- will be an intermediate state in the solution on the outgoing road i for the full Riemann problem at the junction. Now $U_i(x_0+, t) = U^-$ will hold.

PROPOSITION 2.2. *Consider a state $U^+ \neq (0,0)$ and the level curve of the first Riemann invariant $\{w(U) = c\}$ with an arbitrary nonnegative constant c.*

Let $U^\dagger = (\rho^\dagger, \rho^\dagger v^\dagger)$ be the point of intersection, if it exists, of the two Riemann invariants $\{v(U) = v^+\}$ and $\{w(U) = c\}$ with $\rho > 0$ and $v > 0$.

Then the "admissible" states U^- for the Riemann problem satisfying $w(U^-) = c$ and $\rho^- v^- \geq 0$ are given by two cases:

1. *$\rho^\dagger \leq \sigma(w, c)$: U^- is admissible if and only if $0 \leq \rho^- \leq \sigma(w, c)$.*
2. *$\rho^\dagger > \sigma(w, c)$: U^- is admissible if and only if $0 \leq \rho^- < r(\rho^\dagger; w, c)$ or if $U^- \equiv U^\dagger$.*

Note that the set of admissible states U^- depends on the existence of the point $U^\dagger$. Now assume that either $U^+ = (0,0)$ or there is no such point $U^\dagger$ with $\rho^\dagger, v^\dagger > 0$. Then we set $U^\dagger = (0,0)$ and as in Case 1, U^- is admissible if and only if $0 \leq \rho^- \leq \sigma(w, c)$.

In all cases the maximal possible flux associated with any "admissible" state U^- is $s(\rho^\dagger; w, c)$.

Proof. Due to the range of the eigenvalues we can connect a left state U^- to an intermediate state $U^\dagger$ by a 1-shock or a 1-rarefaction wave of positive speed. Then $U^\dagger$ can be connected to U^+ by a 2-contact discontinuity.

If $U^\dagger$ exists, it is well defined, since the curves $\{w(U) = c\}$ and $\{v(U) = v^+\}$ have a unique intersection point such that $\rho > 0, \rho v > 0$. If there is no point $U^\dagger$ with $\rho, v > 0$, then the curves have a unique intersection point at $(0,0)$.

Using the same kind of argument as in Proposition 2.1, we see that either the 1-shock or the 1-rarefaction waves connecting U^- and $U^\dagger$ have a positive speed or the solution is constant.

Next, if $\rho^+ = 0$, we set $U^\dagger = U^+$ and can connect to U^- by waves of the first family only; cf. Case 4 in [2]. □

Combining these two results, we obtain the following proposition.

PROPOSITION 2.3. *Consider an incoming (resp., outgoing) road i, an initial datum $U^- := U_{i,0}$ (resp., $U^+ := U_{i,0}$), and an arbitrary flux $q_0 \geq 0$. Let $w(U) := v + p_i(\rho)$ and $c := w(U_{i,0})$. Assume*

$$q_0 \leq d(\rho_{i,0}; w, c) \ (resp., \ q_0 \leq s(\rho_{i,0}; w, c)).$$

By Propositions 2.1 *and* 2.2*, there exists a unique state U^+ (resp., U^-), such that the corresponding Riemann problem* (1.8) *admits a solution such that $w(U^+) = c$ and $\rho^+ v^+ = q_0$ (resp., $w(U^+) = c$ and $\rho^+ v^+ = q_0$)* and *either all the waves have negative (resp., positive) speed, or the solution is a constant on the corresponding road.*

The reader is advised to pay attention to the notation. In the *full* solution to the Riemann problem at a junction we will have

$$\text{for } i \in \delta^- : \ U_i^+ = U_i(x_0-, t) \quad \text{and} \quad \text{for } i \in \delta^+ : \ U_i^- = U_i(x_0+, t). \tag{2.5}$$

Unfortunately, it seems hard to avoid this possibly misleading notation. Moreover, the state referred to here as U^+ will itself be an intermediate state called $U^\dagger$, defined as in Proposition 2.2.

To summarize, Proposition 2.3 describes the set of "admissible" states for the Riemann data on incoming and outgoing roads. We will refer to Proposition 2.3 regarding these states, which will be intermediate states in the solution of the full problem, satisfying (1.6a) *and* (1.6b). We now turn to the study of the first case.

3. One incoming and one outgoing road. The simplest possible network contains two roads connected by a junction, i.e., one road with two different road conditions.

PROPOSITION 3.1. *Consider two roads $i = 1, 2$ with $a_1 = -\infty$, $b_1 = a_2$, and $b_2 = \infty$ and initial data $U_{i,0} = (\rho_{i,0}, \rho_{i,0} v_{i,0})$, $i = 1, 2$ constant.*

Then there exists a unique solution $U_i(x, t)$ of the Riemann problem at the junction (1.8) *and* (1.9) *with the properties* (1) *and* (2)*. We refer to equation* (3.2) *and to the end of the proof for a description of the structure of this solution.*

(1) *$U_i(x, t)$ is a weak solution of the network problem* (1.5a)–(1.5b)*, where $p_i^\dagger \equiv p_i$, $i = 1, 2$, as given in* (1.1)*. Furthermore* (1.6a)–(1.6b) *are satisfied, and $\rho_i(x, t) v_i(x, t) \geq 0$, $i = 1, 2$.*

(2) *The flux $(\rho_1 v_1)(b_1^-, t)$ is maximal at the interface, subject to the above conditions.*

Proof. Let $U_1^- := U_{1,0}$, $U_2^+ := U_{2,0}$, and $w_i(U) = v + p_i(\rho)$ for $i = 1, 2$. As described in section 2 we construct the *demand* function for the incoming road

$$d(\rho) := d(\rho; w_1, w_1(U_1^-)),$$

and the *supply* function for the outgoing road

$$s(\rho) := s(\rho; w_2, w_1(U_1^-)). \tag{3.1}$$

Note that the supply function is an extension of the nonincreasing part of the curve $\{w_2(U) = w_1(U_1^-)\}$. The expression (3.1) of the supply function $s(\cdot)$ involves the *function* w_2 and the *value* $w_1(U_1^-)$, since the cars which are initially on road 1 and have moved onto road 2 have kept their Lagrangian "color" $w_1(U_1^-)$.

By Proposition 2.2 we obtain $U_2^\dagger$ either as the intersection of the curves $\{v_2(U) = v_2^+\}$ and $\{w_2(U) = w_1(U_1^-)\}$ or by $U_2^\dagger = (0,0)$. Then we solve the maximization problem

$$\begin{aligned} &\max q_1 \text{ subject to} \\ &0 \le q_1 \le d(\rho_1^-), \\ &0 \le q_1 \le s(\rho_2^\dagger). \end{aligned}$$

Denote by $\tilde{q}$ the point where the maximum is attained. Of course the above is equivalent to $\tilde{q} = \min\{d(\rho_1^-), s(\rho_2^\dagger)\}$, but we will need the general form later.

Now, as in Proposition 2.3 there exist U_1^+ and U_2^- such that $\rho_1^+ v_1^+ = \rho_2^- v_2^- = \tilde{q}$.

Knowing the states U_1^+ and U_2^-, we solve the two Riemann problems to obtain weak entropy solutions $U_1(x,t)$ and $U_2(x,t)$:

$$i = 1,2 : \partial_t \begin{bmatrix} \rho_i \\ v_i \end{bmatrix} + \partial_x \begin{bmatrix} \rho_i v_i \\ \rho_i v_i w_i \end{bmatrix} = 0, \tag{3.2a}$$

$$i = 1 : U_1(x,0) = \begin{bmatrix} U_1^- \equiv U_{1,0} & x < b_1 \\ U_1^+ & x \ge b_1 \end{bmatrix}, \tag{3.2b}$$

$$i = 2 : U_2(x,0) = \begin{bmatrix} U_2^- & x \le a_2 \\ U_2^+ \equiv U_{2,0} & x > a_2 \end{bmatrix}. \tag{3.2c}$$

Each solution consists of at most two waves: a 1-rarefaction or a 1-shock wave associated with the first eigenvalue, followed by a 2-contact discontinuity associated with the second eigenvalue.

The conditions (1.6a)–(1.6b) are satisfied since

$$\tilde{q} = \rho_1^+ v_1^+ = \rho_2^- v_2^-$$

and

$$w_1(U_1^+) = w_1(U_1^-) = w_2(U_2^-) = w_2(U_2^\dagger). \qquad \square$$

An example of a solution in the (x,t) plane is depicted in Figure 4.

4. One incoming and two outgoing roads. We now consider the case of one incoming and two outgoing roads. We cannot expect to obtain a unique solution without imposing additional assumptions on the distribution of the flux among the outgoing roads. One could impose an optimization criterion, such as maximizing the total flux at the interface [17, 5].

Here, we impose the proportions (α and $(1-\alpha)$) of cars which go from road 1 to roads 2 and 3. This condition was first introduced in [5] for the first-order LWR model and in [12] for the AR model. In the case of first-order models, the car distribution at junctions has also been studied in [23, 8] and many other works.

PROPOSITION 4.1. *Consider three roads $i = 1,2,3$ with $a_1 = -\infty$, $b_1 = a_2 = a_3$, and $b_2 = b_3 = \infty$ and constant initial data $U_{i,0} = (\rho_{i,0}, \rho_{i,0} v_{i,0})$, $i = 1,2,3$. Let $0 \le \alpha \le 1$ be given.*

Then there exists a unique solution $U_i(x,t)$, $i = 1,2,3$, of the Riemann problem at the junction (1.8) *and* (1.9) *with the following properties* (1) *and* (2). *A description of its structure can be found at the end of the proof.*

(1) *$U_i(x,t)$ is a weak solution of the network problem* (1.5a)–(1.5b), *wherein $p_i^\dagger \equiv p_i$ for all $i = 1,2,3$.*

Furthermore (1.6a)–(1.6b) *are satisfied, and $\rho_i(x,t) v_i(x,t) \ge 0$, $i = 1,2,3$.*

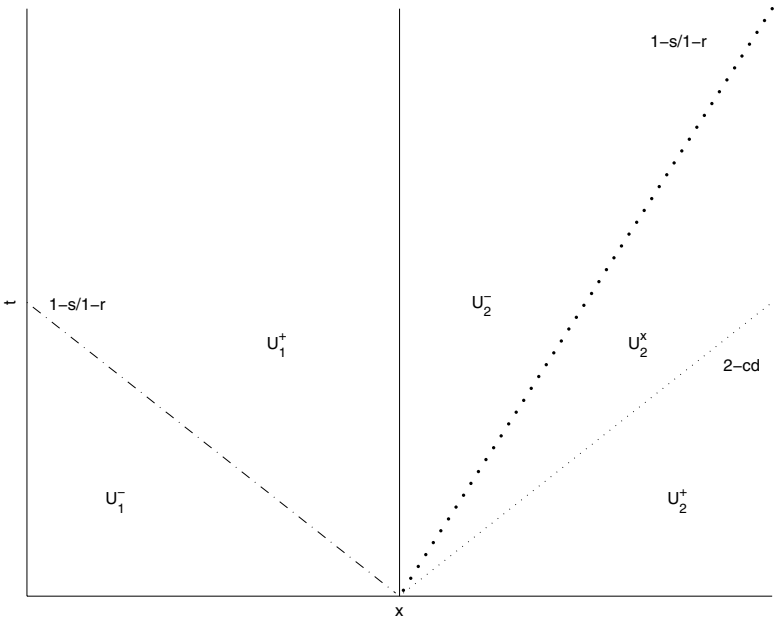


FIG. 4. *Possible solution to the Riemann problems of road* 1 (*left*) *and road* 2 (*right*). $1-s/1-r$ *stands for the* 1*-shock or* 1*-rarefaction wave connecting the left and right states. Similarly,* 2*-cd denotes the* 2*-contact discontinuity.*

(2) *For all* $t > 0$ *the flux is distributed in proportions* α *and* $1-\alpha$ *between roads* 2 *and* 3*:*

$$\alpha(\rho_1 v_1)(b_1^-, t) = (\rho_2 v_2)(a_2^+, t), \tag{4.1a}$$

$$(1-\alpha)(\rho_1 v_1)(b_1^-, t) = (\rho_3 v_3)(a_3^+, t). \tag{4.1b}$$

(3) *The flux* $(\rho_1 v_1)(b_1^-, t)$ *is maximal at the interface, subject to the above conditions.*

Proof. Let $U_1^- = U_{1,0}$, $U_i^+ = U_{i,0}$, $i = 2, 3$, and for $i = 1, 2, 3$ let $w_i(U) := v + p_i(\rho)$. As in section 2 we construct the demand function

$$d(\rho) := d(\rho; w_1, w_1(U_1^-))$$

and the two supply functions

$$s_2(\rho) := s(\rho; w_2, w_1(U_1^-)), \qquad s_3(\rho) := s(\rho; w_3, w_1(U_1^-)).$$

For $i = 2, 3$ we obtain the points $U_i^\dagger$ as the intersection of $\{v(U) = v_i^+\}$ with $\{w_i(U) = w_1(U_1^-)\}$ or as $U_i^\dagger = (0, 0)$; cf. Proposition 2.2. We solve the maximization problem

$$\max q_1 \text{ subject to} \tag{4.2a}$$

$$0 \le q_1 \le d(\rho_1^-), \tag{4.2b}$$

$$0 \le \alpha q_1 \le s_2(\rho_2^\dagger), \tag{4.2c}$$

$$0 \le (1-\alpha) q_1 \le s_3(\rho_3^\dagger). \tag{4.2d}$$

Denote by $\tilde{q}$ the point where the maximum is attained. Of course the above is equivalent to $\tilde{q} = \min\{d(\rho_1^-), s_2(\rho_2^\dagger)/\alpha, s_3(\rho_3^\dagger)/(1-\alpha)\}$. By Proposition 2.3 we conclude that

$$\exists U_1^+ \text{ such that } \rho_1^+ v_1^+ = \tilde{q},\ w_1(U_1^+) = w_1(U_1^-), \tag{4.3a}$$

$$\exists U_2^- \text{ such that } \rho_2^- v_2^- = \alpha\tilde{q},\ w_2(U_2^-) = w_1(U_1^-), \tag{4.3b}$$

$$\text{and } \exists U_3^- \text{such that } \rho_3^- v_3^- = (1-\alpha)\tilde{q},\ w_3(U_3^-) = w_1(U_1^-). \tag{4.3c}$$

Clearly, the conditions (1.6a)–(1.6b) and (4.1a)–(4.1b) are satisfied by (4.3). Again, each solution $U_i(x,t)$ consists of a juxtaposition of rarefaction or shock waves associated with the first eigenvalue and a contact discontinuity associated with the second eigenvalue of (1.1). The construction is similar to (3.2) in Proposition (3.1). In the limit cases $\alpha = 0$ or $\alpha = 1$ we are exactly in the setting of Proposition 3.1. ∎

Before studying the more surprising case of two incoming and one outgoing roads in section 6, we must recall a few basic facts on the Lagrangian version of the model and the corresponding homogenized system.

The reader is advised to take a look at the first part of section 5 and then move to section 6. The second part of section 5 deals with details on the homogenization and can be read after section 6.

5. The Lagrangian model and its homogenized version. The Lagrangian formulation is introduced in [1]. A formal derivation is given in [28] and a mathematical study in [13]. The homogenization of this system is studied in [3]. Proofs of statements below can be found in the above references.

Consider a single road with $p_i := p$. Then it turns out that the weak entropy solutions of

$$\rho_t + (\rho v)_x = 0, \qquad (\rho w)_t + (\rho v w)_x = 0, \qquad w = v + p(\rho)$$

correspond to the weak entropy solutions of the equivalent system in (mass) Lagrangian coordinates (X,t):

$$\tau_t - v_X = 0, \qquad w_t = 0, \qquad w = v + P(\tau), \tag{5.1}$$

with $\tau := 1/\rho, P(\tau) := p(\rho)$. Here, X is the Lagrangian (mass) coordinate, defined by $\partial_x X = \rho$ and $\partial_t X = -\rho\, v$. The existence of $X(\cdot,\cdot)$ follows from the mass conservation equation. For some unspecified t_0, $X(x,t_0) := \int_0^x \rho(y,t_0)\, dy$, where we implicitly define ρ as the dimensionless density, i.e., the *fraction of space* occupied by the cars; see [1]. Therefore X is the position of each car if all cars are parked "nose to tail."

As in [1] consider two different approximations of the system (5.1):

(i) the fully discrete solution of (5.1) constructed with the Godunov scheme, with space and time steps ΔX and Δt;

(ii) the semidiscrete approximation, namely the (infinite) system of ODEs

$$\partial_t \begin{bmatrix} \tau_j \\ w_j \end{bmatrix} - \begin{bmatrix} \frac{v_{j+1}-v_j}{\Delta X} \\ 0 \end{bmatrix} = 0,$$

where ΔX is the length of a car (fixed for simplicity). It is easy to see that this system can be rewritten in the form

$$\dot{x_j} = v_j,\ \dot{w_j} = 0 \text{ with } \tau_j = (x_{j+1} - x_j)/\Delta X,\ w_j = v_j + P(\tau_j). \tag{5.2}$$

In other words, the *semi*discretization of (5.1) is *exactly* the "follow-the-leader model" [15].

The *rigorous* results of convergence in [1, 13] are as follows:

(a) When ΔX and Δt tend to zero with a fixed ratio and satisfy the CFL stability condition, a subsequence of the fully discrete (Godunov) solution converges to a weak entropy solution of (5.1). This limit is viewed as a coarse graining limit, i.e., a "zooming" with the same ratio in X and t ("hyperbolic scaling").

(b) Next, when Δt tends to zero, with ΔX fixed, the Godunov solution converges to the unique solution of the microscopic follow-the-leader model (FLM) system.

(c) Finally, when ΔX tends to zero, this microscopic FLM solution converges to a weak entropy solution to (5.1).

These results were essentially based on uniform a priori BV-estimates (estimates on the total variation) for the Godunov solution. Indeed, this Lagrangian scheme preserves the total variation of the two Riemann invariants if the initial data are BV-functions.

The case of initial data with large oscillations in w, i.e., oscillations in the characteristics of car-driver pairs, is studied in [3]. Oscillations in w also generate oscillations in τ. Note that oscillations in v would be unrealistic (and dangerous!) and would be immediately cancelled by the genuinely nonlinear eigenvalue λ_1.

In the above mentioned (hyperbolic) "zooming," the oscillations in w are wilder and wilder as the zoom parameter goes to 0. Therefore, the corresponding sequence of functions converges only *weakly* to some limit. The above results can be extended, and uniqueness can be proved in this more general setting. The modification involves a homogenized relation between v, w, and τ, which uses the language of Young measures; see [27, 4, 11].

Let us briefly recall a few basic facts on Young measures, adapted to our context. The reader is advised to take a look at the practical example given in section 6.

We introduce a (Lagrangian) grid (X_j) and define $U_j = (\tau_j, w_j)$ and $U_{\Delta X}(X,t) := \sum_j U_j(t)\chi_j(X)$, where χ_j is the characteristic function on $I_j := (X_{j-1/2}, X_{j+1/2})$. For any $\Delta X > 0$, let $U_j^0 = (\tau_j^0, w_j^0)$ be uniformly bounded for all j, and let $U_{\Delta X}^0(X) := \sum_j U_j^0 \chi_j(X)$ be the corresponding sequence of piecewise constant initial data. Of course, this sequence is uniformly bounded in L^∞ when $\Delta X \to 0$.

Therefore [27, 4], there exist a subsequence, still denoted by $U_{\Delta X}^0(\cdot)$, and a family of probability measures $\nu_{X,t}$ in the (v,w) plane, depending on X, such that the weak-$*$ limit of *any* continuous function $F(v_{\Delta X}^0,\, w_{\Delta X}^0)$ is equal a.e. to

$$\langle \nu_{X,t},\, F(v,\, w)\rangle \;:=\; \int F(v,w)\, d\nu_{X,t}(v,\, w). \tag{5.3}$$

Since the sequence $(v_{\Delta X})$ does *not* oscillate, the same subsequence converges pointwise to some strong limit $v^*(X,t)$. Hence, (5.3) can be rewritten as

$$\langle \nu_{X,t},\, F(v,\, w)\rangle \;=\; \langle \mu_X,\, F(v^*(X,t),\, w)\rangle \;:=\; \int F(v^*(X,t),w)\, d\mu_X(w),$$

where the probability measures μ_X describe the weak limit of all functions in the single variable w. Therefore μ depends on X but not on t.

The main result in [3] can be stated as follows.

PROPOSITION 5.1. (i) *Under the above assumptions, the (sub)sequence of weak entropy solutions corresponding to the above (sub)sequence converges in L^∞ weak-$*$*

to the unique weak "entropy" solution $U^* = (\tau^*, w^*)$ *of the* homogenized *problem*

$$\partial_t \tau^* - \partial_X v^* = 0,\ \partial_t w^* = 0. \tag{5.4}$$

(ii) *Furthermore,* $(v_{\Delta X})$ *converges almost everywhere, and the limit state can be characterized as*

$$\tau^*(X,t) = \int P^{-1}(w - v^*(X,t))d\mu_X(w), \tag{5.5a}$$

$$w^*(X,t) = w^*(X,0) = \int w d\mu_X(w), \tag{5.5b}$$

where μ_X *is the Young measure associated with the sequence* $(w_{\Delta X})$.

Moreover, there is a similar result of homogenization for a *multiclass* FLM, similar to (5.2), with *oscillating* data w_j. We again refer to [3] for more details. Proposition 6.1 in the next section deals with a practical example of the above result.

6. Two incoming and one outgoing road. As in section 4, we need an additional assumption to obtain a unique solution at the junction. We introduce a "mixture rule," which describes how cars of the incoming road mix when they enter the outgoing road. One of the most natural assumptions is an equal priority rule: *The cars of both incoming roads enter the outgoing road alternately.*

Note that other assumptions on the mixture of cars are also possible. The discussion below remains valid with obvious changes according to different mixture rules.

PROPOSITION 6.1. *Consider three roads* $i = 1,2,3$ *with* $a_1 = a_2 = -\infty$, $b_1 = b_2 = a_3$, *and* $b_3 = \infty$ *and constant initial data* $U_{i,0} = (\rho_{i,0}\rho_{i,0}v_{i,0})$, $i = 1,2,3$.

Then there exists a unique solution $U_i(x,t)$, $i = 1,2,3$, *of the Riemann problem at the junction* (1.8) *and* (1.9) *with the following properties.*

(1) $U_i(x,t)$ *is a weak solution of the network problem* (1.5a)–(1.5b), *where* $p_i^\dagger \equiv p_i$ *for the incoming roads* $i = 1,2$.
For the outgoing road $i = 3$, *we obtain two different expressions for* $p_3^\dagger$, *depending on the position* (x,t)*:*
   (a) *In the triangle* $\{(x,t) : a_3 \le x \le a_3 + v_{3,0}t\}$ *of the* $x-t$ *plane, we consider the* homogenized solution *described below. Therefore,* $p_i^\dagger(\cdot) := p_i^*(\cdot)$ *is given by* (6.3)–(6.6). *This solution depends on the applied mixture principle, the initial data on* $U_{1,0}, U_{2,0}$, *and the road conditions* p_3. *The triangle is bounded at any fixed time* $t > 0$ *by* $x = a_3$ *and* $x = a_3 + tv_{3,0}$.
   (b) *In the remaining part of the outgoing road we have* $p_3^\dagger \equiv p_3$.

(2) *The equations* (1.6a)–(1.6b) *are satisfied, with* $\rho_i(x,t)v_i(x,t) \ge 0$, $1 \le i \le 3$. *In particular* $U_3(a_3^+, t)$ *satisfies*

$$w_3^\dagger(U_3(a_3^+,t)) := w_3^*(U_3(a_3^+,t)) := v_3(a_3^+,t) + p_3^*(\rho_3(a_3^+,t)) = \bar{w},$$

where $\bar{w}$ *is the* homogenized value*:*

$$\bar{w} := \frac{1}{2}\left(w_1(U_{1,0}) + w_2(U_{2,0})\right). \tag{6.1}$$

(3) *The two incoming fluxes are equal (equal priority rule), and the total flux* $2(\rho_1 v_1)(b_1^-, t) = 2(\rho_2 v_2)(b_2^-, t) = (\rho_3 v_3)(a_3^+, t)$ *is maximal subject to the other conditions.*

Before giving the proof of this result, let us motivate the definition of (6.1) and the necessity of dealing with a function p_3^*. Consider the discrete FLM (5.2), with oscillating $w_j = v_j + P(\tau_j)$:

$$\partial_t \begin{bmatrix} \tau_j \\ w_j \end{bmatrix} - \begin{bmatrix} \frac{v_{j+1}-v_j}{\Delta X} \\ 0 \end{bmatrix} = 0.$$

More precisely, consider a microscopic situation on the outgoing road 3. As in the introduction of this section, assume that the cars coming from each incoming road pass the junction in an alternating way.

Although w was *constant* on each of the roads 1 and 2, the outgoing flow is obviously oscillating. In fact, in *Lagrangian* coordinates,

$$w_j^0 = \begin{bmatrix} w_1 & j \text{ even} \\ w_2 & j \text{ odd} \end{bmatrix},$$

where the constants w_1 and w_2 are given by the two *incoming* flows. The corresponding function P on the outgoing road is the function $P_3(\tau) := p_3(1/\tau)$. Then the piecewise constant approximation $w_{\Delta X}$ alternately takes the two values w_1 and w_2. Consequently, for any continuous function F,

$$F(w_{\Delta X}) \rightharpoonup^* (F(w))^* := \frac{1}{2}(F(w_1) + F(w_2)) = \int F(w) d\mu_X(w),$$

$$\text{where } \mu_X := \frac{1}{2}\left(\delta_{w_1} + \delta_{w_2}\right). \tag{6.2}$$

The value of w has to be given by (6.1), since one car out of two comes from each road 1 or 2 (think of black and white cars producing a grey homogenized flow) *and* since any Lagrangian interval of length ΔX contains one car. Recall that we assumed that all cars have the same length. This assumption could be relaxed, and the formulas would be modified in an obvious way.

Therefore, in the limit $\Delta X \to 0$, the cars passing through the junction have the average property associated with the Young measure μ_X in (6.2). By section 5, the corresponding homogenized solution is the unique weak entropy solution of (5.4), where τ^* is given by (5.5a), i.e., here by

$$\tau^*(X,t) = \frac{1}{2}\left(P_3^{-1}(w_1 - v^*(X,t)) + P_3^{-1}(w_2 - v^*(X,t))\right), \tag{6.3}$$

which (by monotonicity of P_3) defines a one-to-one relation between $v := v^*(X,t)$ and $\tau := \tau^*(X,t)$.

We *choose* to rewrite (6.3) in the form

$$v = w - P_3^*(\tau),\ w := \bar{w}, \tag{6.4}$$

where $\bar{w}$ is given by (6.1). In other words, we *define* P_3^* so that, for each $\tau = \tau^*$, the value $v = v^*$ defined by (6.4) is the *unique* solution of (6.3) to the unknown v. This (convenient) notation could be misleading for an *arbitrary* value w. Indeed, the homogenized relation between v and τ depends on μ_X; see (5.5a). Therefore, it depends on the local proportions of cars coming from each incoming road. In other words, (6.4) would be *wrong* for any value $w \neq \bar{w}$. However, as we see below, on the relevant portion of road 3, the homogenized w *only* takes the value $\bar{w}$.

Now, three questions arise:

(i) How do we express this in Eulerian coordinates?

(ii) In Eulerian coordinates, what is the portion of road 3 concerned with this homogenized flow?

(iii) Does this solution respect the Rankine–Hugoniot relations (1.6a), (1.6b) at the interface $x = b_1 = b_2 = a_3$, and how is it connected with the downstream flow on road 3?

(i) First (see [1]), we can rewrite (5.4), (6.3) in Eulerian coordinates to get the equivalent system (even for weak entropy solutions):

$$\partial_t \rho + \partial_x(\rho v) = 0, \tag{6.5a}$$

$$\partial_t(\rho w) + \partial_x(\rho v w) = 0, \tag{6.5b}$$

$$w(U) \equiv w^*(U) = v + p_3^*(\rho), \tag{6.5c}$$

with $w(U) \equiv \bar{w}$ and

$$p_3^*(\rho) := P_3^*(\rho^{-1}) \tag{6.6}$$

defined by (6.4). Again, for *arbitrary* values of w, we would not recover the correct homogenized solution.

(ii) In the (x,t) plane, at time $t > 0$, the portion of road 3 concerned with this self-similar, homogenized flow is a triangle bounded by $x = b_1 = b_2 = a_3$ and by $x = a_3 + t\, v_{3,0}$. Here, $v_{3,0}$ is the initial datum on road 3.

(iii) On the above portion of road 3, our solution satisfies (6.5), (6.6) and the value of w is *a constant* and is equal to the corresponding average value given by (6.1).

The boundary data specified below preserve the conservation of mass at the intersection and satisfy the equal priority rule on the mixture of the cars:

$$\rho_3\, v_3 = \rho_1\, v_1 + \rho_2\, v_2 = 2\,\rho_1\, v_1.$$

Therefore, combining with (6.1), we see that $\rho_3\, v_3\, w_3 = \rho_1\, v_1\, w_1 + \rho_2\, v_2\, w_2$; i.e., we recover (1.6b): Our solution *also* satisfies the conservation of $y = \rho\, w$ at the junction. Roughly speaking, the total number of white cars is also preserved at the intersections!

Now we can give the proof of Proposition 6.1.

Proof. Let $U_i^- = U_{i,0}$ for $i = 1, 2$ and let $U_3^+ = U_{3,0}$. Denote by $w_i(U) = v + p_i(\rho)$. Let the demand functions d_1 and d_2 be defined by

$$d_1(\rho) := d(\rho; w_1, w_1(U_1^-)), \qquad d_2(\rho) := d(\rho; w_2, w_2(U_2^-)).$$

With all the previous remarks in mind and again with $w_3^\dagger(U) = v + p_3^\dagger(U)$ and $p_3^\dagger(\cdot) := p_3^*(\cdot)$, we consider the following supply function:

$$s_3(\rho) := s_3^\dagger(\rho) := s(\rho; w_3^\dagger, \bar{w}),\ \bar{w} = \frac{1}{2}\left(w_1(U_1^-) + w_2(U_2^-)\right).$$

As in Proposition 2.2 we obtain the intermediate state $U_3^\dagger = (\rho_3^\dagger, \rho_3^\dagger v_3^\dagger)$ as the intersection of $\{v_3(U) = v_3^+\}$ and $\{w_3^\dagger(U) = \bar{w}\}$, or as $U_3^\dagger = (0,0)$. Then we solve for

q_1, q_2:

$$\begin{aligned} &\max q_1 + q_2 \text{ subject to} \\ 0 \le \quad & q_i \quad \le d_i(\rho_i^-), \qquad i = 1, 2, \\ 0 \le \quad & q_1 + q_2 \quad \le s_3(\rho_3^\dagger), \\ & q_1 = q_2. \end{aligned}$$

Clearly, $\tilde{q} = q_1 = q_2 = \min\{s_3(\rho_3^\dagger)/2, d_1(\rho_1^-), d_2(\rho_2^-)\}$ is the unique solution. As in Proposition (2.3) we conclude that

$$\begin{aligned} \exists U_i^+ \text{ such that } \rho_i^+ v_i^+ = \tilde{q}, \quad w_i(U_i^+) = w_i(U_i^-),\ i = 1, 2, \\ \exists U_3^- \text{ such that } \rho_3^- v_3^- = 2\tilde{q}, \quad w_3^\dagger(U_3^-) = \bar{w}. \end{aligned}$$

We recall that $U_i(b_i-, t) = U_i^+$ for $i = 1, 2$ and $U_3(a_3+, t) = U_3^-$.

Then the conditions (1.6a)–(1.6b) are satisfied. Using the considerations above, the function $p_3^\dagger$ is defined in the triangle $\{(x, t) : a_3 \le x \le a_3 + tv_{3,0}\}$ of the $x - t$ plane.

Each solution $U_i(x, t)$ is a juxtaposition of either a rarefaction or a shock wave and a contact discontinuity.

In particular, on the outgoing road $i = 3$, the states U_3^- and $U_3^\dagger$ are connected by a rarefaction or a shock wave associated with the first eigenvalue of system (1.1), with $p_i = p_3^\dagger = p_3^*$. Then $U_3^\dagger$ is connected to $U_3^+ = U_{3,0}$ by a contact discontinuity associated with the second eigenvalue $\lambda_2 = v_{3,0}$, which is *independent* of p_i. Hence, out of the above mentioned triangle, $U_3(x, t) \equiv U_{3,0}$. ∎

An example of a solution is depicted in Figures 5 and 6.

7. Arbitrary number of incoming and outgoing roads. We combine the results of sections 4 to 6 to treat the general case. We consider a fixed junction with m incoming roads $\delta^- = \{1, \dots, m\}$ and n outgoing roads $\delta^+ = \{m + 1, \dots, m + n\}$. We assume constant initial data $U_{i,0}$ for all i and we look for solutions to the Riemann problem (1.8) and (1.9).

In sections 4 to 6 we imposed additional conditions to obtain a unique solution. Here, as in section 6 we introduce a mixture principle for the outgoing traffic which is an extension of the equal priority rule; cf. assumption (H4) below. However, the stated results can be adapted to other mixture rules.

For a set of functions $U_i(x, t) = (\rho_i(x, t), \rho_i(x, t)v_i(x, t))\,,\ i \in \delta^- \cup \delta^+$, we introduce the following abbreviations:

$$\mathbf{q}_i := \rho_i(b_i^-, t)v_i(b_i^-, t)\ \forall i \in \delta^-, \tag{7.1a}$$

$$\mathbf{q}_j := \rho_j(a_j^+, t)v_j(a_j^+, t)\ \forall j \in \delta^+. \tag{7.1b}$$

Next, we introduce real numbers $q_{ji} \in \mathbb{R}$ for $j \in \delta^+$ and $i \in \delta^-$ corresponding to the (a priori unknown) *actual fluxes* of cars coming from road i and going to road j. Since the number of cars entering and leaving the junction is the same,

$$\mathbf{q}_i = \sum_{j \in \delta^+} q_{ji}, \qquad \mathbf{q}_j = \sum_{i \in \delta^-} q_{ji}. \tag{7.2}$$

We look for a solution $U_k(x, t)$ which satisfies the following assumptions and constraints.

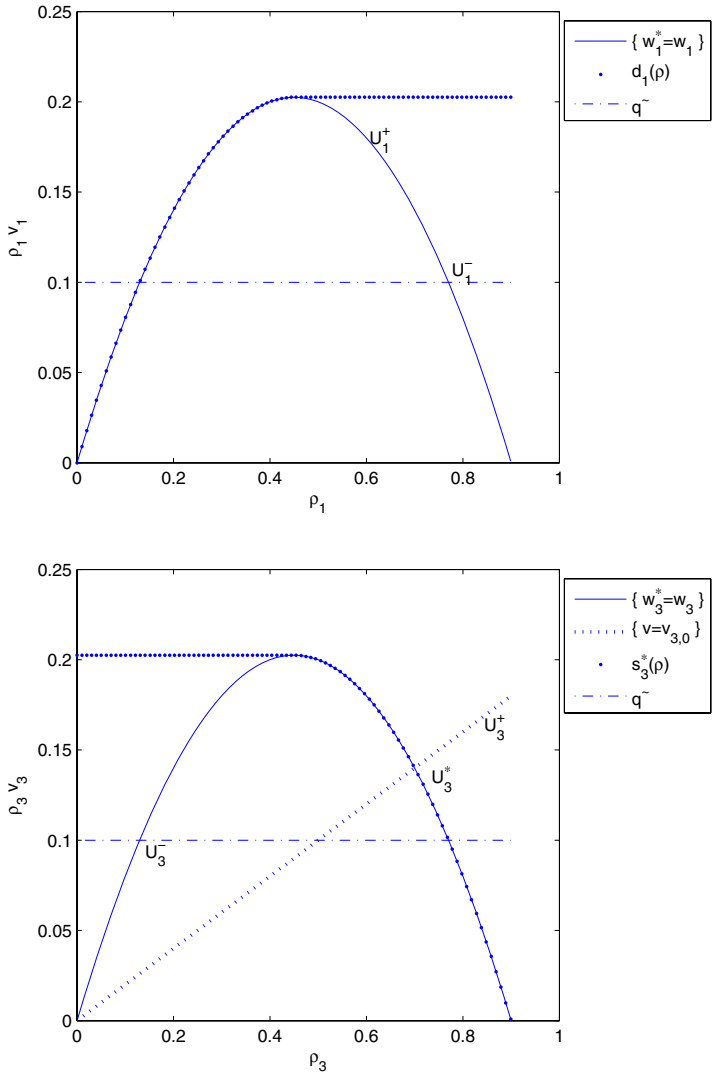


FIG. 5. *An example of intermediate states on road* 1 (*top*) *and road* 3 (*bottom*) *in the case where* $\tilde{q} = d_2(\rho_{2,0}; w_2, w_2(U_{2,0}))$, *i.e.,* $\tilde{q} < d_1(\cdot)$ *and* $\tilde{q} < s_3^\dagger(\cdot)$, *respectively. In this case the solution* U_2 *on road* 2 *is a constant* $U_2(x,t) = U_{2,0} \equiv U_2^-$ *and therefore is omitted from the plots. In the drawings* U_3^*, s_3^*, w_3^*, *and* w_1^* *stand for* $U_3^\dagger, s_3^\dagger, w_3^\dagger$, *and* $w_1^\dagger$, *respectively.*

(H1) Preferred choice of the drivers:
As in [12] we are given a matrix A,

$$A = (\alpha_{ji})_{j\in\delta^+, i\in\delta^-} \in \mathbb{R}^{n\times m}, \tag{7.3}$$

such that $0 \le \alpha_{ji} \le 1$ and $\sum_{j\in\delta^+} \alpha_{ji} = 1 \; \forall i \in \delta^-$.
We introduce $\mathbf{a}_j := \sum_{i\in\delta^-} \alpha_{ji}$ for notational convenience. We impose the constraint

$$q_{ji} = \alpha_{ji}\mathbf{q}_i \qquad \forall j \in \delta^+, \; i \in \delta^-. \tag{7.4}$$

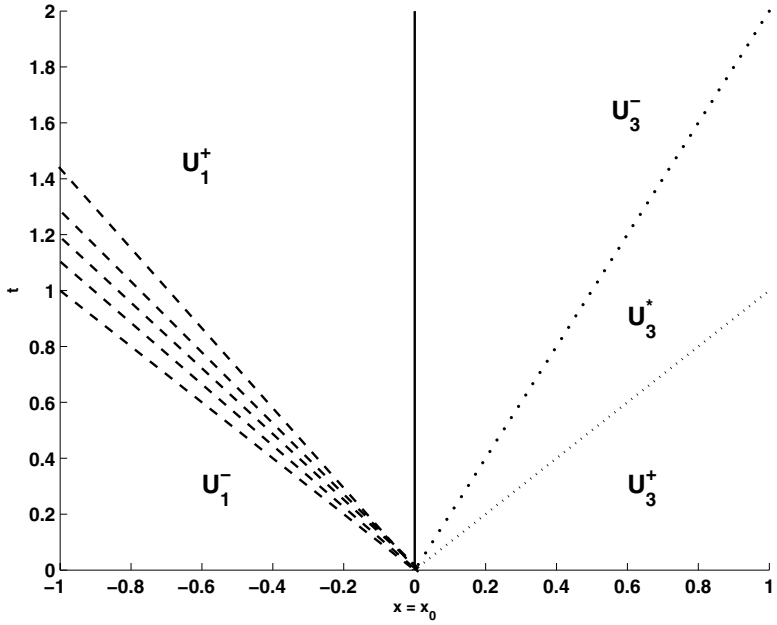


FIG. 6. *Plot of the solution U_1 and U_3 in the $x-t$ plane with data as in Figure 5. Left of the interface, $U_1^- \equiv U_{1,0}$ is connected by a 1-rarefaction to U_1^+. Right of the interface, U_3^- is connected by a 2–shock to $U_3^* \equiv U_3^\dagger$, and this state in turn is connected to $U_3^+ \equiv U_{3,0}$ by a 2–contact discontinuity. We omit the solution U_2 since it is constant.*

(H2) Relation for $w_j^\dagger$ (on the outgoing roads):

$$\forall j \in \delta^+ : w_j^\dagger(U_j(a_j^+,t)) = \sum_{i\in\delta^-} \frac{q_{ji}}{\mathbf{q}_j} w_i(U_i(b_i^-,t)). \tag{7.5}$$

As in the previous sections $w_i(U_i(b_i-,t)) = w_i(U_{i,0})\ \forall i \in \delta^-$ and $\forall j \in \delta^+$:

$$w_j^\dagger(U_j(a_j^+,t)) := v_j(a_j^+,t) + p_j^\dagger(\rho_j(a_j^+,t)) = \bar{w}_j.$$

The functions $p_j^\dagger$ and the homogenized values $\bar{w}_j$ have to be specified later.

(H3) Bounds on the actual fluxes:

$$0 \le \mathbf{q}_i \le d_i(\rho_{i,0})\ \forall i \in \delta^-, \tag{7.6a}$$

$$0 \le \mathbf{q}_j \le s_j(\rho_j^\dagger)\ \forall j \in \delta^+. \tag{7.6b}$$

Here d_i denotes the demand function on road i, i.e., $d_i := d_i(\rho; w_i, w_i(U_{i,0}))$, where $w_i(U) = v + p_i(\rho)$, and $s_j := s_j(\rho; w_j^\dagger, \bar{w}_j)$ is the supply function on road j. The functions $w_j^\dagger$ and the homogenized values $\bar{w}_j$ are specified later and depend on the applied mixture rule. Finally, $(\rho_j^\dagger, \rho_j^\dagger v_j^\dagger)$ is the intermediate state on road j, i.e., the unique intersection of the curves $\{v_j(U) = v_{j,0}\}$ and $\{w_j^\dagger(U) = \bar{w}_j\}$.

In order to define a unique solution, we have to impose a further constraint, i.e., a maximization criterion as in [17, 12]. Here, as in section 6, we choose to impose the following rule.

(H4) The mixture rule:
The *actual* incoming fluxes $(\mathbf{q}_i)_{i\in\delta^-}$ are proportional to a *given* nonnegative vector $(\tilde{q}_i)_{i\in\delta^-}$. The equal priority rule introduced in section 6 is a particular subcase, with $(\tilde{q}_i)_{i\in\delta^-} = (1,\dots,1)$. So we impose in general

$$\mathbf{q}_i = \tilde{q}\,\tilde{q}_i \geq 0, \tag{7.7}$$

where $\tilde{q} > 0$ is *a priori unknown*, but $(\tilde{q}_i)_{i\in\delta^-}$ is given.

THEOREM 7.1. *Consider a junction with m incoming and n outgoing roads, with constant initial data $U_{i,0} = (\rho_{i,0}, \rho_{i,0}v_{i,0})$ $\forall i \in \delta^- \cup \delta^+$ under the assumptions* (H1)–(H4).

Then there exists a unique solution $\{U_i(x,t)\}_{i\in\delta^-\cup\delta^+}$ to the Riemann problems (1.8)–(1.9) *which is described below, and which satisfies the following properties.*

(1) *$\{U_i(x,t)\}_{i\in\delta^-\cup\delta^+}$ is a weak entropy solution of the network problem* (1.5a)–(1.5b) *and for $i \in \delta^- : p_i^\dagger \equiv p_i$.*

For the outgoing roads $j \in \delta^+$ we obtain two different expressions for $p_j^\dagger$, depending on the region. In the $x-t$ plane in a triangle near the junction, we consider the homogenized solution and hence $p_j^\dagger(\cdot) = p_j^*(\cdot)$ defined below in (7.14). *This triangle is defined by $\{(x,t) : a_j \leq x \leq tv_{j,0}\}$ for any fixed time $t > 0$. Beyond this triangle we have $p_j^\dagger(\cdot) \equiv p_j(\cdot)$.*

(2) *The constraints* (7.4)–(7.6) *are satisfied, and the homogenized values $\bar{w}_j$ are given by*

$$\bar{w}_j := \sum_{i\in\delta^-} \frac{q_{ji}}{\mathbf{q}_j} w_i(U_{i,0})\ \forall j \in \delta^+. \tag{7.8}$$

The ratios $q_{ji}/\mathbf{q}_j$ are defined below in (7.15).

(3) *Moreover, the incoming fluxes satisfy* (7.7) *and are maximal subject to the other conditions.*

For simplicity we restrict ourselves the case of the equal priority rule. Obviously the proof can be extended to the general case (7.7). Note that the matrix A plays the same role as in [12], but we do *not* need the same restrictions on A.

Proof. With the discussion in section 6 in mind we consider the following supply functions for $j \in \delta^+$:

$$s_j(\rho) := s(\rho; w_j^\dagger; \bar{w}_j), \tag{7.9}$$

$$\bar{w}_j := \sum_{i\in\delta^-} \frac{q_{ji}}{\mathbf{q}_j} w_i(U_{i,0}), \tag{7.10}$$

where $w_i(U) = v + p_i(U)$ $\forall i \in \delta^-$ and where $w_j^\dagger(U) = v + p_j^\dagger(U)$ and $p_j^\dagger(\cdot) := p_j^*(\cdot)$ $\forall j \in \delta^+$. For each $j \in \delta^+$, $p_j^*(\cdot)$ is defined as in section 6. Namely, we first define the function

$$P_j(\tau) := p_j(1/\tau). \tag{7.11}$$

Then we set

$$v \to \tau := \sum_{i\in\delta^-} \frac{q_{ji}}{\mathbf{q}_j} P_j^{-1}(w_i(U_{i,0}) - v). \tag{7.12}$$

Next, we *choose* to define a new invertible function P_j^* by rewriting the relation (7.12) under the form

$$\tau := (P_j^*)^{-1}(\bar{w}_j - v), \tag{7.13}$$

which we use *only* with the particular value $\bar{w}_j$ defined in (7.8). Finally, we set

$$p_j^\dagger(\rho) := p_j^*(\rho) := P_j^*(1/\rho). \tag{7.14}$$

Of course this construction assumes that the proportions $q_{ji}/\mathbf{q}_j$ are known. Here, thanks to the crucial assumption (H4), we can determine them:

$$\frac{q_{ji}}{\mathbf{q}_j} = \frac{\alpha_{ji}\mathbf{q}_i}{\sum_{i\in\delta^-} q_{ji}} = \frac{\alpha_{ji}\tilde{q}_i}{\sum_{i\in\delta^-}\alpha_{ji}\tilde{q}_i}\ \forall i\in\delta^-,\ \forall j\in\delta^+. \tag{7.15}$$

In particular in the case of the equal priority rule, $\tilde{q}_i = 1\ \forall i\in\delta^-$ holds true. Therefore, $\mathbf{q}_i = \tilde{q}$, $\mathbf{q}_j = \mathbf{a}_j\tilde{q}$, and $q_{ji}/\mathbf{q}_j = \alpha_{ji}/\mathbf{a}_j$ for $i\in\delta^-$, $j\in\delta^+$ and for some unknown $\tilde{q}\in\mathbb{R}$.

Before we turn to the determination of $\tilde{q}$ we define $U_j^\dagger$. As in Proposition 2.2 we obtain for each j the intermediate state $U_j^\dagger$ as the intersection of $\{v_j(U) = v_{j,0}\}$ and $\{w_j^\dagger(U) = \bar{w}_j\}$.

Now, we obtain $\tilde{q}$ as the unique solution to the following maximization problem:

$$\max_{q\in\mathbb{R}} q \ \text{ subject to} \tag{7.16a}$$

$$0 \le \mathbf{q}_i = q \le d_i(\rho_{i,0}; w_i; w_i(U_{i,0}))\ \forall i\in\delta^-, \tag{7.16b}$$

$$0 \le \mathbf{q}_j = \mathbf{a}_j q \le s_j(\rho_j^\dagger; w_j^\dagger; \bar{w}_j)\ \forall j\in\delta^+, \tag{7.16c}$$

where the functions $s_j(\cdot), w_j^\dagger(\cdot)$ and the values $\bar{w}_j$ are well defined since the proportions $q_{ji}/\mathbf{q}_j$ are known.

We conclude as before that

$$\begin{aligned}
&\exists U_i^+ \text{ such that } \rho_i^+ v_i^+ = \tilde{q},\quad w_i(U_i^+) = w_i(U_{i,0})\ \forall i\in\delta^+,\\
&\exists U_j^- \text{ such that } \rho_j^- v_j^- = \mathbf{a}_j\tilde{q},\quad w_j^\dagger(U_j^-) = \bar{w}_j\ \forall j\in\delta^+.
\end{aligned}$$

The conditions (1.6a)–(1.6b) are satisfied. Also (7.5) and (7.6) are fulfilled.

Again, each $U_i(x,t)$ consists of a juxtaposition of rarefaction or shock waves associated with the first eigenvalue and, for $i\in\delta^+$, an additional contact discontinuity associated with the second eigenvalue. Furthermore, the solution satisfies on the incoming roads $i\in\delta^- : U_i^+ = U_i(b_i-,t)$ and on the outgoing roads $j\in\delta^+ : U_j^- = U_i(a_j+,t)$.

For general $\tilde{q}_i$, (7.16b) and (7.16c), respectively, become

$$\begin{aligned}
&0 \le \mathbf{q}_i = \tilde{q}_i\, q \le d_i(\rho_{i,0}; w_i, w_i(U_{i,0}))\ \ \forall i\in\delta^-,\\
&\text{and } 0 \le \mathbf{q}_j = \left(\sum_{i\in\delta^-}\alpha_{ji}\tilde{q}_i\right) q \le s_j^\dagger(\rho_{j,0}; w_j^\dagger, \bar{w}_j)\ \ \forall j\in\delta^+. \qquad \square
\end{aligned}$$

8. Conclusion. In this paper, we have introduced coupling conditions for the AR traffic flow model. Contrary to [12] the total "momentum" (e.g., the total number of white cars) is conserved at each junction. We have presented the full solution to Riemann problems for different cases and have given a microscopic motivation and validation of the approach. Last, we have discussed the general case of arbitrary numbers of incoming and outgoing roads. The most striking fact is the role of the homogenized flow on some part of the outgoing roads. It is worth noting that, even with Riemann data and with the *same* function $p_j \equiv p$ on *all* the roads, after some time, due to the mixture of cars at each junction, the flow is associated with a *new homogenized* pseudopressure $p_j^\dagger$, which depends on the proportions of the mixture.

As we already said in section 2, the model presented is too sophisticated for real life applications. But it contains as a particular case the classical first-order models.

Acknowledgments. M. Herty would like to thank the Laboratoire J. A. Dieudonné at Université Nice Sophia-Antipolis for its hospitality. The authors thank M. Garavello and B. Piccoli for their preprint [12].

REFERENCES

[1] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models*, SIAM J. Appl. Math., 63 (2002), pp. 259–278.
[2] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.
[3] P. Bagnerini and M. Rascle, *A multiclass homogenized hyperbolic model of traffic flow*, SIAM J. Math. Anal., 35 (2003), pp. 949–973.
[4] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDEs and Continuum Models of Phase Transitions, Lecture Notes in Phys. 344, M. Rascle, D. Serre, and M. Slemrod, eds., Springer-Verlag, Berlin, 1989, pp. 207–215.
[5] G. M. Coclite, M. Garavello, and B. Piccoli, *Traffic flow on a road network*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.
[6] R. M. Colombo, *Hyperbolic phase transitions in traffic flow*, SIAM J. Appl. Math., 63 (2002), pp. 708–721.
[7] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, Heidelberg, New York, 2000.
[8] C. F. Daganzo, *A behavioral theory of multi-lane traffic flow part* I: *Long homogeneous freeway sections*, Trans. Res. B, 36 (2002), pp. 131–158.
[9] C. F. Daganzo, *A behavioral theory of multi-lane traffic flow part* II: *Merges and the onset of congestion*, Trans. Res. B, 36 (2002), pp. 159–169.
[10] C. F. Daganzo, *Requiem for second order fluid approximations of traffic flow*, Trans. Res. B, 29 (1995), pp. 277–286.
[11] R. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27–70.
[12] M. Garavello and B. Piccoli, *Traffic Flow on a Road Network Using the Aw-Rascle Model*, preprint, 2004.
[13] J. M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.
[14] J. M. Greenberg, A. Klar, and M. Rascle, *Congestion on multilane highways*, SIAM J. Appl. Math., 63 (2003), pp. 818–833.
[15] D. Helbing, *Verkehrsdynamik*, Springer-Verlag, Berlin, Heidelberg, New York, 1997.
[16] M. Herty and A. Klar, *Modeling, simulation, and optimization of traffic flow networks*, SIAM J. Sci. Comput., 25 (2003), pp. 1066–1087.
[17] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.
[18] R. Illner, A. Klar, and M. Materne, *Vlasov-Fokker-Planck models for multilane traffic flow*, Commun. Math. Sci., 1 (2003), pp. 1–12.
[19] A. Klar, R. Kühne, and R. Wegener, *Mathematical models for vehicular traffic*, Surveys Math. Indust., 6 (1996), pp. 215–239.
[20] A. Klar and R. Wegener, *Kinetic derivation of macroscopic anticipation models for vehicular traffic*, SIAM J. Appl. Math., 60 (2000), pp. 1749–1766.

[21] A. Klar and R. Wegener, *A hierarchy of models for multilane vehicular traffic* I: *Modeling*, SIAM J. Appl. Math., 59 (1999), pp. 983–1001.
[22] A. Klar and R. Wegener, *A hierarchy of models for multilane vehicular traffic* II: *Numerical investigations*, SIAM J. Appl. Math., 59 (1999), pp. 1002–1011.
[23] J. P. Lebacque, *Les modèles macroscopiques du trafic*, Annales des Ponts, 67 (1993), pp. 28–45.
[24] T. Li, *Well-posedness theory of an inhomogeneous traffic flow model*, Discrete Contin. Dyn. Syst. Ser. B, 2 (2002), pp. 401–414.
[25] M. J. Lighthill and J. B. Whitham, *On kinematic waves*, Proc. Roy. Soc. Edinburgh Sect. A, 229 (1955), pp. 281–316.
[26] H. J. Payne, *FREFLO: A macroscopic simulation model for freeway traffic*, Trans. Res. Rec., 722 (1979), pp. 68–77.
[27] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, Res. Notes in Math. 39, Pitman, Boston, London, 1979, pp. 136–212.
[28] H. M. Zhang, *A non-equilibrium traffic model devoid of gas-like behavior*, Trans. Res. B, 36 (2002), pp. 275–290.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 617–636

HIGH FREQUENCY ANALYSIS OF HELMHOLTZ EQUATIONS: CASE OF TWO POINT SOURCES*

ELISE FOUASSIER†

Abstract. We derive the high frequency limit of the Helmholtz equation with source term when the source is the sum of two point sources. We study it in terms of Wigner measures (quadratic observables). We prove that the Wigner measure associated with the solution satisfies a Liouville equation with, as source term, the sum of the source terms that would be created by each of the two point sources taken separately. The first step, and main difficulty, in our study is to obtain uniform estimates on the solution. Then, from these bounds, we derive the source term in the Liouville equation together with the radiation condition at infinity satisfied by the Wigner measure.

Key words. Wigner measures, high frequencies, energy propagation

AMS subject classifications. 35Q60, 35J05, 81S30

DOI. 10.1137/050629999

1. Introduction.

In this article, we are interested in the analysis of the high frequency limit of the Helmholtz equation

$$-i\frac{\alpha_\varepsilon}{\varepsilon}u^\varepsilon + \Delta u^\varepsilon + \frac{n(x)^2}{\varepsilon^2}u^\varepsilon = S^\varepsilon(x), \qquad x \in \mathbb{R}^3, \tag{1.1}$$

with

$$S^\varepsilon(x) = S_0^\varepsilon(x) + S_1^\varepsilon(x) = \frac{1}{\varepsilon^3}S_0\Big(\frac{x}{\varepsilon}\Big) + \frac{1}{\varepsilon^3}S_1\Big(\frac{x-q_1}{\varepsilon}\Big),$$

where q_1 is a point in $\mathbb{R}^3$ different from the origin.

In what follows, we assume that the refraction index n is constant, $n(x) \equiv 1$.

Equation (1.1) models the propagation of a source wave in a medium with refraction index $n(x)$. There, the small positive parameter ε is related to the frequency $\omega = \frac{1}{2\pi\varepsilon}$ of u^ε. In this paper, we study the high frequency limit, i.e., the asymptotics $\varepsilon \to 0$. We assume that the regularizing parameter α_ε is positive, with $\alpha_\varepsilon \to 0$ as $\varepsilon \to 0$. The positivity of α_ε ensures the existence and uniqueness of a solution u^ε to the Helmholtz equation (1.1) in $L^2(\mathbb{R}^3)$ for any $\varepsilon > 0$.

The source term S^ε models a source signal that is the sum of two source signals concentrating, respectively, close to the origin and close to the point q_1 at the scale ε. The concentration profiles S_0 and S_1 are given functions. Since ε is also the scale of the oscillations dictated by the Helmholtz operator $\Delta + \frac{1}{\varepsilon^2}$, resonant interactions can occur between these oscillations and the oscillations due to the sources S_0^ε and S_1^ε. On the other hand, since the two sources are concentrating close to two different points in $\mathbb{R}^3$, one can guess that they do not interact when $\varepsilon \to 0$. These are the phenomena that the present paper aims at studying quantitatively. We refer to section 3 for the precise assumptions we need on the sources.

*Received by the editors April 26, 2005; accepted for publication (in revised form) January 31, 2006; published electronically June 23, 2006.
http://www.siam.org/journals/sima/38-2/62999.html
†Université de Rennes 1, IRMAR, Campus de Beaulieu, 35042 Rennes cedex, France. Current address: Ecole Normale Superieure de Lyon, UMPA, 69364 Lyon, France (elise.fouassier@umpa.ens-lyon.fr).

In some sense, the sign of the term $-i\alpha_\varepsilon u^\varepsilon/\varepsilon$ prescribes a radiation condition at infinity for u^ε, i.e., the oscillations of the solution at infinity (as $e^{-in|x|}$). One of the key difficulties in our problem is following this condition in the limiting process $\varepsilon \to 0$. We refer to section 3 for more details.

We study the high frequency limit in terms of Wigner measures (or semiclassical measures). This is a means of describing the propagation of quadratic quantities, like the local energy density $|u^\varepsilon(x)|^2$, as $\varepsilon \to 0$. The Wigner measure $\mu(x,\xi)$ is the energy carried by rays at the point x with frequency ξ. These measures were introduced by Wigner [14] and mathematically developed by Gérard [6] and Lions and Paul [9] (see also the surveys [3] and [7]). They are relevant when a typical length ε is prescribed. They have already proven to be an efficient tool in the study of high frequencies; see, for instance, [2], [4] for Helmholtz equations, Gérard et al. [7] for periodic media, Papanicolaou and Ryzhik [11] for a formal analysis of general wave equations, Erdös and Yau [5] for an approach linked to statistical physics, and Miller [10] for a study in the case with sharp interface.

The high frequency limit of Helmholtz equations has been studied in Benamou et al. [2] and Castella, Perthame, and Runborg [4]. In [2], the authors considered the case of one point source and a general index of refraction, whereas in [4] they treated the case of a source concentrating close to a general manifold with a constant refraction index. In the present paper, we borrow the methods used in both articles.

In the case of one point source, for instance, S_0^ε only, with a constant index of refraction, it is proved in [2] that the corresponding Wigner measure μ_0 is the solution to the Liouville equation

$$0^+\mu_0(x,\xi) + \xi\cdot\nabla_x\mu_0(x,\xi) = Q_0(x,\xi) = \frac{1}{(4\pi)^2}\delta(x)\delta(|\xi|^2-1)|\widehat{S_0}(\xi)|^2,$$

the term 0^+ meaning that μ is the outgoing solution given by

$$\mu_0(x,\xi) = \int_{-\infty}^0 Q_0(x+t\xi,\xi)dt.$$

In particular, the energy source created by S_0^ε is supported at $x=0$. Similarly, the energy source created by the source S_1^ε is supported at $x=q_1$. Thinking of the orthogonality property on Wigner measures, one can guess that the energy source generated by the sum $S_0^\varepsilon + S_1^\varepsilon$ is the sum of the two energy sources created asymptotically by S_0^ε and S_1^ε.

Indeed, we prove in this paper that the Wigner measure μ associated with the sequence (u^ε) satisfies

$$0^+\mu(x,\xi) + \xi\cdot\nabla_x\mu(x,\xi) = Q_0 + Q_1, \tag{1.2}$$

where Q_0 and Q_1 are the source terms obtained in [2] in the case of one point source. However, our proof does not rest on the mere orthogonality property.

Let us now give some details about our proof. Our strategy is borrowed from [2]. First, we prove uniform estimates on the sequence of solutions (u^ε). It turns out that we also need to study the limiting behavior of (and to estimate) the rescaled solutions $\varepsilon u^\varepsilon(\varepsilon x)$ and $\varepsilon u^\varepsilon(q_1+\varepsilon x)$. The latter point is the key difficulty in our paper. It relies on the study of the sequence (a^ε) such that

$$-i\alpha_\varepsilon\varepsilon a^\varepsilon + \Delta a^\varepsilon + a^\varepsilon = S_1\Big(x - \frac{q_1}{\varepsilon}\Big).$$

Using the explicit formula for the Fourier transform of a^ε, we prove that a^ε is uniformly bounded in a suitable space and that $a^\varepsilon \to 0$ as $\varepsilon \to 0$ weakly. We would like to point out that our analysis, based on a study in Fourier space, strongly rests on the assumption of a constant index of refraction.

Second, our results on the Wigner measure then follow from the properties proved in [2]. They are essentially consequences of the uniform bounds on (u^ε): we write the equation satisfied by the Wigner transform associated with (u^ε) and pass to the limit $\varepsilon \to 0$ in the various terms that appear in this equation. The only difficult (and new) term to handle is the source term.

Third, we prove an improved version of the radiation condition of [2]. Our argument relies on the observation that μ is localized on the energy set $\{|\xi|^2 = 1\}$, a property that was not exploited in [2].

The paper is organized as follows. In section 2, we recall some definitions and state our assumptions. Section 3 is devoted to the proof of uniform bounds on the sequence of solutions (u^ε) and of the convergence of the rescaled solutions. Then, in section 4, we establish the transport equation satisfied by the Wigner measure μ together with the radiation condition at infinity.

2. Notation and assumptions. In this section, we recall the definitions of Wigner transforms and of the B, B^* norms introduced by Agmon and Hörmander [1] for the study of Helmholtz equations. Then, we give our assumptions.

2.1. Wigner transform and Wigner measures. We use the following definition for the Fourier transform:

$$\hat{u}(\xi) = (\mathcal{F}_{x\to\xi}u)(\xi) = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^d} e^{-ix\cdot\xi}u(x)dx.$$

For $u, v \in \mathcal{S}(\mathbb{R}^3)$ and $\varepsilon > 0$, we define the Wigner transform

$$W^\varepsilon(u,v)(x,\xi) = (\mathcal{F}_{y\to\xi})\left(u\left(x+\frac{\varepsilon}{2}y\right)\bar{v}\left(x-\frac{\varepsilon}{2}y\right)\right),$$
$$W^\varepsilon(u) = W^\varepsilon(u,u).$$

In what follows, we denote $W^\varepsilon = W^\varepsilon(u^\varepsilon)$.

If (u^ε) is a bounded sequence in $L^2(\mathbb{R}^d)$ (or in some weighted-L^2 space, as we will see later on), it turns out that (see [6], [9]), up to extracting a subsequence, the sequence $(W^\varepsilon(u^\varepsilon))$ converges weakly to a positive Radon measure μ on the phase space $T^*\mathbb{R}^3 = \mathbb{R}^3_x \times \mathbb{R}^3_\xi$ called the Wigner measure (or semiclassical measure) associated with (u^ε):

$$\forall \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^6),\ \lim_{\varepsilon\to 0}\langle W^\varepsilon(u^\varepsilon),\varphi\rangle = \int \varphi(x,\xi)d\mu. \tag{2.1}$$

We recall that these measures can be obtained using pseudodifferential operators. The Weyl semiclassical operator $a^W(x,\varepsilon D_x)$ (or $Op_\varepsilon^W(a)$) is the continuous operator from $\mathcal{S}(\mathbb{R}^d)$ to $\mathcal{S}'(\mathbb{R}^d)$ associated with the symbol $a \in \mathcal{S}'(T^*\mathbb{R}^d)$ by Weyl quantization rule

$$(a^W(x,\varepsilon D_x)u)(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d_\xi}\int_{\mathbb{R}^d_y} a\left(\frac{x+y}{2},\varepsilon\xi\right)f(y)e^{i(x-y)\cdot\xi}d\xi dy. \tag{2.2}$$

We have the following formula: for $u, v \in \mathcal{S}'(\mathbb{R}^d)$ and $a \in \mathcal{S}(\mathbb{R}^d \times \mathbb{R}^d)$,

$$\langle W^\varepsilon(u,v), a\rangle_{\mathcal{S}',\mathcal{S}} = \langle \bar{v}, a^W(x,\varepsilon D_x)\bar{u}\rangle_{\mathcal{S}',\mathcal{S}}, \tag{2.3}$$

where the duality brackets $\langle .,. \rangle$ are semilinear with respect to the second argument. This formula is also valid for u, v lying in other spaces, as we will see in section 3.

2.2. Besov-like norms. In order to get uniform (in ε) bounds on the sequence (u^ε), we shall use the following Besov-like norms, introduced by Agmon and Hörmander [1]: for $u, f \in L^2_{loc}(\mathbb{R}^3)$, we denote

$$\|u\|_{B^*} = \sup_{j \geq -1} \left(2^{-j} \int_{C(j)} |u|^2 dx \right)^{1/2},$$
$$\|f\|_B = \sum_{j \geq -1} \left(2^{j+1} \int_{C(j)} |f|^2 dx \right)^{1/2},$$

where $C(j)$ denotes the ring $\{x \in \mathbb{R}^3 / 2^j \leq |x| < 2^{j+1}\}$ for $j \geq 0$ and $C(-1)$ is the unit ball.

These norms are adapted to the study of Helmholtz operators. Indeed, Agmon and Hörmander [1] proved that if v is the solution to

$$-i\alpha v + \Delta v + v = f,$$

where $\alpha > 0$, then there exists a constant C independent of α such that

$$\|v\|_{B^*} \leq C \|f\|_B.$$

Perthame and Vega [12] generalized this result to Helmholtz equations with general indices of refraction.

We denote, for $x \in \mathbb{R}^3$, $|x| = \sqrt{\sum_{j=1}^3 x_j^2}$ and $\langle x \rangle = (1 + |x|^2)^{1/2}$. For all $\delta > \frac{1}{2}$, we have

$$\|u\|_{L^2_{-\delta}} := \|\langle x \rangle^{-\delta} u\|_{L^2} \leq C(\delta) \|u\|_{B^*}, \tag{2.4}$$

and

$$\|f\|_B \leq C(\delta) \|f\|_{L^2_\delta}. \tag{2.5}$$

We end this section by stating two properties of these spaces that will be useful for our purpose (the reader can find the proofs in [1]). The first proposition states that, in some sense, we can define the trace of a function in B on a linear manifold of codimension 1.

PROPOSITION 2.1. *There exists a constant C such that for all $f \in B$, we have*

$$\int_{\mathbb{R}} \|f(x_1, .)\|_{L^2(\mathbb{R}^2)} dx_1 \leq C \|f\|_B.$$

The second property gives the stability of the space B by change of variables in Fourier space.

PROPOSITION 2.2. *Let Ω_1, Ω_2 be two open sets in $\mathbb{R}^3$, $\psi : \Omega_1 \to \Omega_2$ a $\mathcal{C}^2$ diffeomorphism, $\chi \in \mathcal{C}^1_c(\mathbb{R}^3)$. For all $u \in B$, we denote*

$$Tu = \mathcal{F}^{-1}\big(\chi(\hat{u} \circ \psi)\big).$$

Then

$$\|Tu\|_B \leq C \|\chi\|_{\mathcal{C}^1_b} \|\psi\|_{\mathcal{C}^2_b} \|u\|_B.$$

2.3. Assumptions. We are now ready to state our assumptions. Our first assumption, borrowed from [2], concerns the regularizing parameter $\alpha_\varepsilon > 0$.

(H1) $\alpha_\varepsilon \geq \varepsilon^\gamma$ for some $\gamma > 0$.

This assumption is technical and is used to get a radiation condition at infinity in the limit $\varepsilon \to 0$. Next, in order to compute the limit of the energy source, we shall need the following assumption.

(H2) $\langle x\rangle^N S_0 \in L^2(\mathbb{R}^3)$ and $\langle x\rangle^N S_1 \in L^2(\mathbb{R}^3)$ for some $N > \frac{1}{2} + \frac{3\gamma}{\gamma+1}$.

Note that (H2) implies that the source terms S_0 and S_1 belong to the natural Besov space that is needed to actually solve the Helmholtz equation (1.1):

$$\|S_0\|_B, \|S_1\|_B < \infty.$$

3. Bounds on solutions to Helmholtz equations. In this section, we first establish uniform bounds on the sequence (u^ε) that will imply estimates on the sequence of Wigner transforms (W^ε). It turns out that we shall also need to compute the limit of the rescaled solutions w_0^ε and w_1^ε defined below in order to obtain the energy source in the equation satisfied by the Wigner measure μ.

Before stating our two results, let us define these rescaled solutions. Following [2] and [4], we denote

$$\left\{ \begin{array}{l} w_0^\varepsilon(x) = \varepsilon u^\varepsilon(\varepsilon x), \\ w_1^\varepsilon(x) = \varepsilon u^\varepsilon(q_1 + \varepsilon x). \end{array} \right. \tag{3.1}$$

They, respectively, satisfy

$$\left\{ \begin{array}{l} -i\alpha_\varepsilon \varepsilon w_0^\varepsilon + \Delta w_0^\varepsilon + w_0^\varepsilon = S_0(x) + S_1\left(x - \dfrac{q_1}{\varepsilon}\right), \\ -i\alpha_\varepsilon \varepsilon w_1^\varepsilon + \Delta w_1^\varepsilon + w_1^\varepsilon = S_0\left(x + \dfrac{q_1}{\varepsilon}\right) + S_1(x). \end{array} \right.$$

We are ready to state our results on u^ε, w_0^ε, and w_1^ε.

PROPOSITION 3.1. *Assume S_0, $S_1 \in B$. Then*

(i) *the solution u^ε to the Helmholtz equation* (1.1) *satisfies the bound*

$$\|u^\varepsilon\|_{B^*} \leq C(\|S_0\|_B + \|S_1\|_B);$$

(ii) *the rescaled solutions (w_0^ε) and (w_1^ε) are uniformly bounded in B^*:*

$$\|w_0^\varepsilon\|_{B^*} + \|w_1^\varepsilon\|_{B^*} \leq C(\|S_0\|_B + \|S_1\|_B).$$

(In both points, C denotes a constant independent of ε.)

PROPOSITION 3.2. *The sequences of rescaled solutions (w_0^ε) and (w_1^ε) converge weakly-$*$ in B^* to the outgoing solutions w_0 and w_1 to the Helmholtz equations*

$$\left\{ \begin{array}{l} \Delta w_0 + w_0 = S_0, \\ \Delta w_1 + w_1 = S_1, \end{array} \right.$$

i.e., w_0 and w_1 are given in Fourier space by

$$\widehat{w_j}(\xi) = \frac{-\widehat{S_j}(\xi)}{|\xi|^2 - 1 + i0} = -\left(p.v.\left(\frac{1}{|\xi|^2 - 1}\right) + i\pi\delta(|\xi|^2 - 1)\right)\widehat{S_j}(\xi), \quad j = 0, 1.$$

Remark. The Helmholtz equation $\Delta w + w = S$ does not uniquely specify the solution w. An extra condition is necessary, for instance, the Sommerfeld radiation condition. When the refraction index is constant equal to 1, this condition reads

$$\lim_{r\to\infty} \frac{1}{r} \int_{S_r} \left| \frac{\partial w}{\partial r} + iw \right|^2 d\sigma = 0. \tag{3.2}$$

Such a solution is called an outgoing solution.

Alternatively, still assuming that the refraction index is constant, the outgoing solution to the Helmholtz equation may be defined as the weak limit, as $\delta \to 0$, w of the sequence (w^δ) such that

$$-i\delta w^\delta + \Delta w^\delta + w^\delta = S(x).$$

We point out that the two points of view are equivalent in the case of a constant index of refraction (which is not true for a general index of refraction).

We prove the two propositions in the following two sections. As we will see in the proofs, our main difficulties are linked to the rays that are emitted by the source at 0 towards the point q_1 (and conversely). Hopefully, the interaction between those rays is "destructive" and not constructive.

3.1. Proof of Proposition 3.1. In what follows, C will denote any constant independent of ε.

Proof of point (i). Since u^ε is a solution to the Helmholtz equation (1.1), and *since the refraction index is constant*, we may write $u^\varepsilon(x) = u_0^\varepsilon(x) + u_1^\varepsilon(x - q_1)$, where u_0^ε and u_1^ε satisfy

$$\begin{cases} -i\dfrac{\alpha_\varepsilon}{\varepsilon} u_0^\varepsilon + \Delta u_0^\varepsilon + \dfrac{1}{\varepsilon^2} u_0^\varepsilon = \dfrac{1}{\varepsilon^3} S_0\left(\dfrac{x}{\varepsilon}\right), \\ -i\dfrac{\alpha_\varepsilon}{\varepsilon} u_1^\varepsilon + \Delta u_1^\varepsilon + \dfrac{1}{\varepsilon^2} u_1^\varepsilon = \dfrac{1}{\varepsilon^3} S_1\left(\dfrac{x}{\varepsilon}\right). \end{cases}$$

If we denote, for $j = 0, 1$, $\widetilde{w_j^\varepsilon}(x) = \varepsilon u_j^\varepsilon(\varepsilon x)$, then

$$\begin{cases} -i\alpha_\varepsilon \varepsilon \widetilde{w_0^\varepsilon} + \Delta \widetilde{w_0^\varepsilon} + \widetilde{w_0^\varepsilon} = S_0(x), \\ -i\alpha_\varepsilon \varepsilon \widetilde{w_1^\varepsilon} + \Delta \widetilde{w_1^\varepsilon} + \widetilde{w_1^\varepsilon} = S_1(x). \end{cases}$$

Moreover, the bound $\|\widetilde{w_j^\varepsilon}\|_{B^*} \le C\|S_j\|_B$ is established in Agmon and Hörmander [1] (see also Perthame and Vega [12]). Hence, using the scaling invariance

$$\|u_j^\varepsilon\|_{B^*} \le \|\widetilde{w_j^\varepsilon}\|_{B^*},$$

we get

$$\|u_j^\varepsilon\|_{B^*} \le C\|S_j\|_B.$$

Finally, we obtain the uniform estimate on u^ε by noting that, since q_1 is a fixed point, we have

$$\|u_1^\varepsilon(. - q_1)\|_{B^*} \le C(q_1)\|u_1^\varepsilon\|_{B^*},$$

which implies

$$\|u^\varepsilon\|_{B^*} \le \|u_0^\varepsilon\|_{B^*} + C(q_1)\|u_1^\varepsilon\|_{B^*} \le C(q_1)\left(\|S_0\|_B + \|S_1\|_B\right).$$

Proof of point (ii). Unfortunately, to get the bound on the rescaled sequences w_0^ε and w_1^ε, we cannot use a method similar to the one used in the proof of point (i) (roughly, because of the translation q_1/ε and not q_1). We instead use the explicit formula for w_j^ε that is available. In what follows, we prove the result for w_0^ε; the estimate on w_1^ε can be obtained similarly.

Since w_0^ε is a solution to

$$-i\alpha_\varepsilon \varepsilon w_0^\varepsilon + \Delta w_0^\varepsilon + w_0^\varepsilon = S_0(x) + S_1\left(x - \frac{q_1}{\varepsilon}\right)$$

we may decompose $w_0^\varepsilon = \widetilde{w_0^\varepsilon} + a^\varepsilon$, where $\widetilde{w_0^\varepsilon}$ is defined above and a^ε satisfy

$$-i\alpha_\varepsilon \varepsilon a^\varepsilon + \Delta a^\varepsilon + a^\varepsilon = S_1\left(x - \frac{q_1}{\varepsilon}\right).$$

Since $\|\widetilde{w_0^\varepsilon}\|_{B^*} \leq C\|S_0\|_B$, the proof of point (ii) reduces to the proof of the following lemma.

LEMMA 3.3. *If a^ε is the solution to*

$$-i\alpha_\varepsilon \varepsilon a^\varepsilon + \Delta a^\varepsilon + a^\varepsilon = S_1\left(x - \frac{q_1}{\varepsilon}\right),$$

then a^ε is uniformly (in ε) bounded in $B^\star$:

$$\|a^\varepsilon\|_{B^*} \leq C\|S_1\|_B.$$

Proof. We want to prove that

$$\forall v \in B, \quad |\langle a^\varepsilon, v\rangle| \leq C\|S_1\|_B\|v\|_B.$$

Using Parseval's equality, we write

$$\langle a^\varepsilon, v\rangle = \int_{\mathbb{R}^3} \frac{e^{-i\frac{q_1\cdot\xi}{\varepsilon}}\widehat{S_1}(\xi)\bar{\hat{v}}(\xi)}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} d\xi. \tag{3.3}$$

To estimate this integral, we shall distinguish the values of ξ close to or far from two critical sets: the sphere $\{|\xi|^2 = 1\}$ (the set where the denominator in (3.3) vanishes when $\varepsilon \to 0$) and the line $\{\xi$ collinear to $q_1\}$ (the set where we cannot apply directly the stationary phase theorem to (3.3)).

More precisely, we first take a small parameter $\delta \in]0,1[$, and we distinguish in the integral (3.3) the contributions due to the values of ξ such that $|\xi^2 - 1| \geq \delta$ or $|\xi^2 - 1| \leq \delta$. Let $\chi \in \mathcal{C}_c^\infty(\mathbb{R})$ be a truncation function such that $\chi(\lambda) = 0$ for $|\lambda| \geq 1$. We denote $\chi_\delta(\xi) = \chi\big(\frac{|\xi|^2-1}{\delta}\big)$. We accordingly decompose

$$\begin{aligned}\langle a^\varepsilon, v\rangle &= \int_{\mathbb{R}^3} \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\widehat{S_1}(\xi)\bar{\hat{v}}(\xi)\chi_\delta(\xi)}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} d\xi + \int_{\mathbb{R}^3} \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\widehat{S_1}(\xi)\bar{\hat{v}}(\xi)(1-\chi_\delta(\xi))}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} d\xi \\ &= I^\varepsilon + II^\varepsilon.\end{aligned}$$

First, since the denominator is not singular on the support of χ_δ, we easily bound the first part with the L^2 norms,

$$|I^\varepsilon| \leq \frac{\|\chi\|_{L^\infty}}{\delta}\|\widehat{S_1}\|_{L^2}\|\hat{v}\|_{L^2},$$

and using $B \hookrightarrow L^2$, we obtain the desired bound:

$$|I^\varepsilon| \leq C\|S_1\|_B\|v\|_B. \tag{3.4}$$

Let us now study the second part, II^ε, where the denominator is singular. Up to a rotation, we may assume $q_1 = |q_1|e_1$, where e_1 is the first vector of the canonical base. We make the polar change of variables,

$$\xi = \begin{cases} r\sin\theta\cos\varphi, \\ r\sin\theta\sin\varphi, \\ r\cos\theta. \end{cases}$$

Hence, $q_1 \cdot \xi = |q_1| r\sin\theta\cos\varphi$, and we get

$$II^\varepsilon = \int \frac{e^{-i\frac{|q_1|}{\varepsilon}r\sin\theta\cos\varphi}}{-r^2+1-i\varepsilon\alpha_\varepsilon}\big(\widehat{S_1\bar{v}}(1-\chi_\delta)\big)(\xi(r,\theta,\varphi))r^2\sin\theta dr d\theta d\varphi.$$

Now, we distinguish the contributions to the integral $d\theta d\phi$ linked to the values close to or far from the critical direction $\{\theta = \frac{\pi}{2},\ \varphi = 0,\ \text{or } \varphi = \pi\}$ (which corresponds to the case $\{\xi \text{ collinear to } q_1\}$). To that purpose, let $\eta > 0$ be a small parameter and denote

$$\Omega_0 = \left\{(r,\theta,\varphi) \;\middle|\; \left(1-\chi\left(\frac{r^2-1}{\delta}\right)\right) \neq 0, \chi\left(\frac{\theta-\frac{\pi}{2}}{\eta}\right) \neq 0, \chi\left(\frac{\varphi}{\eta}\right) \neq 0\right\},$$
$$\Omega_\pi = \left\{(r,\theta,\varphi) \;\middle|\; \left(1-\chi\left(\frac{r^2-1}{\delta}\right)\right) \neq 0, \chi\left(\frac{\theta-\frac{\pi}{2}}{\eta}\right) \neq 0, \chi\left(\frac{\varphi-\pi}{\eta}\right) \neq 0\right\}.$$

Let $k_0,\ k_\pi \in \mathcal{C}_c^\infty$ be such that $(1-\chi_\delta)k_0(\theta,\varphi)$ is a localization function on Ω_0 and $(1-\chi_\delta)k_\pi(\theta,\varphi)$ is a localization function on Ω_π. We denote $k = k_0 + k_\pi$. We write

$$\begin{aligned} II^\varepsilon = &\int \frac{e^{-i\frac{|q_1|}{\varepsilon}r\sin\theta\cos\varphi}}{-r^2+1-i\varepsilon\alpha_\varepsilon}(\widehat{S_1\bar{v}})(\xi(r,\theta,\varphi))(1-\chi_\delta(r))k_0(\theta,\varphi)r^2\sin\theta dr d\theta d\varphi \\ &+ \int \frac{e^{-i\frac{|q_1|}{\varepsilon}r\sin\theta\cos\varphi}}{-r^2+1-i\varepsilon\alpha_\varepsilon}(\widehat{S_1\bar{v}})(\xi(r,\theta,\varphi))(1-\chi_\delta(r))k_\pi(\theta,\varphi)r^2\sin\theta dr d\theta d\varphi \\ &+ \int \frac{e^{-i\frac{|q_1|}{\varepsilon}r\sin\theta\cos\varphi}}{-r^2+1-i\varepsilon\alpha_\varepsilon}(\widehat{S_1\bar{v}})(\xi(r,\theta,\varphi))(1-\chi_\delta(r))(1-k(\theta,\varphi))r^2\sin\theta dr d\theta d\varphi, \end{aligned}$$

$$II^\varepsilon = III_0^\varepsilon + III_\pi^\varepsilon + IV^\varepsilon.$$

The two parts III_0^ε and III_π^ε being similar, we write only how to estimate III_0^ε. In order to translate the stationary point in $(0,0)$, we consider the new variable $\alpha = \theta - \frac{\pi}{2}$. The phase function is $rg(\alpha,\varphi) = r\cos\alpha\cos\varphi$, so

$$\frac{\partial g}{\partial \alpha} = -\sin\alpha\cos\varphi = 0 \ \text{ at } (\alpha,\varphi) = (0,0),$$
$$\frac{\partial g}{\partial \varphi} = -\cos\alpha\sin\varphi = 0 \ \text{ at } (\alpha,\varphi) = (0,0),$$

and the Hessian at the point $(0,0)$ is

$$D^2 g(0,0) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

We apply the Morse lemma: upon choosing $\eta > 0$ small enough, there exists a $\mathcal{C}^\infty$ change of variables on Ω_0, $(\alpha, \varphi) \mapsto (\alpha', \varphi')$, such that

$$g(\alpha, \varphi) = 1 - \frac{\alpha'^2}{2} - \frac{\varphi'^2}{2}.$$

Then, we make the change of variables $\alpha'' = \sqrt{\frac{r}{2}}\alpha'$, $\varphi'' = \sqrt{\frac{r}{2}}\varphi'$. Finally, we decompose $(1-\chi_\delta)k_0 = \chi^1\chi^2$, with $\chi^1,\ \chi^2 \in \mathcal{C}_c^\infty$. Thus, we obtain, for the contribution III_0^ε, the formula

$$III_0^\varepsilon = \int \frac{e^{i\frac{q_1}{\varepsilon}(-r+\alpha''^2+\varphi''^2)}}{-r+1+i\varepsilon\alpha_\varepsilon} \widehat{T^1 S_1}(r,\alpha'',\varphi'')\overline{\widehat{T^2 v}}(r,\alpha'',\varphi'')drd\alpha''d\varphi'', \tag{3.5}$$

where

$$\begin{aligned} T^1 S_1 &:= \mathcal{F}^{-1}\Big((\chi^1\widehat{S_1}) \circ \xi(r,\alpha(\alpha'',\varphi''),\varphi(\alpha'',\varphi''))\Big), \\ T^2 v &:= \mathcal{F}^{-1}\Bigg(\frac{-r+1+i\varepsilon\alpha_\varepsilon}{-r^2+1-i\varepsilon\alpha_\varepsilon}(\chi^2\hat{v}) \circ \xi(r,\alpha(\alpha'',\varphi''),\varphi(\alpha'',\varphi'')) \\ &\qquad \times \frac{2}{r}\Big|\frac{\mathrm{d}\xi}{\mathrm{d}(r,\alpha,\varphi)}\Big|\Big|\frac{\mathrm{d}(\alpha,\varphi)}{\mathrm{d}(\alpha',\varphi')}\Big|\Bigg). \end{aligned}$$

As a first step, using Proposition 2.2, we directly get $T^1 S_1 \in B$ with

$$\|T^1 S_1\|_B \leq C\|S_1\|_B.$$

As a second step, we study $T^2 v$. Since for r close to 1,

$$\Big|\frac{-r+1+i\varepsilon\alpha_\varepsilon}{-r^2+1-i\varepsilon\alpha_\varepsilon}\Big| \leq 1,$$

we recover, from Proposition 2.2,

$$T^2 v \in B \qquad \text{and} \qquad \|T^2 v\|_B \leq C\|v\|_B.$$

Now, we apply Parseval's equality with respect to the r variable in the formula (3.5)

$$\begin{aligned} III_0^\varepsilon &= \int \frac{e^{i\frac{|q_1|}{\varepsilon}(\alpha''^2+\varphi''^2)}}{-r+1+i\varepsilon\alpha_\varepsilon} T^1 S_1\left(\widehat{. - \frac{|q_1|}{\varepsilon}, ., .}\right)\overline{\widehat{T^2 v}}drd\alpha''d\varphi'' \\ &= \int e^{i\frac{|q_1|}{\varepsilon}(\alpha''^2+\varphi''^2)}\mathbf{1}_{\{t>0\}}e^{-(\varepsilon\alpha_\varepsilon - i)t}\mathcal{F}_{r\to\rho}\left(T^1 S_1\left(\widehat{. - \frac{|q_1|}{\varepsilon}, ., .}\right)\right)(\rho - t, \alpha'', \varphi'') \\ &\qquad \times \mathcal{F}_{r\to\rho}(\overline{\widehat{T^2 v}})(\rho,\alpha'',\varphi'')dtd\rho d\alpha''d\varphi'', \end{aligned}$$

where $\mathbf{1}_{\{t>0\}}$ denotes the characteristic function of the set $\{t > 0\}$. Hence, we obtain

$$\begin{aligned} |III_0^\varepsilon| &\leq \left(\int \left\|\mathcal{F}_{r\to\rho}\left(T^1 S_1\left(\widehat{. - \frac{|q_1|}{\varepsilon}, ., .}\right)\right)(\rho)\right\|_{L^2} d\rho\right) \\ &\qquad \times \left(\int \|\mathcal{F}_{r\to\rho}(\widehat{T^2 v})(\rho)\|_{L^2} d\rho\right), \\ |III^\varepsilon| &\leq \left(\int \left\|T^1 S_1\left(\rho - \frac{|q_1|}{\varepsilon}\right)\right\|_{L^2} d\rho\right)\left(\int \|T^2 v(\rho)\|_{L^2} d\rho\right), \\ |III^\varepsilon| &\leq \sum_{j=1}^n \left(\int \|T^1 S_1(\rho)\|_{L^2} d\rho\right)\left(\int \|T^2 v(\rho)\|_{L^2} d\rho\right). \end{aligned}$$

Now, using Proposition 2.1, we get

$$|III_0^\varepsilon| \le C\|T^1 S_1\|_B \|T^2 v\|_B,$$
$$|III_0^\varepsilon| \le C\|S_1\|_B \|v\|_B,$$

which is the desired estimate.

We are left with the part IV^ε, which corresponds to the directions ξ that are not collinear to q_1. We denote by K' the support of $(1-\chi_\delta)(1-k)$, which is a compact set. If we denote

$$\eta_1 = |\xi|^2 - 1, \quad \eta_2 = -q_1 \cdot \xi, \tag{3.6}$$

then

$$\frac{\mathrm{d}(\eta_1,\eta_2)}{\mathrm{d}\xi} = \begin{pmatrix} 2\xi \\ -q_1 \end{pmatrix}$$

is of maximal rank 2 for $\xi \in K'$. Hence, there exists a finite covering $(\Omega_j)_{j=1,m}$ ($m \in \mathbb{N}$) of K' such that in Ω_j, we can make the change of variables $\xi \mapsto \eta$, where η_1, η_2 are given by (3.6) and η_3 is one of the components of ξ (depending on Ω_j). We denote by $\chi_j = \chi_j^3 \chi_j^4$ some localization functions on Ω_j such that $(1-\chi_\delta)(1-k) = \sum_{j=1}^m \chi_j$. Thus, for $j = 1, \dots, m$,

$$\int \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}}{-|\xi|^2 + 1 + i\varepsilon\alpha_\varepsilon} \widehat{S_1 \bar{\hat{v}}} \chi_j d\xi = \int \frac{e^{i\frac{\eta_2}{\varepsilon}}}{-\eta_1 + i\varepsilon\alpha_\varepsilon} (\widehat{S_1 \bar{\hat{v}}} \chi_j)(\xi(\eta)) \left|\frac{\mathrm{d}\xi}{\mathrm{d}\eta}\right| d\eta.$$

If we denote

$$T_j^3 S_1 := \mathcal{F}^{-1}((\chi_j^3 \widehat{S_1}) \circ \xi),$$
$$T_j^4 v := \mathcal{F}^{-1}((\chi_j^4 \hat{v}) \circ \xi) \left|\frac{\mathrm{d}\xi}{\mathrm{d}\eta}\right|,$$

and if $\mathcal{F}_1$ denotes the Fourier transform with respect to the η_1 variable, Parseval's equality with respect to η_1 gives

$$\begin{aligned}\left| \int \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}}{-|\xi|^2 + 1 + i\varepsilon\alpha_\varepsilon} \widehat{S_1 \bar{\hat{v}}} \chi_j d\xi \right| &= (2\pi)^d \Bigg| \int \chi_{\{t>0\}} e^{-\varepsilon\alpha_\varepsilon t} (\mathcal{F}_1^{-1}(\widehat{T_j^3 S_1}))(x_1 - t) \\ &\qquad \times (\mathcal{F}_1^{-1}(\overline{\widehat{T_j^4 v}}))(x_1) e^{i\eta_2/\varepsilon} dt dx_1 d\eta_2 d\eta_3 \Bigg| \\ &\le C\|S_1\|_B \|v\|_B,\end{aligned}$$

using Proposition 2.1 again. Summing over j, we obtain

$$|IV^\varepsilon| \le C\|S_1\|_B \|v\|_B,$$

which ends the proof of the bound

$$|\langle a^\varepsilon, v\rangle| \le C\|S_1\|_B \|v\|_B. \qquad \square$$

3.2. Proof of Proposition 3.2. As before, we prove the result for the sequence (w_0^ε) only. As we did in the proof of Proposition 3.1, we write $w_0^\varepsilon = \widetilde{w_0^\varepsilon} + a^\varepsilon$. Since $\widetilde{w_0^\varepsilon}$ is the solution to a Helmholtz equation with constant index of refraction and fixed source, it converges weakly-$*$ to the outgoing solution w_0 to $\Delta w + w = S_0$. Hence, it suffices to show the following result.

LEMMA 3.4. *If $a^\varepsilon \in B^\star$ is the solution to*

$$-i\alpha_\varepsilon \varepsilon a^\varepsilon + \Delta a^\varepsilon + a^\varepsilon = S_1\left(x - \frac{q_1}{\varepsilon}\right),$$

then $a^\varepsilon \to 0$ in $B^\star$.

Proof. The proof of this result requires two steps (using a density argument):

1. For $v \in B$, we have the bound $\left|\langle a^\varepsilon, v\rangle\right| \leq C\|S_1\|_B\|v\|_B$.
2. If S_1 and v are smooth, then $\langle a^\varepsilon, v\rangle \to 0$.

The first point is exactly the result in Lemma 3.3. It remains to prove the convergence in the smooth case (the second point above). We write

$$\langle a^\varepsilon, v\rangle = \int_{\mathbb{R}^3} \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\widehat{S_1}(\xi)\bar{\hat{v}}(\xi)}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} d\xi.$$

We are thus left with the study of

$$R_\varepsilon(\psi) = \int_{\mathbb{R}^3} \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\psi(\xi)}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} d\xi, \tag{3.7}$$

where $\psi = \widehat{S_1}\widehat{\bar{v}}$ belongs to $\mathcal{S}(\mathbb{R}^3)$.

As in the proof of Lemma 3.3, we distinguish the contributions of various values of ξ. We shall use exactly the same partition, according to the values of ξ close to or far from the sphere $|\xi| = 1$ and collinear or not to q_1. We shall use the same notation for the various truncation functions.

We first separate the contributions of ξ such that $|\xi^2 - 1| \leq \delta$ and $|\xi^2 - 1| \geq \delta$ using the truncation function χ_δ:

$$\begin{aligned} R_\varepsilon(\psi) &= \int_{\mathbb{R}^3} \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\psi(\xi)\chi_\delta(\xi)}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} d\xi + \int_{\mathbb{R}^3} \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\psi(\xi)(1-\chi_\delta(\xi))}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} d\xi \\ &= I^\varepsilon + II^\varepsilon. \end{aligned}$$

In the support of χ_δ, since the denominator is not singular, we can apply the nonstationary phase method. Since $q_1 \neq 0$, we may assume $q_1^1 \neq 0$, and we have

$$\begin{aligned} I^\varepsilon &= \frac{\varepsilon}{iq_1^1}\int_{\mathbb{R}^3} e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\partial_{\xi_1}\left(\frac{\psi(\xi)\chi_\delta(\xi)}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon}\right)d\xi \\ &= \frac{\varepsilon}{iq_1^1}\int_{\mathbb{R}^3} e^{-i\frac{q_1}{\varepsilon}\cdot\xi}\left(\frac{\partial_{\xi_1}(\psi(\xi)\chi_\delta(\xi))}{-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon} - \frac{2\psi(\xi)\chi_\delta(\xi)\xi_1}{(-|\xi|^2 + 1 - i\varepsilon\alpha_\varepsilon)^2}\right). \end{aligned}$$

Hence, we obtain the bound

$$|I^\varepsilon| \leq \frac{\varepsilon}{|q_1^1|}\int_{\mathbb{R}^3}\left(\frac{1}{\delta}|\partial_{\xi_1}(\chi\psi)| + \frac{2}{\delta^2}|\xi_1\chi\psi|\right)d\xi.$$

Since $\partial_{\xi_1}(\chi\psi)$ and $\xi_1\chi\psi$ belongs to $\mathcal{S}$, we have, as $\varepsilon \to 0$,

$$I^\varepsilon \to 0.$$

Let us now study the second term, II^ε. As in section 3.1, we first decompose II^ε into the sum $III_0^\varepsilon + III_\pi^\varepsilon + IV^\varepsilon$. We then use the same changes of variables. It leads to the following formula for III_0^ε:

$$III_0^\varepsilon = \int \frac{e^{-i\frac{q_1}{\varepsilon}(r-\alpha''^2-\varphi''^2)}}{-r+1+i\varepsilon\alpha_\varepsilon}\widetilde{\chi}(r,\alpha'',\varphi'')\widetilde{\psi}(r,\alpha'',\varphi'')drd\alpha''d\varphi'',$$

where

$$\begin{aligned}\widetilde{\chi}(r,\alpha'',\varphi'') &= ((1-\chi_\delta)k_0)\circ\xi(r,\alpha(\alpha'',\varphi''),\varphi(\alpha'',\varphi''))\\ &\qquad\times\frac{2(-r+1+i\varepsilon\alpha_\varepsilon)}{r(-r^2+1-i\varepsilon\alpha_\varepsilon)}\left|\frac{\mathrm{d}(\alpha,\varphi)}{\mathrm{d}(\alpha',\varphi')}\right|,\\ \widetilde{\psi}(r,\alpha'',\varphi'') &= \psi\circ\xi(r,\alpha(\alpha'',\varphi''),\varphi(\alpha'',\varphi''))\end{aligned}$$

are still smooth functions that are bounded independently from ε.

Using Parseval's inequality with respect to the variables (α'',φ''), we obtain the bound

$$|III_0^\varepsilon| \le C\varepsilon\left|\int\frac{e^{-i\frac{|q_1|}{\varepsilon}r}e^{-i\varepsilon(\lambda^2+\mu^2)}}{-r+1+i\varepsilon\alpha_\varepsilon}\mathcal{F}_{\lambda,\mu}(\widetilde{\chi}\widetilde{\psi})drd\lambda d\mu\right|.$$

To obtain the convergence of III_0^ε, it remains to study an integral of the type

$$\int_{|r-1|\le\delta}\frac{e^{-i\frac{|q_1|}{\varepsilon}r}w(r)}{-r+1+i\varepsilon\alpha_\varepsilon}dr,\ \text{where } w\in\mathcal{S}.$$

This is done in the following lemma.

LEMMA 3.5. *For all $w\in\mathcal{S}$ and all $\theta\in(0,1)$, we have*

$$\int_{|r|\le\delta}\frac{e^{-i\frac{|q_1|}{\varepsilon}r}w(r)}{-r+i\varepsilon\alpha_\varepsilon}dr = -i\pi w(0)+O_{\varepsilon\to0}(\varepsilon^{-\theta}).$$

Using this lemma, we readily get the estimate

$$|III^\varepsilon| \le C\varepsilon^{1-\theta}\quad\forall\theta\in(0,1), \tag{3.8}$$

which proves that $III^\varepsilon\to0$ as $\varepsilon\to0$.

Remark. In order to make the calculations easier, we write this paper in dimension equal to 3, but the proof would be similar in any dimension $d\ge2$. For a general dimension $d\ge2$, we will obtain an estimate in $O(\varepsilon^{(d-1)/2-\theta})$ instead of $O(\varepsilon^{1-\theta})$ above. The result fails in dimension $d=1$.

It remains to give the following proof.

Proof of Lemma 3.5. We write

$$\begin{aligned}\int_{-\delta}^{\delta}\frac{e^{-i\frac{|q_1|}{\varepsilon}r}w(r)}{-r+i\varepsilon\alpha_\varepsilon}dr &= \int_{-\delta}^{\delta}\frac{e^{-i\frac{|q_1|}{\varepsilon}r}w(r)}{r^2+(\varepsilon\alpha_\varepsilon)^2}(r-i\varepsilon\alpha_\varepsilon)dr\\ &= -i\varepsilon\alpha_\varepsilon\int_{-\delta}^{\delta}\frac{e^{-i\frac{|q_1|}{\varepsilon}r}w(r)}{r^2+(\varepsilon\alpha_\varepsilon)^2}dr+\int_{-\delta}^{\delta}e^{-i\frac{|q_1|}{\varepsilon}r}\frac{rw(r)}{r^2+(\varepsilon\alpha_\varepsilon)^2}dr\\ &= I+II.\end{aligned}$$

We have

$$I = -i \int_{-\frac{\delta}{\varepsilon\alpha_\varepsilon}}^{\frac{\delta}{\varepsilon\alpha_\varepsilon}} \frac{e^{-i|q_1|\alpha_\varepsilon y} w(\varepsilon\alpha_\varepsilon y)}{y^2+1} dy \to -i\pi w(0)$$

and

$$II = \int_{-\delta}^{\delta} \big(e^{-i\frac{|q_1|}{\varepsilon}r} w(r) - w(0)\big) \frac{r}{r^2 + (\varepsilon\alpha_\varepsilon)^2} dr + \int_{-\delta}^{\delta} w(0) \frac{r}{r^2+(\varepsilon\alpha_\varepsilon)^2} dr \ .$$

The last term vanishes because the integrand is odd. Moreover, using the smoothness of w, we easily obtain that for all $\theta \in (0,1)$,

$$\big|e^{-i\frac{|q_1|}{\varepsilon}r} w(r) - w(0)\big| \le C_\theta \Big(\frac{r}{\varepsilon}\Big)^\theta .$$

Thus,

$$\left| \int_{-\delta}^{\delta} \big(e^{-i\frac{|q_1|}{\varepsilon}r} w(r) - w(0)\big) \frac{r}{r^2+(\varepsilon\alpha_\varepsilon)^2} dr \right| \le \frac{C}{\varepsilon^\theta} \int_{-\delta}^{\delta} |r|^{\theta-1} dr$$

and the result is proved. □

We are left with the study of IV^ε. We use the same change of variables as in section 3.1:

$$\begin{aligned} IV^\varepsilon &= \sum_{j=1}^m \int \frac{e^{-i\frac{q_1}{\varepsilon}\cdot\xi}}{-|\xi|^2 + 1 + i\varepsilon\alpha_\varepsilon} \psi(\xi)\chi_j(\xi) d\xi \\ &= \sum_{j=1}^m \int \frac{e^{i\frac{\eta_2}{\varepsilon}}}{-\eta_1 + i\varepsilon\alpha_\varepsilon} (\psi\chi_j)(\xi(\eta)) \Big|\frac{\mathrm{d}\xi}{\mathrm{d}\eta}\Big| d\eta \\ &= i\varepsilon \sum_{j=1}^m \int \frac{e^{i\frac{\eta_2}{\varepsilon}}}{-\eta_1 + i\varepsilon\alpha_\varepsilon} \partial_{\eta_2}\Big((\psi\chi_j)(\xi(\eta)) \Big|\frac{\mathrm{d}\xi}{\mathrm{d}\eta}\Big|\Big) d\eta. \end{aligned}$$

The integral obviously converges with respect to all the variables except η_1. It remains to prove the convergence with respect to the η_1 variable, i.e., the convergence of

$$\int \frac{\phi(\eta)}{-\eta_1 + i\varepsilon\alpha_\varepsilon} d\eta_1,$$

where

$$\phi = \partial_{\eta_2}\Big((\psi\chi_j)(\xi(\eta)) \Big|\frac{\mathrm{d}\xi}{\mathrm{d}\eta}\Big|\Big)$$

is smooth and compactly supported with respect to η. It is a consequence of the fact that the distribution $(x+i0)^{-1}$ is well defined on $\mathbb{R}$ by

$$\frac{1}{x+i0} = p.v.\left(\frac{1}{x}\right) - i\pi\delta(x).$$

We conclude that $IV^\varepsilon \to 0$ and $\langle a^\varepsilon, v\rangle \to 0$ as $\varepsilon \to 0$. □

4. Transport equation and radiation condition on μ. In this section, we state and prove our results on the Wigner measure associated with (u^ε). Since we established the uniform bounds on (u^ε) and the convergence of (w_0^ε), (w_1^ε), these results now essentially follow from the results proved in [2]. We first prove bounds on the sequence of Wigner transforms (W^ε) that allow us to define a Wigner measure μ associated to (u^ε). Then, we get the transport equation satisfied by μ together with the radiation condition at infinity, which uniquely determines μ.

4.1. Results.

THEOREM 4.1. *Let $S_0,\ S_1 \in B$ and $\lambda > 0$. The sequence (W^ε) is bounded in the Banach space $X_\lambda^\star$, and up to extracting a subsequence, it converges weakly-$*$ to a positive and locally bounded measure μ such that*

$$\sup_{R>0} \frac{1}{R} \int_{|x|<R} \int_{\xi\in\mathbb{R}^3} \mu(x,\xi)\, dxd\xi \leq C(\|S_0\|_B + \|S_1\|_B)^2. \tag{4.1}$$

The Banach space X_λ^* is defined as the dual space of the set X_λ of functions $\hat\varphi(x,\xi)$ such that $\varphi(x,y) := \mathcal{F}_{\xi\to y}(\hat\varphi(x,\xi))$ satisfies

$$\int_{\mathbb{R}^3} \sup_{x\in\mathbb{R}^3} (1+|x|+|y|)^{1+\lambda} |\varphi(x,y)| dy < \infty. \tag{4.2}$$

THEOREM 4.2. *Assume* (H1), (H2). *Then the Wigner measure μ associated with (u^ε) satisfies the transport equation*

$$\xi\cdot\nabla_x\mu = \frac{1}{(4\pi)^2}\Big(\delta(x)|\widehat{S_0}(\xi)|^2 + \delta(x-q_1)|\widehat{S_1}(\xi)|^2\Big)\delta(|\xi|^2-1) := Q(x). \tag{4.3}$$

Moreover, μ is the outgoing solution to (4.3) *in the following sense: for all test functions $R \in \mathcal{C}_c^\infty(\mathbb{R}^6)$, if we denote $g(x,\xi) = \int_0^\infty R(x-\xi t,\xi)dt$, then*

$$\int_{\mathbb{R}^6} R(x,\xi) d\mu(x,\xi) = -\int_{\mathbb{R}^6} Q(x,\xi) g(x,\xi) dxd\xi. \tag{4.4}$$

Remark. Here the support of the test function R contains 0, contrary to [2].

4.2. Proof of Theorem 4.1. This theorem, which is proved in [2], is a consequence of the uniform estimate on the sequence (u^ε) in the space B^* obtained in Proposition 3.1. We observe that for any $\lambda > 0$,

$$\|\langle x\rangle^{-\frac12-\lambda} u^\varepsilon(x)\|_{L^2} \leq C\|u^\varepsilon\|_{B^*} \leq C\left(\|S_0\|_B + (\|S_1\|_B\right); \tag{4.5}$$

hence, for any function φ satisfying (4.2), we have

$$\begin{aligned}
&|\langle W^\varepsilon(u^\varepsilon), \hat\varphi\rangle| \\
&\leq \int_{\mathbb{R}^6} \frac{|u^\varepsilon|(x+\frac{\varepsilon}{2}y)|\overline{u^\varepsilon}|(x-\frac{\varepsilon}{2}y)}{\left\langle x+\frac{\varepsilon}{2}y\right\rangle^{\frac12+0}\left\langle x-\frac{\varepsilon}{2}y\right\rangle^{\frac12+0}} \left\langle x+\frac{\varepsilon}{2}y\right\rangle^{\frac12+0}\left\langle x-\frac{\varepsilon}{2}y\right\rangle^{\frac12+0} |\varphi|(x,y)dxdy \\
&\leq C\left(\|S_0\|_B + \|S_1\|_B\right)^2 \int_{\mathbb{R}^3} \sup_{x\in\mathbb{R}^3} \langle|x|+|y|\rangle^{1+0} |\varphi(x,y)| dy.
\end{aligned}$$

So $(W^\varepsilon(u^\varepsilon))$ is bounded in X_λ^*, $\lambda > 0$. We deduce that, up to extracting a subsequence, $(W^\varepsilon(u^\varepsilon))$ converges weakly-$*$ to a nonnegative measure μ satisfying

$$|\langle\mu,\hat\varphi\rangle| \leq C\left(\|S_0\|_B + \|S_1\|_B\right)^2 \int_{\mathbb{R}^3} \sup_{x\in\mathbb{R}^3} \langle|x|+|y|\rangle^{1+0} |\varphi(x,y)| dy. \tag{4.6}$$

We refer, for instance, to Lions and Paul [9] for the proof of the nonnegativity of μ.

The bound (4.1) is obtained using the family of functions

$$\varphi_\delta^R(x,y) = \frac{1}{\delta^{3/2}} e^{-|y|^2/\delta} \frac{1}{R} \chi(\langle x \rangle \leq R)$$

and letting $\delta \to 0$, $R \to \infty$. □

4.3. Proof of the transport equation (4.3). This section is devoted to the proof of the transport equation satisfied by μ. We first write the transport equation satisfied by W^ε in a dual form. Then we study the convergence of the source term (the convergence of the other terms is obvious). Finally, choosing an appropriate test function in the limiting process, we get the radiation condition at infinity satisfied by μ. Proving first a localization property, we improve the radiation condition proved in [2].

4.3.1. Transport equation satisfied by W^ε. W^ε satisfies the equation

$$\alpha_\varepsilon W^\varepsilon + \xi \cdot \nabla_x W^\varepsilon = Q^\varepsilon, \tag{4.7}$$

where, for $\psi \in \mathcal{S}(T^*\mathbb{R}^d)$, if $\varphi(x,y) = \mathcal{F}^{-1}_{y \to \xi}(\psi(x,\xi))$,

$$\begin{aligned}
\langle Q^\varepsilon, \psi \rangle &= \frac{i\varepsilon}{2} \mathcal{I}m \, \langle W^\varepsilon(S^\varepsilon, u^\varepsilon), \psi \rangle \\
&= \frac{i}{2} \mathcal{I}m \left(\int_{\mathbb{R}^6} \overline{w_0^\varepsilon}(x+y) S_0^\varepsilon(y) \varphi \left(\varepsilon \left(x + \frac{y}{2} \right), y \right) dxdy \right) \\
&+ \frac{i}{2} \mathcal{I}m \left(\int_{\mathbb{R}^6} \overline{w_1^\varepsilon}(x+y) S_1^\varepsilon(y) \varphi \left(q_1 + \varepsilon \left(x + \frac{y}{2} \right), y \right) dxdy \right).
\end{aligned}$$

This equation can be obtained writing first the equation satisfied by

$$v^\varepsilon(x,y) = u^\varepsilon \left(x + \frac{\varepsilon}{2} y \right) \overline{u^\varepsilon} \left(x - \frac{\varepsilon}{2} y \right).$$

From the equality

$$\nabla_y \cdot \nabla_x v^\varepsilon = \frac{\varepsilon}{2} \left[\Delta u^\varepsilon \left(x + \frac{\varepsilon}{2} y \right) \overline{u^\varepsilon} \left(x - \frac{\varepsilon}{2} y \right) - \Delta \overline{u^\varepsilon} \left(x - \frac{\varepsilon}{2} y \right) u^\varepsilon \left(x + \frac{\varepsilon}{2} y \right) \right],$$

we deduce

$$\alpha_\varepsilon v^\varepsilon + i \nabla_y \cdot \nabla_x v^\varepsilon + \frac{i}{2\varepsilon} \left[n^2 \left(x + \frac{\varepsilon}{2} y \right) - n^2 \left(x - \frac{\varepsilon}{2} y \right) \right] v^\varepsilon = \sigma_\varepsilon(x,y),$$

where

$$\sigma_\varepsilon(x,y) := \frac{i\varepsilon}{2} \left[S^\varepsilon \left(x + \frac{\varepsilon}{2} y \right) \overline{u^\varepsilon} \left(x - \frac{\varepsilon}{2} y \right) - \overline{S^\varepsilon} \left(x - \frac{\varepsilon}{2} y \right) u^\varepsilon \left(x + \frac{\varepsilon}{2} y \right) \right].$$

After a Fourier transform, we obtain (4.7).

Then we write the dual form of this equation. Letting $\psi \in \mathcal{S}(\mathbb{R}^6)$, we have

$$\alpha_\varepsilon \langle W^\varepsilon, \psi \rangle - \langle W^\varepsilon, \xi \cdot \nabla_x \psi \rangle = \langle Q^\varepsilon, \psi \rangle. \tag{4.8}$$

By the definition of the Wigner measure μ, we get

$$\alpha_\varepsilon \langle W^\varepsilon, \psi \rangle \to 0 \qquad \text{and} \qquad \langle W^\varepsilon, \xi \cdot \nabla_x \psi \rangle \to \langle \mu, \xi \cdot \nabla_x \psi \rangle.$$

Hence we are left with the study of the source term $\langle Q^\varepsilon, \psi \rangle$.

4.3.2. Convergence of the source term. In order to compute the limit of the source term in (4.7), we develop

$$\langle Q^\varepsilon,\psi\rangle = \frac{i\varepsilon}{2}\mathcal{I}m\Big(\langle W^\varepsilon(S_0^\varepsilon,u^\varepsilon),\psi\rangle + \langle W^\varepsilon(S_1^\varepsilon,u^\varepsilon),\psi\rangle\Big).$$

Thus, the result is contained in the following proposition.

PROPOSITION 4.3. *The sequences* $\big(\varepsilon W^\varepsilon(S_0^\varepsilon,u^\varepsilon)\big)$ *and* $\big(\varepsilon W^\varepsilon(S_1^\varepsilon,u^\varepsilon)\big)$ *are bounded in* $\mathcal{S}'(\mathbb{R}^6)$*, and for all real-valued* $\psi\in\mathcal{S}(\mathbb{R}^6)$*, we have*

$$\lim_{\varepsilon\to0}\varepsilon\langle W^\varepsilon(S_0^\varepsilon,u^\varepsilon),\psi\rangle_{\mathcal{S}',\mathcal{S}} = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3}\overline{\widehat{w_0}}(\xi)\widehat{S_0}(\xi)\psi(0,\xi)d\xi, \tag{4.9}$$

$$\lim_{\varepsilon\to0}\varepsilon\langle W^\varepsilon(S_1^\varepsilon,u^\varepsilon),\psi\rangle_{\mathcal{S}',\mathcal{S}} = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3}\overline{\widehat{w_1}}(\xi)\widehat{S_1}(\xi)\psi(q_1,\xi)d\xi, \tag{4.10}$$

where w_0 *and* w_1 *are defined in Proposition* 3.2.

Using Proposition 4.3, we readily get, for any real-valued test function ψ,

$$\begin{aligned}\lim_{\varepsilon\to0}\langle Q^\varepsilon,\psi\rangle &= \frac{i}{2(2\pi)^3}\mathcal{I}m\left(\int_{\mathbb{R}^3}\overline{\widehat{w_0}}(\xi)\widehat{S_0}(\xi)\psi(0,\xi)d\xi+\int_{\mathbb{R}^3}\overline{\widehat{w_1}}(\xi)\widehat{S_1}(\xi)\psi(q_1,\xi)d\xi\right)\\ &= \frac{1}{(4\pi)^2}\bigg(\int_{\mathbb{R}^3}|\widehat{S_0}(\xi)|^2\delta(\xi^2-1)\psi(0,\xi)d\xi\\ &\qquad\qquad+\int_{\mathbb{R}^3}|\widehat{S_1}(\xi)|^2\delta(\xi^2-1)\psi(q_1,\xi)d\xi\bigg),\end{aligned}$$

which is the result in Theorem 4.2. ∎

Let us now prove Proposition 4.3.

Proof of Proposition 4.3. The two terms to study being of the same type, we consider only the first one in our proof. Let $\psi\in\mathcal{S}(T^*\mathbb{R}^d)$ and $\varphi(x,y)=\mathcal{F}^{-1}_{y\to\xi}(\psi(x,\xi))$. Then we have

$$\begin{aligned}\varepsilon\langle W^\varepsilon(S_0^\varepsilon,u^\varepsilon),\psi\rangle_{\mathcal{S}',\mathcal{S}} &= \varepsilon\int S_0^\varepsilon\left(x+\frac{\varepsilon}{2}y\right)\overline{u^\varepsilon}\left(x-\frac{\varepsilon}{2}y\right)\varphi(x,y)dxdy\\ &= \int S_0(x)\overline{w_0^\varepsilon}(x+y)\varphi\left(\varepsilon\left(x+\frac{y}{2}\right),y\right)dxdy.\end{aligned}$$

As a first step, let us prove that $\varepsilon\langle W^\varepsilon(S_0^\varepsilon,u^\varepsilon),\psi\rangle_{\mathcal{S}',\mathcal{S}}$ is bounded. Using the fact that $\psi\in\mathcal{S}(\mathbb{R}^{2d})$, we get

$$\begin{aligned}\big|\varepsilon\langle W^\varepsilon(S_0^\varepsilon,u^\varepsilon),\psi\rangle_{\mathcal{S}',\mathcal{S}}\big| &\le C\int\langle x\rangle^N|S_0(x)|\frac{|w_0^\varepsilon(x+y)|}{\langle x+y\rangle^\beta}\frac{\langle x+y\rangle^\beta}{\langle x\rangle^N\langle y\rangle^k}dxdy\\ &\le C\|\langle x\rangle^N S_0\|_{L^2}\|w_0^\varepsilon\|_{B^*}\int_{\mathbb{R}^3_y}\sup_{x\in\mathbb{R}^3}\frac{\langle x+y\rangle^\beta}{\langle x\rangle^N\langle y\rangle^k}dy\end{aligned}$$

for any $k\ge0$ and $\frac12<\beta<N$, upon using the Cauchy–Schwarz inequality in x. Then we distinguish the cases $|x|\le|y|$ and $|x|\ge|y|$: the term stemming from the first case gives a contribution which is bounded by $C\int\frac{dy}{\langle y\rangle^{k-\beta}}$, and the second contribution is bounded by $C\int\frac{dy}{\langle y\rangle^k}$. Hence, upon choosing k large enough, we obtain

$$|\varepsilon\langle W^\varepsilon(S_0^\varepsilon,u^\varepsilon),\psi\rangle_{\mathcal{S}',\mathcal{S}}|\le C\|\langle x\rangle^N S_0\|_{L^2}\|w^\varepsilon\|_{B^*}.$$

As a second step, we compute the limit (4.9). We write

$$\begin{aligned}\varepsilon\langle W^\varepsilon(S_0^\varepsilon,u^\varepsilon),\psi\rangle &= \int S_0(x)\overline{w_0^\varepsilon}(x+y)\Big(\varphi\Big(\varepsilon\Big(x+\frac{y}{2}\Big),y\Big)-\varphi(0,y)\Big)dxdy\\ &\quad+\int \overline{w_0^\varepsilon}(x)S_0(x-y)\varphi(0,y)dxdy\\ &= I_\varepsilon + II_\varepsilon.\end{aligned}$$

Reasoning as above, we readily get that $\lim_{\varepsilon\to 0} I_\varepsilon = 0$. Indeed, since $\varphi\in\mathcal{S}(\mathbb{R}^{2d})$, we have, for all $k\in\mathbb{N}$ and $x,\ y\in\mathbb{R}^d$,

$$\Big|\varphi\Big(\varepsilon\Big(x+\frac{y}{2}\Big),y\Big)-\varphi(0,y)\Big| \le C\varepsilon\frac{|x|+|y|}{\langle y\rangle^k} \le C\varepsilon\frac{\langle|x|+|y|\rangle}{\langle y\rangle^k}.$$

Hence,

$$\begin{aligned}|I^\varepsilon| &\le C\varepsilon\int\langle x\rangle^N|S_0(x)|\frac{|w_0^\varepsilon(x+y)|}{\langle x+y\rangle^\beta}\frac{\langle|x|+|y|\rangle^{\beta+1}}{\langle x\rangle^N\langle y\rangle^k}dxdy\\ &\le C\varepsilon\|\langle x\rangle^N S_0\|_{L^2}\|w_0^\varepsilon\|_{B^*}\int_{\mathbb{R}^3_y}\sup_{x\in\mathbb{R}^3}\frac{\langle|x|+|y|\rangle^{\beta+1}}{\langle x\rangle^N\langle y\rangle^k}dy\end{aligned}$$

for any $k\ge 0$ and $1/2<\beta<N-1$. As above, the previous integral converges for k large enough. Therefore, $I^\varepsilon\to 0$.

We end the proof by proving that the second term, II^ε, converges to $\int S_0(x)\overline{w_0}(x+y)\widehat{\psi}(0,y)dxdy$. We have

$$II_\varepsilon = \int \overline{w_0^\varepsilon}(x)\big(S_0*\varphi(0,.)\big)(x)dx.$$

Hence, since w_0^ε converges weakly-$*$ in B^*, it suffices to prove that $S_0*\varphi(0,.)$ belongs to B. We denote $\phi=\varphi(0,.)$. Then, $\phi\in\mathcal{S}(\mathbb{R}^3)$. Let $1/2<\beta<N$. We have, using (2.5),

$$\|S_0*\phi\|_B^2 \le C\|S_0*\phi\|_{L^2_\beta}^2 = C\int\langle x\rangle^{2\beta}|S_0*\phi(x)|^2dx.$$

Moreover, upon using the Cauchy–Schwarz inequality, we get, for all $x\in\mathbb{R}^3$,

$$\begin{aligned}|S_0*\phi(x)|^2 &\le \left(\int_{\mathbb{R}^d_y}|S_0(x-y)||\phi(y)|dy\right)^2\\ &\le \left(\int_{\mathbb{R}^d_y}|S_0(x-y)|^2|\phi(y)|dy\right)\left(\int_{\mathbb{R}^d_y}|\phi(y)|dy\right)\\ &= \|\phi\|_{L^1}|S_0|^2*|\phi|(x).\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\|S_0*\phi\|_B^2 &\le C\|\phi\|_{L^1}\int\langle x\rangle^{2\beta}|S_0(x-y)|^2|\phi(y)|dydx\\ &\le C\|\langle x\rangle^N S_0\|_{L^2}^2\int_{\mathbb{R}^3_y}\sup_{x\in\mathbb{R}^3}\frac{\langle x+y\rangle^{2\beta}}{\langle x\rangle^{2N}\langle y\rangle^k}dy\end{aligned}$$

for any k. As before, this integral converges. Thus, we have established that $S_0 * \widehat{\psi}(0,.)$ belongs to B, which implies that

$$II_\varepsilon \to \int S_0(x)\overline{w_0}(x+y)\widehat{\psi}(0,y)dxdy. \qquad \square$$

4.4. Proof of the radiation condition (4.4). It remains to prove that μ satisfies the weak radiation condition (4.4).

4.4.1. Support of μ. In order to prove the radiation condition without restriction on the test function R (as assumed in [2]), we first prove a localization property on the Wigner measure μ. This property is well known when u^ε satisfies a Helmholtz equation without source term. It is still valid here thanks to the scaling of S^ε.

PROPOSITION 4.4. *Under hypotheses* (H1), (H2)*, the Wigner measure μ satisfies*

$$\mathrm{supp}(\mu) \subset \{(x,\xi) \in \mathbb{R}^6 / \ |\xi|^2 = 1\}.$$

Proof. Let $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^6)$ and $\phi^\varepsilon = \phi^W(x, \varepsilon D_x)$. Let us denote $H^\varepsilon = -\varepsilon^2\Delta - 1$. Since u^ε satisfies the Helmholtz equation (1.1), we have

$$i\alpha_\varepsilon \varepsilon u^\varepsilon + H^\varepsilon u^\varepsilon = \varepsilon^2 S^\varepsilon. \tag{4.11}$$

Moreover, H^ε is a pseudodifferential operator with symbol $|\xi|^2 - 1$. By pseudodifferential calculus, $\phi^\varepsilon H^\varepsilon = Op_\varepsilon^W(\phi(x,\xi)(|\xi|^2-1)) + O(\varepsilon)$; thus, using the definition of the measure μ, we get that

$$\begin{aligned}\lim_{\varepsilon\to 0}(\phi^\varepsilon H^\varepsilon u^\varepsilon, u^\varepsilon) &= \lim_{\varepsilon\to 0}(Op_\varepsilon^W(\phi(x,\xi)(|\xi|^2-1))u^\varepsilon, u^\varepsilon) \\ &= \int \phi(x,\xi)(|\xi|^2-1)d\mu.\end{aligned}$$

Using (4.11), we write

$$(\phi^\varepsilon H^\varepsilon u^\varepsilon, u^\varepsilon) = \varepsilon^2(\phi^\varepsilon S^\varepsilon, u^\varepsilon) - i\alpha_\varepsilon\varepsilon(\phi^\varepsilon u^\varepsilon, u^\varepsilon) = \varepsilon^2(W^\varepsilon(S^\varepsilon, u^\varepsilon), \phi) - i\alpha_\varepsilon\varepsilon(\phi^\varepsilon u^\varepsilon, u^\varepsilon).$$

On the one hand, Proposition 4.3 gives that $\lim_{\varepsilon\to 0}\varepsilon^2(W^\varepsilon(S^\varepsilon,u^\varepsilon),\phi) = 0$. On the other hand, $(\phi^\varepsilon u^\varepsilon, u^\varepsilon)$ is bounded, so $\lim_{\varepsilon\to 0}\alpha_\varepsilon\varepsilon(\phi^\varepsilon u^\varepsilon, u^\varepsilon) = 0$. Therefore, for any $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^6)$, we have $\int \phi(|\xi|^2-1)d\mu = 0$, so $\mathrm{supp}(\mu) \subset \{|\xi|^2 = 1\}$. $\square$

4.4.2. Proof of the condition (4.4). Using the previous localization property, in order to prove the radiation condition (4.4), one may only use test functions $R \in \mathcal{C}_c^\infty(\mathbb{R}^6)$ such that $\mathrm{supp}(R) \subset \mathbb{R}^6 \backslash \{\xi = 0\}$.

Let R be such a test function. We associate with R the solution g^ε to

$$-\alpha_\varepsilon g^\varepsilon + \xi \cdot \nabla_x g^\varepsilon = R(x,\xi).$$

By duality, we have

$$\langle Q^\varepsilon, g^\varepsilon\rangle = \langle W^\varepsilon, R\rangle,$$

so that it suffices to establish the following two convergences:

$$\lim_{\varepsilon\to 0}\langle Q^\varepsilon, g^\varepsilon\rangle = \langle Q, g\rangle, \tag{4.12}$$

$$\lim_{\varepsilon\to 0}\langle W^\varepsilon, R\rangle = \langle f, R\rangle, \tag{4.13}$$

where Q and g are defined in Theorem 4.2.

As before, since $R \in X_\lambda$ for any $\lambda > 0$, the limit (4.13) follows from the weak-$*$ convergence of W^ε in $X_\lambda^\star$.

On the other hand,

$$\langle Q^\varepsilon, g^\varepsilon \rangle = \mathcal{I}m \int_{\mathbb{R}^6} \overline{S_0}(x) w_0^\varepsilon(x+y) \widehat{g^\varepsilon}\left(\varepsilon\left[x+\frac{y}{2}\right], y\right) dxdy + \mathcal{I}m \int_{\mathbb{R}^6} \overline{S_1}(x) w_1^\varepsilon(x+y) \widehat{g^\varepsilon}\left(q_1 + \varepsilon\left[x+\frac{y}{2}\right], y\right) dxdy, \tag{4.14}$$

so $\langle Q^\varepsilon, g^\varepsilon \rangle$ is the sum of two terms of the same type. Such a term has been studied in [2], where the following result is proved.

PROPOSITION 4.5. *Assume that (w^ε) is bounded in B^* and that (w^ε) converges weakly-$*$ in B^* to w_0. Assume that S_0 satisfy* (H2). *Let $R \in \mathcal{C}_c^\infty(\mathbb{R}^6)$ be such that* $\mathrm{supp}(R) \subset \mathbb{R}^6 \backslash \{\xi = 0\}$. *Let g^ε be the solution to*

$$-\alpha_\varepsilon g^\varepsilon + \xi \cdot \nabla_x g^\varepsilon = R(x,\xi)$$

and $g(x,\xi) = \int_0^\infty R(x+t\xi,\xi)dt$. Then we have

$$\lim_{\varepsilon\to 0} \int_{\mathbb{R}^6} \overline{S_0}(x) w_0^\varepsilon(x+y) \widehat{g^\varepsilon}\left(\varepsilon\left[x+\frac{y}{2}\right], y\right) dxdy = \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \overline{\widehat{S_0}}(\xi)\widehat{w_0}(\xi) g(0,\xi) d\xi.$$

Using the proposition above together with Proposition 3.2, we get that

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \langle Q^\varepsilon, g^\varepsilon \rangle \\
&= \mathcal{I}m\left(\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \overline{\widehat{S_0}}(\xi)\widehat{w_0}(\xi) g(0,\xi) d\xi + \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \overline{\widehat{S_1}}(\xi)\widehat{w_1}(\xi) g(q_1,\xi) d\xi \right) \\
&= \frac{1}{(4\pi)^2} \left(\int_{\mathbb{R}^3} |\widehat{S_0}(\xi)|^2 \delta(\xi^2 - 1) g(0,\xi) d\xi + \int_{\mathbb{R}^3} |\widehat{S_1}(\xi)|^2 \delta(\xi^2-1) g(q_1,\xi) d\xi \right).
\end{aligned}$$

Thus, the radiation condition (4.4) is proved.

Acknowledgments. I would like to thank my advisor François Castella for having guided this work. I would also like to thank the referee for having suggested the short proof of the estimate of u^ε. Finally, I would like to thank Patrick Gérard for having pointed out to me that the measures μ_0 and μ_1 are mutually singular, using a dimension argument which is analogous to that of our proof so that the result follows from the orthogonality property.

REFERENCES

[1] S. AGMON AND L. HÖRMANDER, *Asymptotic properties of solutions of differential equations with simple characteristics*, J. Anal. Math., 30 (1976), pp. 1–38.

[2] J.D. BENAMOU, F. CASTELLA, T. KATSAOUNIS, AND B. PERTHAME, *High frequency limit of the Helmholtz equation*, Rev. Math. Iberoamericana, 18 (2002), pp. 187–209.

[3] N. BURQ, *Mesures semi-classiques et mesures de défaut*, Séminaire Bourbaki, Vol. 1996/97, Astérisque, 245 (1997), pp. 167–195.

[4] F. CASTELLA, B. PERTHAME, AND O. RUNBORG, *High frequency limit of the Helmholtz equation. Source on a general manifold*, Comm. Partial Differential Equations, 27 (2002), pp. 607–651.

[5] L. ERDÖS AND H.T. YAU, *Linear Boltzmann equation as scaling limit of the quantum Lorentz gas*, in Advances in Differential Equations and Mathematical Physics (Atlanta, GA, 1997), Contemp. Math. 217, AMS, Providence, RI, 1998, pp. 137–155.

[6] P. Gérard, *Mesures semi-classiques et ondes de Bloch*, in Séminaire sur les Équations aux Dérivées Partielles 1988–1989, Exp. XVI, Ecole Polytechnique, Palaiseau, France, 1988.
[7] P. Gérard, P.A. Markowich, N.J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 321–357.
[8] L. Hörmander, *The Analysis of Linear Partial Differential Operators.* I, Springer-Verlag, Berlin, 1983.
[9] P.-L. Lions and T. Paul, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.
[10] L. Miller, *Refraction of high-frequency waves density by sharp interfaces and semiclassical measures at the boundary*, J. Math. Pures Appl. (9), 79 (2000), pp. 227–269.
[11] G. Papanicolaou and L. Ryzhik, *Waves and Transport*, in Hyperbolic Equations and Frequency Interactions, IAS/Park City Math. Ser. 5, AMS, Providence, RI, 1999, pp. 305–382.
[12] B. Perthame and L. Vega, *Morrey-Campanato estimates for the Helmholtz equation*, J. Funct. Anal., 164 (1999), pp. 340–355.
[13] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.
[14] E. Wigner, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 637–656

WEAK SHOCKS FOR A ONE-DIMENSIONAL BGK KINETIC MODEL FOR CONSERVATION LAWS*

C. M. CUESTA† AND C. SCHMEISER‡

Abstract. For one-dimensional kinetic BGK models, regarded as relaxation models for scalar conservation laws with genuinely nonlinear fluxes, existence of small amplitude traveling waves is proven. Dynamic stability of these kinetic shock profiles is shown by extending a classical energy method for viscous regularizations of conservation laws.

Key words. weak shocks, kinetic BGK model, micro-macro decomposition, scalar conservation laws

AMS subject classifications. 35Q99, 82C99, 35L65, 35A99, 35B40

DOI. 10.1137/050638059

1. Introduction.

In this paper we study small amplitude traveling wave solutions of the following one-dimensional Bhatnagar–Gross–Krook (BGK) type equation:

$$\partial_t f + v\partial_x f = M(\rho_f, v) - f\,, \quad \text{with } t > 0\,,\ x \in \mathbb{R}\,,\ v \in \Omega\,. \tag{1.1}$$

Here $f(t,x,v)$ can be interpreted (in analogy with the Boltzmann equation) as a time-dependent phase space density of particles with time t, position x, and velocity v. We shall assume that $\Omega \subset \mathbb{R}$ is the support of a measure $d\mu(v)$. In particular, both continuous velocity distributions as well as discrete velocity models, where (1.1) is a hyperbolic system, are included in our assumptions.

The function $\rho_f(t,x)$ in (1.1) is the macroscopic density corresponding to the distribution f, i.e., the zeroth order velocity moment

$$\rho_f(t,x) = \int f(t,x,v)d\mu(v)\,. \tag{1.2}$$

Here and in the following we refrain from writing Ω under the integral sign in integrals with respect to the measure $d\mu(v)$. Note that in the case of a discrete velocity model, Ω is a discrete set, and the integral above is a sum. The "Maxwellian" $M(\rho,v)$ is an equilibrium distribution satisfying the moment conditions

$$\int M(\rho,v)d\mu(v) = \rho \quad \text{and} \quad \int v\, M(\rho,v)d\mu(v) = a(\rho) \tag{1.3}$$

for a macroscopic flux function $a(\rho)$ that will be assumed smooth and genuinely nonlinear, actually (without loss of generality) concave: $a''(\rho) < 0$. The properties (1.3) ensure, at least formally, that the macroscopic limit equation (scaling with $(t,x) \to (t/\varepsilon, x/\varepsilon)$ and taking $\varepsilon \to 0$) of (1.1) is the scalar conservation law

$$\partial_t \rho + \partial_x a(\rho) = 0\,. \tag{1.4}$$

*Received by the editors August 12, 2005; accepted for publication (in revised form) February 1, 2006; published electronically June 23, 2006.
http://www.siam.org/journals/sima/38-2/63805.html
†Faculty of Mathematics, University of Vienna, Nordbergstrasse 15, 1090 Vienna, Austria (pmxcc@maths.nottingham.ac.uk).
‡Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria (christian.schmeiser@tuwien.ac.at).

It is well known that initial value problems for (1.4) do not possess smooth solutions in general, and that weak solutions are not unique. Uniqueness can be obtained by considering (1.4) as the limit of an appropriately regularized problem. Classically this is done by introducing an artificial viscosity and carrying out the limit $\nu \to 0^+$ in

$$\partial_t \rho + \partial_x a(\rho) = \nu \partial_x^2 \rho \,; \tag{1.5}$$

see, e.g., [10]. In this work, instead of (1.5), the kinetic regularization (1.1) is studied.

Typical weak solutions of (1.4) are shock waves of the form

$$\rho(t,x) = \begin{cases} \rho_- & \text{if } x - st < x_0 \,, \\ \rho_+ & \text{if } x - st > x_0 \,, \end{cases}$$

where the constants $\rho_\pm$ and the wave speed s are related by the Rankine–Hugoniot condition

$$s = \frac{a(\rho_+) - a(\rho_-)}{\rho_+ - \rho_-} \,. \tag{1.6}$$

The admissibility condition

$$\frac{a(\rho) - a(\rho_-)}{\rho - \rho_-} - s > 0 \quad \text{for all } \rho \in (\min(\rho_+, \rho_-), \max(\rho_+, \rho_-)) \tag{1.7}$$

can be derived by constructing viscous profiles, i.e., traveling wave solutions of (1.5). In this framework, (1.7) gives a necessary and sufficient condition for existence of traveling wave solutions connecting the values $\rho_\pm$ at $x - st = \pm\infty$. For the concave flux functions $a(\rho)$ considered here, (1.7) reduces to the condition $\rho_- < \rho_+$. This is called an *entropy condition* since it can also be obtained from the (distributional) entropy inequality

$$\partial_t \phi(\rho) + \partial_x \psi(\rho) \le 0 \,,$$

which can be derived for every convex entropy density $\phi(\rho)$ and corresponding entropy flux $\psi(\rho)$ (satisfying $\psi' = \phi' a$) in the limit $\nu \to 0$ from (1.5).

An entropy inequality can also be derived for solutions of the kinetic equation (1.1) under an additional structure condition on the equilibrium distribution. We shall assume that the Maxwellian is a smooth and strictly increasing function of ρ:

$$\partial_\rho M(\rho, v) > 0 \,. \tag{1.8}$$

Then there exists a function $\theta(f,v)$ such that $f = M(\rho,v)$ is equivalent to $\rho = \theta(f,v)$. With the primitive $\Theta(f,v)$ ($\partial_f \Theta = \theta$), solutions of (1.1) formally satisfy the entropy inequality

$$\begin{aligned} &\partial_t \int \Theta(f,v) d\mu(v) + \partial_x \int v\, \Theta(f,v)\, d\mu(v) \\ &= \int (M(\rho_f, v) - f)(\theta(f,v) - \rho_f)\, d\mu(v) \le 0 \,. \end{aligned}$$

In the context of relaxation systems, condition (1.8) can be seen as a subcharacteristic condition. It can be used for proving stability results such as a total variation

diminishing property corresponding to that for entropy solutions of the macroscopic equation (1.4); see [1], [5]. A class of examples of Maxwellians $M(\rho, v)$ satisfying the moment conditions (1.3) as well as (1.8) has been given by the authors in [3]:

$$M(\rho, v) = \int_0^\rho m(v - a'(r))dr\,,$$

where $m(v) > 0$ for $v \in \mathbb{R}$ is an even function satisfying $\int_{-\infty}^{\infty} m(v)d\mu(v) = 1$ ($\Omega = \mathbb{R}$, $d\mu(v) = dv$).

It is our aim to study small amplitude kinetic shock profiles as traveling wave solutions of kinetic models of the form (1.1). Assuming (1.8), we shall prove their existence under the same entropy condition as required for the viscous regularization in section 3. This is no surprise considering that our constructive existence proof shows asymptotic closeness of viscous and kinetic profiles for small shocks. A main ingredient of the proof is a fluid-kinetic (or micro-macro) decomposition in the spirit of the one introduced by Caflisch and Nicolaenko [2] for the gas dynamics Boltzmann equation.

A well-known kinetic model for scalar conservation laws is the Perthame–Tadmor model [9]. There the Maxwellian is a discontinuous function. This lack of smoothness is an obstacle for the study of small waves by perturbation arguments as carried out here. Existence of big traveling waves has been studied by compactness arguments in [4]. The same approach has been carried out for (1.1) by the authors of this work [3]. In this parallel, the results of the present study are reviewed, and the existence result is extended to large amplitude waves. As opposed to the results here, the existence proof for large waves is nonconstructive, and their stability is still open.

In section 4, local dynamic stability of the constructed traveling waves is proven. Again, a micro-macro decomposition (now in the spirit of Liu and Yu [7]) is at the heart of the argument. A classical energy method for proving stability on the macroscopic level is combined with entropy estimates for the kinetic perturbations.

In the remainder of this section, we present the formal asymptotics for the construction of small amplitude waves as well as the energy method for proving stability of viscous profiles.

Formal construction of kinetic shock profiles. We look for solutions of (1.1), whose dependence on x and t is only through the traveling wave variable $\xi = x - st$, with s being the wave speed:

$$(v - s)\partial_\xi f = M(\rho_f, v) - f\,, \qquad \xi \in \mathbb{R},\ v \in \Omega\,, \tag{1.9}$$

subject to the far-field conditions

$$f(\pm\infty, v) = M(\rho^\pm, v)\,, \qquad v \in \Omega\,. \tag{1.10}$$

We are interested in small amplitude waves and assume

$$\rho^+ - \rho^- = \varepsilon \qquad \text{with } 0 < \varepsilon \ll 1\,. \tag{1.11}$$

The positivity of ε reflects the entropy condition (1.7). It turns out that it is appropriate to rescale the traveling wave variable by $\xi \to \xi/\varepsilon$, to get

$$\varepsilon(v - s)\partial_\xi f = M(\rho_f, v) - f, \qquad \xi \in \mathbb{R},\ v \in \Omega\,. \tag{1.12}$$

The Rankine–Hugoniot condition (1.6) is derived as a necessary condition for existence by integrating (1.12) with respect to v and by (1.3).

The formal asymptotics below is a variant of the Chapman–Enskog expansion procedure. We start by introducing the decomposition

$$f = M(\rho_f, v) + \varepsilon^2 f^\perp, \quad \text{with} \int f^\perp d\mu(v) = 0\,, \tag{1.13}$$

of the solution into an equilibrium part and a remainder (or into a macroscopic and a microscopic contribution, or into a fluid and a kinetic part). As a next step, the traveling wave equation (1.12) is integrated with respect to v and ξ:

$$\int v f \, d\mu(v) - s\rho_f = a(\rho_-) - s\rho_-\,. \tag{1.14}$$

Essentially, this equation is considered as an equation for the macroscopic density ρ_f, and the full kinetic equation (1.12) should determine $f^\perp$. The smallness of the wave is reflected in the fact that the macroscopic density is everywhere close to its far-field value at $\xi = -\infty$ and that the wave speed is close to the characteristic speed there:

$$\rho_f = \rho_- + \varepsilon u, \quad s = a'(\rho_-) + \varepsilon\sigma\,. \tag{1.15}$$

Substitution of this and (1.13) into (1.12) and (1.14) give the leading order ($O(\varepsilon^2)$) term equation

$$f^\perp = -(v - a'(\rho_-))\partial_\rho M(\rho_-, v)\partial_\xi u\,, \tag{1.16}$$

$$-\int v f^\perp d\mu(v) = \frac{a''(\rho_-)}{2} u^2 - \sigma u\,. \tag{1.17}$$

The limiting version of the Rankine–Hugoniot condition (1.6) is given by $\sigma = a''(\rho_-)/2 < 0$. After elimination of $f^\perp$, this becomes the traveling wave equation of the viscous Burgers equation

$$D_0\,\partial_\xi u = -\sigma u(1-u)\,, \tag{1.18}$$

with the diffusivity

$$D_0 = \int (v - a'(\rho_-))^2\,\partial_\rho M(\rho_-, v)\, d\mu(v) > 0\,, \tag{1.19}$$

by $\partial_\rho M > 0$. Obviously, solutions of (1.18) connecting $u = 0$ at $\xi = -\infty$ to $u = 1$ at $\xi = +\infty$ exist. The lack of uniqueness due to the translation invariance of the traveling wave problem will be an issue below.

It is far from obvious how to make this argument rigorous, since (1.16) for $f^\perp$ is a singular limit and, even worse, its solution is a differentiation problem. In the existence proof in section 3 we adapt an idea from Caflisch and Nicolaenko [2], where existence of weak shock profiles for the Boltzmann equation of gas dynamics has been proven. It is based on a slight modification of the micro-macro decomposition such that the fluid and the kinetic terms satisfy a system of equations with separated derivatives.

Stability of viscous shock profiles. In section 4 we prove local dynamic stability of small amplitude traveling waves. The idea is to decouple the equation into a macroscopic part and a small microscopic part. Then we use L^2-type energy (actually entropy) methods for the macroscopic equation, which can be extended to also control the microscopic part. Similar techniques have been used by Liu and Yu [7] for the Boltzmann equation. For the Broadwell model, a discrete velocity model for the Boltzmann equation, energy estimates have also been used in [6].

We expand briefly on the ideas behind the L^2-estimates at the macroscopic level. If ϕ is a traveling wave solution of the diffusive regularization (1.5), then the perturbation $\tilde{\rho} = \rho - \phi$ satisfies

$$\partial_t \tilde{\rho} - s\partial_\xi \tilde{\rho} + \partial_\xi (a(\phi + \tilde{\rho}) - a(\phi)) = \nu \partial_\xi^2 \tilde{\rho}.$$

Linearizing this equation and testing it with $\tilde{\rho}$ produces a term with the wrong sign, which is not possible to control. The usual trick (see, e.g., [8]) for overcoming this problem is to introduce the new macroscopic unknown

$$W(\xi, t) = \int_{-\infty}^{\xi} \tilde{\rho}(x, t)\, dx \tag{1.20}$$

after choosing the shift in ϕ such that $\int_{\mathbb{R}} (\rho - \phi) d\xi = 0$. Testing the integrated perturbation equation

$$\partial_t W - s\partial_\xi W + a(\phi + \tilde{\rho}) - a(\phi) = \nu \partial_\xi^2 W\,,$$

with W, gives in particular the term

$$\int_{\mathbb{R}} (a(\phi + \tilde{\rho}) - a(\phi)) W \; d\xi = -\frac{1}{2} \int_{\mathbb{R}} \partial_\xi (a'(\phi)) W^2 \, d\xi + n.l.t.$$

(*n.l.t.* stands for *nonlinear terms*). By the monotonicity of the wave profile the term on the right-hand side is positive, indicating decay of the L^2-norm of W if the nonlinear terms can be controlled.

The idea now consists of combining the energy estimates for W and for $\tilde{\rho} = \partial_\xi W$ to get an estimate on the H^1-norm of W. Clearly in both cases the contribution of the diffusion term has the good sign. This way we can also control the term with the wrong sign for $\tilde{\rho}$ by the term coming from diffusion for W. The basic estimate reads

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}} \left(W^2 + \gamma (\partial_\xi W)^2 \right) \, d\xi \leq - \left(\nu - C_0 \sup |W| - \gamma C_1 \right) \int_{\mathbb{R}} (\partial_\xi W)^2 \, d\xi \tag{1.21}$$

for some arbitrary $\gamma > 0$. Here C_0 and C_1 are positive constants depending on pointwise bounds for the density ρ which are a consequence of the maximum principle. The supremum norm of W is controlled by the H^1-norm in one dimension. Hence, starting with initial data such that $\sup |W(t = 0)|$ is small enough, and choosing γ small enough, the right-hand side of (1.21) is negative initially and remains so. This implies global existence for $W \in H^1(\mathbb{R})$, as well as stability of macroscopic traveling waves. To achieve an analogous result for the kinetic equation, the argument will be similar, but one has to take care of the contribution of the microscopic part, which shall, however, be small by assumption (1.11). Additional difficulties will rise from the absence of a maximum principle requiring estimates for $\tilde{\rho}$ in H^1 (W in H^2) for pointwise control of $\tilde{\rho}$, as well as from the fact that the monotonicity of the macroscopic density of the kinetic traveling wave is not obvious.

2. Notation and assumptions. Since we shall linearize around the state $M(\rho_-, v)$, we introduce the notation $F(v) := \partial_\rho M(\rho_-, v)$ for simplicity. We shall sometimes skip the dependence on v in the function M, and write $M'(\rho)$ instead of $\partial_\rho M(\rho, v)$ (i.e., $F = M'(\rho_-)$).

We shall work in the weighted Hilbert space L^2_v of functions of the velocity, defined by the scalar product

$$\langle f, g\rangle_v = \int \frac{fg}{F} d\mu(v)\,,$$

where $\|\cdot\|_v$ denotes the induced norm. We also consider the L^2- and H^k-norms for functions of ξ. We write these spaces as L^2_ξ and H^k_ξ, and their norms as $\|\cdot\|_\xi$ and $\|\cdot\|_{H^k}$, respectively.

The Hilbert space $L^2_{\xi,v}$ is then naturally defined by the scalar product

$$\langle f, g\rangle_{\xi,v} = \int_{\mathbb{R}} \langle f, g\rangle_v d\xi\,,$$

with the induced norm $\|\cdot\|_{\xi,v}$. Similarly, we shall denote by $H^k_\xi(L^2_v)$ the space of functions with derivatives with respect to ξ up to order k in L^2_v, and the corresponding norm

$$\|f\|_{H^k_\xi(L^2_v)} = \left(\|f\|^2_{\xi,v} + \cdots + \|\partial^k_\xi f\|^2_{\xi,v}\right)^{1/2}\,. \tag{2.1}$$

The linearization of the collision operator on the right-hand side of (1.1) around $M(\rho_-, v)$ is given by

$$\mathcal{L}f := F\rho_f - f\,. \tag{2.2}$$

It is symmetric and negative semidefinite in L^2_v. These properties are easily seen from the identity

$$\langle \mathcal{L}f_1, f_2\rangle_v = -\frac{1}{2}\iint FF'\left(\frac{f_1}{F} - \frac{f_1'}{F'}\right)\left(\frac{f_2}{F} - \frac{f_2'}{F'}\right) d\mu(v)\, d\mu(v')\,,$$

where $'$ denotes evaluation at v'. The entropy inequality $\langle \mathcal{L}f, f\rangle_v \le 0$ is a straightforward consequence.

Apart from the essential requirements (1.3) and (1.8), our existence and stability proofs rely on additional technical assumptions on M. For fixed $v \in \Omega$, we assume that $M(\rho, v)$ is a C^3-function of ρ. Moreover, for a given ρ, up to fourth order moments of the derivatives exist, i.e.,

$$\int \left|v^m \partial^k_\rho M(\rho, v)\right| d\mu(v) < \infty \qquad \text{for } k \le 3,\ m \le 4\,. \tag{2.3}$$

For the second and third order derivatives of M with respect to ρ we require that for given ρ_1 and ρ_2,

$$\int |v|^m \frac{(\partial^k_\rho M(\rho_1, v))^2}{\partial_\rho M(\rho_2, v)}\, d\mu(v) < \infty \qquad \text{for } k \le 3,\ m \le 4\,. \tag{2.4}$$

As a consequence of (2.3), up to second order moments of velocity distributions can be bound by their L^2_v-norm:

$$\left|\int (v-s)^m f\, d\mu(v)\right| \le \left(\int |v-s|^{2m} F\, d\mu(v)\right)^{1/2} \|f\|_v \qquad \text{for } m \le 2\,. \tag{2.5}$$

This will be used repeatedly in the following. For simplicity we also adopt the notation

$$\hat{D} := \int (v-s)^2 F \, d\mu(v) \,. \tag{2.6}$$

Finally, we assume that for fixed ρ, $\partial_\rho M(\rho, v)$ is a continuous function of $v \in \Omega$.

3. Existence of small amplitude traveling waves.

3.1. An approximate solution. In this section we prove existence of solutions of (1.12) subject to (1.10) for $\varepsilon \ll 1$. We start by returning to the problem of constructing a formal approximation. Instead of formally passing to the limit as in section 1, we avoid expansion errors wherever possible and produce a residual whose v-integral vanishes. We start with the ansatz

$$f_{as} := M(\rho, v) + \varepsilon^2 f^\perp \,, \tag{3.1}$$

formally resembling (1.13). The residual is then given by

$$\varepsilon^3 h := M(\rho_{as}) - f_{as} - \varepsilon(v-s)\partial_\xi f_{as} \,, \qquad \text{with } \rho_{as} = \rho + \varepsilon^2 \rho^\perp \,. \tag{3.2}$$

Recalling (1.16), we eliminate two terms in the right-hand side by the choice

$$f^\perp := -\frac{1}{\varepsilon}(v-s)\partial_\xi M(\rho, v) \,. \tag{3.3}$$

Finally, the requirement that the v-integral of h vanish and that f_{as} satisfy the far-field conditions (1.10) leads to an ordinary differential equation (ODE) for ρ:

$$\frac{a(\rho) - a(\rho_-) - s(\rho - \rho_-)}{\varepsilon^2} = \frac{1}{\varepsilon}\left(\int (v-s)^2 M'(\rho, v) d\mu(v) \right) \partial_\xi \rho \tag{3.4}$$

subject to

$$\rho(-\infty) = \rho_- \quad \text{and} \quad \rho(+\infty) = \rho_+ \,. \tag{3.5}$$

With $\rho = \rho_- + \varepsilon u$ the problem for u formally tends to (1.18) in the limit $\varepsilon \to 0$. Actually, since the diffusivity $D(\rho) := \int (v-s)^2 M'(\rho, v)\, d\mu(v)$ is obviously positive, (3.4) has the same qualitative properties as (1.18), and a solution of (3.4), (3.5) exists, which is determined uniquely by the condition

$$\rho(0) = \frac{\rho_- + \rho_+}{2} \,. \tag{3.6}$$

It is easily shown that u and $\partial_\xi u$ are uniformly bounded as $\varepsilon \to 0$ and for $\xi \in \mathbb{R}$, and, therefore, the same holds for ρ and $\partial_\xi \rho/\varepsilon$. As a consequence, $D(\rho)$ is uniformly bounded away from zero. Division of (3.4) by $D(\rho)$ and differentiation shows that also $\partial_\xi^k \rho/\varepsilon$ is uniformly bounded for $k = 2, 3$ (here assumption (2.3) is used). Furthermore, the convergence of all these terms as $\xi \to \pm\infty$ is exponential. Recalling $s = a'(\rho_-) + O(\varepsilon)$,

$$\rho^\perp = -\frac{1}{\varepsilon}(a'(\rho) - s)\partial_\xi \rho = O(\varepsilon)$$

holds uniformly for $\xi \in \mathbb{R}$. This shows that the scaling of the residual in (3.2) has been chosen correctly in terms of the sup-norm:

$$h = \frac{M(\rho + \varepsilon^2 \rho^\perp) - M(\rho)}{\varepsilon^3} - (v-s)\partial_\xi f^\perp \tag{3.7}$$

is uniformly bounded in ε and ξ and decays exponentially as $\xi \to \pm\infty$. However, we shall need the following result also in other norms.

LEMMA 3.1. *Let the assumptions* (2.3) *and* (2.4) *be satisfied and let* f_{as} *be determined by* (3.1) *and* (3.3)–(3.5). *Then* f_{as} *satisfies the far-field conditions* (1.10), *and the traveling wave equation* (1.12) *up to the residual* $\varepsilon^3 h$, *where* h *is in* $H^1_\xi(L^2_v)$ *uniformly in* ε, *and* $\int h\, d\mu(v) = 0$.

Proof. The far-field conditions and the last statement are a direct consequence of the construction of f_{as}.

The boundedness of the first term on the right-hand side of (3.7) in $H^1_\xi(L^2_v)$ is a consequence of our observations above and of (2.4) ($m = 0$, $k = 2$). For the second term (2.4) is used with $m = 4$, $k = 2$. Of course the exponential decay of all terms suffices for integrability with respect to ξ. ∎

3.2. The micro-macro decomposition of the correction term. In terms of the correction term $\varepsilon^2 g = f - f_{as}$, the traveling wave problem reads

$$\begin{aligned} &\varepsilon(v-s)\partial_\xi g - \mathcal{L}g \\ &= (M'(\rho_{as}) - F)\rho_g + \frac{M(\rho_{as}+\varepsilon^2\rho_g) - M(\rho_{as}) - \varepsilon^2 M'(\rho_{as})\rho_g}{\varepsilon^2} + \varepsilon h \end{aligned} \tag{3.8}$$

subject to

$$g(\pm\infty, v) = 0 \quad \text{for all } v \in \Omega\,. \tag{3.9}$$

The left-hand side of (3.8) is the linearization of the traveling wave equation (1.12) around $M(\rho_-)$ with the linearized collision operator $\mathcal{L}$ defined in (2.2). The right-hand side contains an $O(\varepsilon)$ linear correction (since we should actually linearize around $M(\rho_{as})$), an $O(\varepsilon^2\rho_g^2)$ nonlinear term, and the residual. The homogeneous far-field conditions and Lemma 3.1 imply, after integration of (3.8) with respect to ξ, that the flux of g vanishes:

$$\int (v-s) g\, d\mu(v) = 0\,. \tag{3.10}$$

The problem (3.8), (3.9) will be solved in several steps. First, we introduce a splitting of g into a macroscopic part and a microscopic part. Then, in the following two subsections, we solve the linear equations associated to the decomposition of g, and finally solve the nonlinear problem.

In the first step, two ideas from the work by Caflisch and Nicolaenko [2] on the Boltzmann equation will be adapted to the present situation. The first one is a special micro-macro decomposition defined by

$$g(\xi, v) = z(\xi)\Phi(v) + \varepsilon w(\xi, v)\,, \tag{3.11}$$

where $\Phi := F\big(1 + \varepsilon\frac{\sigma}{\hat{D}}(v-s)\big)$, and the orthogonality condition $\langle (v-s)\Phi, w\rangle_v = 0$ holds.

The choice of the coefficient $\sigma/\hat{D}$ (see (1.15) and (2.6) for the definition of the constants) in front of the correction term in Φ guarantees that Φ shares the property (3.10) with g:

$$\int (v-s)\Phi\, d\mu(v) = 0\,. \tag{3.12}$$

This and the definition of the decomposition imply several properties of z and w.

LEMMA 3.2. *If g satisfies* (3.8), (3.9), *then*

$$w(\pm\infty, v) \equiv 0, \ z(\pm\infty) = 0\,, \tag{3.13}$$

and

$$\int (v-s)w(\xi,v)\,d\mu(v) = 0, \qquad \int (v-s)^2 w(\xi,v)\,d\mu(v) = 0 \quad \textit{for all } \xi \in \mathbb{R}\,. \tag{3.14}$$

Substitution of (3.11) into (3.8) and division by ε gives

$$(v-s)\Phi\,\partial_\xi z + \varepsilon(v-s)\partial_\xi w - \mathcal{L}w - \Lambda z = \varepsilon\Gamma\rho_w + \varepsilon R(\rho_g) + h\,, \tag{3.15}$$

where again the right-hand side contains a linear correction, the nonlinearity, and the residual, with

$$\Lambda := \frac{M'(\rho_{as})\rho_\Phi - \Phi}{\varepsilon}\,, \qquad \Gamma := \frac{M'(\rho_{as}) - F}{\varepsilon}\,,$$
$$R(\rho_g) := \frac{1}{\varepsilon^4}[M(\rho_{as} + \varepsilon^2\rho_g) - M(\rho_{as}) - \varepsilon^2 M'(\rho_{as})\rho_g]\,.$$

These terms are formally $O(1)$ such that the ε-powers in (3.15) reflect the expected orders of magnitude.

Observe that, in terms of z and w, the nonlinearity should be written as $R(\rho_g) = R(z\rho_\Phi + \varepsilon\rho_w)$ with $\rho_\Phi = 1 - \varepsilon^2\sigma^2/\hat{D}$ (a constant). The identities $\int \Lambda\,d\mu(v) = \int \Gamma\,d\mu(v) = \int R(\rho_g)\,d\mu(v) = 0$ hold.

In order to get an equation for z (the macroscopic equation), we apply an approximation of the macroscopic projection to (3.15), i.e., we multiply (3.15) by $(v-s)$ and integrate with respect to v:

$$\tilde{D}\,\partial_\xi z - r(\xi)z = \varepsilon\frac{a'(\rho_{as}) - s}{\varepsilon}\rho_w + \varepsilon\int (v-s)R(\rho_g)\,d\mu(v) + \int (v-s)h\,d\mu(v)\,, \tag{3.16}$$

with

$$\tilde{D} := \int (v-s)^2\Phi\,d\mu(v) = D_0 + O(\varepsilon) > 0\,, \tag{3.17}$$

$$r(\xi) := \int (v-s)\Lambda(\xi,v)\,d\mu(v) = \frac{a'(\rho_{as}(\xi)) - s}{\varepsilon}\rho_\Phi\,. \tag{3.18}$$

Here we have used (3.12) and Lemma 3.2. This equation already reveals the magic of the micro-macro decomposition (3.11). It does not contain derivatives of w, and actually becomes independent of w as $\varepsilon \to 0$. The formal limit is the linearization of the viscous Burgers traveling wave equation (1.18). In particular, $r(\xi) = \sigma(2u-1) + O(\varepsilon)$ and, consequently, there exist $\gamma, \bar{\xi} > 0$ such that

$$r(\xi) \le -\gamma \quad \text{for } \xi \ge \bar{\xi}\,, \qquad r(\xi) \ge \gamma \quad \text{for } \xi \le -\bar{\xi}\,. \tag{3.19}$$

Now an equation for w (the microscopic equation) is derived by substituting (3.16) into (3.15), which actually amounts to applying the projection

$$Pf := f - \frac{(v-s)\Phi}{\tilde{D}}\int (v-s)f\,d\mu(v) \tag{3.20}$$

to (3.15):

$$\varepsilon(v-s)\partial_\xi w - \mathcal{L}w - P\Lambda z = \varepsilon\tilde{\Gamma}\rho_w + \varepsilon PR(\rho_g) + Ph\,, \tag{3.21}$$

with

$$\tilde{\Gamma} = P\Gamma - \frac{(v-s)\Phi}{\tilde{D}}\int (v-s)F\,d\mu(v) = \Gamma - \frac{(v-s)\Phi}{\tilde{D}}(a'(\rho_{as}) - s)\,.$$

We make one more manipulation to get a final equation for w. This corresponds to the second idea from [2]. As we observed in section 2 the operator $\mathcal{L}$ is symmetric negative semidefinite, but not strictly negative. We introduce a new operator $\mathcal{M}$, which is strictly negative and coincides with $\mathcal{L}$ on the set of functions w satisfying the property (3.14):

$$\mathcal{M}w := \mathcal{L}w - (v-s)^2 F\int (v-s)^2 w\,d\mu(v)\,.$$

LEMMA 3.3. *The operator $\mathcal{M}$ is symmetric and negative definite in L^2_v, i.e., there exists a $\kappa > 0$ such that*

$$-\langle \mathcal{M}w, w\rangle_v > \kappa\|w\|_v^2 \qquad \textit{for all } w \in L^2_v\,.$$

Proof. The symmetry follows from the symmetry of $\mathcal{L}$ and from

$$\langle \mathcal{M}w_1, w_2\rangle_v = \langle \mathcal{L}w_1, w_2\rangle_v - \int (v-s)^2 w_1\,d\mu(v)\int (v-s)^2 w_2\,d\mu(v)\,.$$

To prove that $\mathcal{M}$ is negative definite, we write $w = F\rho_w + w^\perp$ and observe that $\mathcal{L}w = -w^\perp$:

$$\begin{aligned}
-\langle \mathcal{M}w, w\rangle_v &= \|w^\perp\|_v^2 + \left(D_0\,\rho_w + \int (v-s)^2 w^\perp d\mu(v)\right)^2 \\
&= \|w^\perp\|_v^2 + \gamma D_0^2\,\rho_w^2 + (1-\gamma)D_0^2\,\rho_w^2 \\
&\quad + 2D_0\,\rho_w\int (v-s)^2 w^\perp\,d\mu(v) + \left(\int (v-s)^2 w^\perp\,d\mu(v)\right)^2
\end{aligned}$$

for $\gamma \in (0,1)$. Hence

$$\begin{aligned}
-\langle \mathcal{M}w, w\rangle_v &= \|w^\perp\|_v^2 + \gamma D_0^2\rho_w^2 \\
&\quad + (1-\gamma)\left[D_0\,\rho_w + \frac{1}{1-\gamma}\int (v-s)^2 w^\perp\,d\mu(v)\right]^2 - \frac{\gamma}{1-\gamma}\left(\int (v-s)^2 w^\perp\,d\mu(v)\right)^2 \\
&\geq \gamma D_0^2\,\rho_w^2 + \|w^\perp\|_v^2\left(1 - \frac{\gamma}{1-\gamma}\|(v-s)^2 F\|_v^2\right) \geq \kappa(\rho_w^2 + \|w^\perp\|_v^2) = \kappa\|w\|_v^2\,,
\end{aligned}$$

with $\kappa > 0$ for γ small enough. Here we have used (2.5) with $m = 2$. ∎

We shall prove existence of solutions of equations (3.16) and

$$\varepsilon(v-s)\partial_\xi w - \mathcal{M}w - zP\Lambda = \varepsilon\tilde{\Gamma}\rho_w + \varepsilon PR(\rho_g) + Ph \tag{3.22}$$

subject to (3.13). The relation to the original problem (3.8), (3.9) is not obvious.

LEMMA 3.4. *The function* $g = z\Phi + \varepsilon w$ *is a solution of* (3.8), (3.9) *iff* z *and* w *solve* (3.16), (3.22) *subject to* (3.13).

Proof. The problem (3.16), (3.22), (3.13) has been derived from (3.8), (3.9) using the property (3.14) of solutions of the latter. The proof relies on showing that (3.14) also holds for solutions of (3.16), (3.22), (3.13) without requiring it as a side condition.

The properties of Φ imply that for an $f(v)$ satisfying $\int f\,d\mu(v) = 0$, also $\int Pf\,d\mu(v) = 0$ and $\int (v-s)Pf\,d\mu(v) = 0$ hold. Since the v-integrals of Λ, Γ, $R(\rho_g)$, and h vanish, these terms do note contribute when we integrate (3.22) and its product with $v - s$ with respect to v:

$$\varepsilon\partial_\xi \int (v-s)w\,d\mu(v) = -D_0 \int (v-s)^2 w\,d\mu(v)\,,$$
$$\varepsilon\partial_\xi \int (v-s)^2 w\,d\mu(v) = -\int (v-s)w\,d\mu(v) - \int (v-s)^3 F\,d\mu(v) \int (v-s)^2 w\,d\mu(v)\,.$$

This is a system of linear ODEs with constant coefficients for the unknowns $\int (v-s)w\,d\mu(v)$ and $\int (v-s)^2 w\,d\mu(v)$. The decay of w at $\xi = \pm\infty$ implies homogeneous far-field conditions for these quantities and, thus, $\int (v-s)w\,d\mu(v) \equiv \int (v-s)^2 w\,d\mu(v) \equiv 0$. ∎

3.3. The linear problem. In this section we prove solvability of (3.16) and (3.22) regarding the right-hand sides as given inhomogeneities. In particular, we look for solutions of

$$\varepsilon(v-s)\partial_\xi w - \mathcal{M}w = h_w \qquad \text{with } h_w \in H^1_\xi(L^2_v) \tag{3.23}$$

and

$$\partial_\xi z - r(\xi)z = h_z \qquad \text{with } h_z \in L^2_\xi\,. \tag{3.24}$$

We shall look for solutions in the same spaces as the inhomogeneities. This will replace homogeneous far-field conditions in the following. Whereas this requirement provides uniqueness for the solution of (3.23), it permits a one-parameter set of solutions of (3.24). This reflects the arbitrary shift in traveling wave solutions. Uniqueness will be guaranteed by the additional requirement

$$z(0) = z_0\,, \tag{3.25}$$

where $z_0 \in \mathbb{R}$ parametrizes the set of solutions.

LEMMA 3.5. *Let* z *be the solution of* (3.24), (3.25) *with* r *bounded and satisfying* (3.19). *Then there exists a positive constant* C, *such that*

$$\|z\|_{H^1_\xi} \le C(|z_0| + \|h_z\|_\xi)\,.$$

Proof. The solution of (3.24), (3.25) is given by

$$z(\xi) = E(\xi,0)z_0 + \int_0^\xi E(\xi,y)h_z(y)dy \qquad \text{with} \quad E(\xi,y) = \exp\left(\int_y^\xi r(\eta)d\eta\right).$$

For $|\xi| < \bar{\xi}$, $E(\xi,y)$ is bounded and thus, obviously,

$$\|z\|_{L^2_\xi(-\bar{\xi},\bar{\xi})} \le C\|h_z\|_{L^2_\xi(-\bar{\xi},\bar{\xi})}\,.$$

For $\xi > \bar{\xi}$, by (3.19), we have

$$|z(\xi)| \le E(\bar{\xi},0)e^{\gamma(\bar{\xi}-\xi)}|z_0| + \int_0^{\bar{\xi}} E(\bar{\xi},y)e^{\gamma(\bar{\xi}-\xi)}|h_z(y)|dy + \int_{\bar{\xi}}^{\xi} e^{\gamma(y-\xi)}|h_z(y)|dy$$
$$\le Ce^{-\gamma\xi}(|z_0| + \|h_z\|_{L^2_\xi(0,\bar{\xi})}) + z_1(\xi)\,,$$

where z_1 solves $\partial_\xi z_1 = -\gamma z_1 + |h_z|$, with $z_1(\bar{\xi}) = 0$. Multiplying by z_1 and integrating over $(\bar{\xi},\infty)$ gives $\|z_1\|_{L^2_\xi(\bar{\xi},\infty)} \le \frac{1}{\gamma}\|h_z\|_{L^2_\xi(\bar{\xi},\infty)}$, and hence

$$\|z\|_{L^2_\xi(\bar{\xi},\infty)} \le C(|z_0| + \|h_z\|_{L^2_\xi(0,\infty)})\,.$$

The interval $(-\infty,-\bar{\xi})$ is treated analogously, completing the estimation of $\|z\|_\xi$.

The estimate on $\|\partial_\xi z\|_\xi$ is an obvious consequence of the differential equation and the boundedness of r. ∎

THEOREM 3.6. *There exists a unique solution $w \in H^1_\xi(L^2_v)$ of (3.23). Moreover w satisfies*

$$\|\partial^k_\xi w\|_{\xi,v} \le \frac{1}{\kappa}\|\partial^k_\xi h_w\|_{\xi,v} \quad \textit{for } k = 0,1\,,$$

with κ as in Lemma 3.3.

Proof. We introduce an approximation by discrete velocity models. We choose an increasing sequence $\{\Omega^N\}$ of bounded measurable subsets of the support of the velocity measure exhausting it:

$$\Omega^N \subset \Omega^{N+1}\,, \qquad \bigcup_{N=1}^{\infty} \Omega^N = \Omega\,.$$

Each of the Ω^N is written as a finite disjoint union $\Omega^N = \bigcup_{j=1}^N \Omega^N_j$ of connected measurable subsets Ω^N_j, and the discrete velocities are chosen from these subsets: $v^N_j \in \Omega^N_j$ such that

$$\frac{1}{F(v^N_j)} = \frac{1}{\mu(\Omega^N_j)}\int_{\Omega^N_j} \frac{d\mu(v)}{F(v)}\,,$$

which is possible by the continuity of F. A quadrature formula for v-integrals is then defined by

$$\int f(v)\,d\mu(v) \approx \sum_{j=1}^N f(v^N_j)\mu(\Omega^N_j)\,.$$

These choices imply that for functions f and g, whose support is a subset of Ω^N and which are piecewise constant, i.e., constant on each Ω^N_j, the quadrature formula is exact both for the scalar product $\langle f,g\rangle_v$ and for the integrals $\int f\,d\mu(v)$ and $\int g\,d\mu(v)$. Finally, we make the decomposition of Ω^N fine enough such that

$$\lim_{N\to\infty}\sup_{1\le j\le N}\left|(v^N_j - s)^2 - \frac{1}{\mu(\Omega^N_j)}\int_{\Omega^N_j}(v-s)^2 d\mu(v)\right| = 0\,. \tag{3.26}$$

Now we approximate (3.23) by the discrete velocity model

(3.27)
$$\varepsilon(v_j^N - s)\partial_\xi w_j^N - (\mathcal{M}^N w^N)_j = h_{wj}^N := \frac{F(v_j^N)}{\mu(\Omega_j^N)} \int_{\Omega_j^N} \frac{h_w(v)}{F(v)} d\mu(v)\,, \quad j = 1, \dots, N\,,$$

where w^N denotes both the vector $(w_1^N, \dots, w_N^N)$ and the piecewise constant function in L_v^2 defined by

$$w^N(v) = \begin{cases} w_j^N & \text{for } v \in \Omega_j^N\,, \\ 0 & \text{for } v \in \Omega \setminus \Omega^N\,. \end{cases}$$

Note that by the construction of the quadrature we have

$$\langle f^N, g^N \rangle_N := \sum_{j=1}^N \frac{f_j^N g_j^N}{F(v_j^N)} \mu(\Omega_j^N) = \langle f^N, g^N \rangle_v\,.$$

The matrix $\mathcal{M}^N$ is defined by

(3.28)
$$(\mathcal{M}^N w^N)_j := F(v_j^N) \sum_{l=1}^N w_l^N \mu(\Omega_l^N) - w_j^N - (v_j^N - s)^2 F(v_j^N) \sum_{l=1}^N (v_l^N - s)^2 w_l^N \mu(\Omega_l^N)\,.$$

Equations (3.27) are a system of linear constant coefficient ODEs. A proof analogous to that of Lemma 3.3 shows that $\mathcal{M}^N$ is symmetric and negative definite with respect to $\langle \cdot, \cdot \rangle_N$ and, as a consequence, the generalized eigenvalue problem

$$(\mathcal{M}^N - \lambda(\mathcal{V}^N - s\mathrm{Id}^N))\phi = 0\,, \qquad \text{with } \mathcal{V}^N = \mathrm{diag}(v_1^N, \dots, v_N^N)\,,$$

corresponding to the left-hand side of (3.27) has only real eigenvalues away from zero. Thus, a unique bounded solution exists which converges to zero as $\xi \to \pm\infty$. Computing the scalar product of the resulting equation with w^N and integration with respect to ξ gives

$$-\int_{-\infty}^{\infty} \langle \mathcal{M}^N w^N, w^N \rangle_N \, d\xi = \int_{-\infty}^{\infty} \langle h_w^N, w^N \rangle_N \, d\xi\,.$$

With the definiteness of $\mathcal{M}^N$ and the properties of the quadrature, this implies

$$\|w^N\|_{\xi,v} \le \frac{1}{\kappa} \|h_w\|_{\xi,v}\,.$$

The uniform boundedness of w^N in $L_{\xi,v}^2$ implies its weak convergence (for a subsequence) to $w \in L_{\xi,v}^2$. We can also pass to the limit in (3.27). Here we use (3.26) for proving convergence of the last term in (3.28). The above estimate carries over to the limit w. Then the estimate for $\partial_\xi w$ is obtained by differentiating equation (3.23). Uniqueness is an obvious consequence. ∎

3.4. The nonlinear problem. In this section we prove existence and uniqueness of solutions of the nonlinear problem (3.16), (3.22) subject to $z(0) = z_0$ and to the requirement that the solution is square integrable with respect to ξ, replacing homogeneous far-field conditions.

After the preparations in the previous subsections, the proof is a straightforward contraction argument. We need, however, estimates for the right-hand sides of (3.16) and (3.22). In the following, C denotes (possibly different) ε-independent constants.

LEMMA 3.7.

(i) *The coefficients* Λ *and* Γ *satisfy*

$$\|\Lambda\|_{C^1_\xi(L^2_v)} + \|\Gamma\|_{C^1_\xi(L^2_v)} \le C\,.$$

(ii) *The nonlinear term* $R(\rho)$ *is quadratic in* ρ*: Let* $\rho_1, \rho_2 \in H^1_\xi$ *satisfy* $|\rho_1|, |\rho_2| \le C\varepsilon^{-2}$*. Then*

$$\|R(\rho_1) - R(\rho_2)\|_{H^1_\xi(L^2_v)} \le C\left(\|\rho_1\|_{H^1_\xi} + \|\rho_2\|_{H^1_\xi}\right)\|\rho_1 - \rho_2\|_{H^1_\xi}\,.$$

Proof. The proofs of the statements are straightforward. All that is needed is the boundedness in L^2_v of the derivatives $\partial^k_\rho M(\rho_{as} + \varepsilon^2\tilde\rho)$ for $k \le 3$ with $\tilde\rho$ between values of ρ_1 and ρ_2, as well as the continuous embedding $L^\infty_\xi \to H^1_\xi$. ∎

LEMMA 3.8. *The projection* $P : L^2_v \to L^2_v$*, defined by* (3.20)*, is a bounded operator.*

Proof. The proof is a straightforward consequence of (2.3)–(2.5). ∎

Before stating the existence and uniqueness result for traveling waves we recall the micro-macro decomposition $g(\xi, v) = z(\xi)\Phi(v) + \varepsilon w(\xi, v)$ of functions $g \in H^1_\xi(L^2_v)$, made unique by the requirement $\langle (v-s)\Phi, w\rangle_v = 0$, since

$$\langle (v-s)\Phi, \Phi\rangle_v = \frac{\varepsilon\sigma\tilde D}{\hat D} \ne 0\,.$$

We define a norm on $H^1_\xi(L^2_v)$ by

$$\|g\| := \|z\|_{H^1_\xi} + \varepsilon\|w\|_{H^1_\xi(L^2_v)}\,. \tag{3.29}$$

We also note that in terms of the original unknown $f = f_{as} + \varepsilon^2 g$, the condition $z(0) = z_0$ reads

$$\langle (v-s)\Phi, f - f_{as}\rangle_v(\xi = 0) = \frac{\varepsilon^3\sigma\tilde D}{\hat D} z_0\,. \tag{3.30}$$

THEOREM 3.9. *Let the assumptions stated in section* 2 *be satisfied. Then for every* $z_0 \in \mathbb{R}$ *and for* ε *small enough, there exists a solution* f *of* (1.12) *satisfying* (3.30)*, unique in a ball in* $(H^1_\xi(L^2_v), \|\ \|)$ *with center* f_{as} *and an* $O(\varepsilon)$ *radius. It satisfies*

$$\|f - M(\rho)\|_{H^1_\xi(L^2_v)} = O(\varepsilon^2)\,,$$

where ρ *is the solution of* (3.4)–(3.6)*, or, more precisely,*

$$f = M(\rho) - \varepsilon(v-s)\partial_\xi M(\rho) + \varepsilon^2 z\Phi + \varepsilon^3 w\,,$$

where $\|z\|_{H^1_\xi}$ *and* $\|w\|_{H^1_\xi(L^2_v)}$ *are uniformly bounded as* $\varepsilon \to 0$.

Proof. As a consequence of Lemmas 3.7 and 3.8 we have

$$\left\| \frac{a'(\rho_{as}) - s}{\varepsilon} \rho_w \right\|_{H^1_\xi} \le C \|w\|_{H^1_\xi(L^2_v)} ,$$
$$\|P\Lambda z\|_{H^1_\xi(L^2_v)} \le C\|z\|_{H^1_\xi} , \qquad \|\tilde{\Gamma}\rho_w\|_{H^1_\xi(L^2_v)} \le C\|w\|_{H^1_\xi(L^2_v)} .$$

This implies that for ε small enough, the results from Lemmas 3.5 and 3.6 can be extended to the linear system

$$\tilde{D}\,\partial_\xi z - r(\xi) z = \varepsilon \frac{a'(\rho_{as}) - s}{\varepsilon} \rho_w + h_z ,$$
$$\varepsilon(v - s)\partial_\xi w - \mathcal{M} w - zP\Lambda = \varepsilon \tilde{\Gamma} \rho_w + h_w ,$$

with inhomogeneities h_z, h_w and $z_0 = z(0)$. Applying the solution operator for this system to (3.16), (3.22), we obtain a fixed point problem of the form

$$z = \varepsilon R_z(z\rho_\Phi + \varepsilon\rho_w) + \tilde{h}_z , \tag{3.31}$$
$$w = \varepsilon R_w(z\rho_\Phi + \varepsilon\rho_w) + \tilde{h}_w , \tag{3.32}$$

with R_z and R_w sharing the properties of R given in Lemma 3.7, and $\tilde{h}_z$ and $\tilde{h}_w$ are given and bounded. Using $\|z\rho_\Phi + \varepsilon\rho_w\|_{H^1_\xi} \le \|(z, w)\|$ (see (3.29)), the estimate

$$\|(\varepsilon R_z(z\rho_\Phi + \varepsilon\rho_w) + \tilde{h}_z, \varepsilon R_w(z\rho_\Phi + \varepsilon\rho_w) + \tilde{h}_w)\| \le c(1 + \varepsilon\|(z, w)\|^2)$$

follows. Here we have identified $g = z\Phi + \varepsilon w$ with the pair (z, w). The estimate implies that for ε small enough both the ball with radius $2c$ and the ball with radius $1/(2\varepsilon c)$ are mapped into themselves by the right-hand side of (3.31), (3.32). Also, with the property of the nonlinearity from Lemma 3.7, the fixed point operator is a contraction on a ball with an $O(\varepsilon^{-1})$ radius. We conclude that for ε small enough, (3.31), (3.32) has a solution with $\|(z, w)\| \le 2c$ which is unique in a ball with an $O(\varepsilon^{-1})$ radius. Knowing this and returning to (3.32), the boundedness of $\|w\|_{H^1_\xi(L^2_v)}$ follows. ∎

By its construction the approximating density ρ is strictly monotone. It will be important for the stability proof below to extend this property to the exact density ρ_f.

LEMMA 3.10. *Let the assumptions of Theorem* 3.9 *hold and let f be the solution of* (1.12), (3.30). *Then the macroscopic density $\rho_f(\xi)$ is strictly monotone.*

Proof. The previous proof is easily extended to show that the dependence of z and w on z_0 is Lipschitz continuous with ε-independent Lipschitz constant. Actually, the difference of two solutions (z, w) and $(\hat{z}, \hat{w})$ with different z_0-values z_0 and $\hat{z}_0$, respectively, satisfies a system similar to (3.31), (3.31) with inhomogeneities proportional to $z_0 - \hat{z}_0$. With the properties of the nonlinearities from Lemma 3.7 it is straightforward that

$$\|z - \hat{z}\|_{H^1_\xi}, \ \|w - \hat{w}\|_{H^1_\xi(L^2_v)} \le C|z_0 - \hat{z}_0| . \tag{3.33}$$

For the corresponding solutions f and $\hat{f}$ of (1.12), (3.30),

$$\rho_f(0) - \rho_{\hat{f}}(0) = \varepsilon^2(z_0 - \hat{z}_0)\rho_\Phi + \varepsilon^3(\rho_w(0) - \rho_{\hat{w}}(0))$$

holds. The continuous embedding of $C(\mathbb{R})$ in H^1_ξ and (3.33) imply

$$|\rho_w(0) - \rho_{\hat{w}}(0)| \le C|z_0 - \hat{z}_0|\,,$$

and, thus, strict monotonicity (and therefore invertibility) of the map $z_0 \mapsto \rho_f(0)$ for ε small enough. This in turn implies that the traveling wave can also be made locally unique by prescribing the value of $\rho_f(0)$ instead of z_0. This argument can of course be repeated with $\rho_f(\xi_0)$ for every $\xi_0 \in \mathbb{R}$ instead of the origin.

Now assume that ρ_f is not strictly monotone. Then there are two ξ-values ξ_0 and $\xi_0 + \delta$ with arbitrarily small positive δ such that $\rho_f(\xi_0) = \rho_f(\xi_0 + \delta)$. Now $\tilde{f}(\xi, v) = f(\xi + \delta, v)$ is a traveling wave with $\rho_{\tilde{f}}(\xi_0) = \rho_f(\xi_0)$ and $\tilde{f}$ arbitrarily close to f by making δ small. By the uniqueness result $\tilde{f} \equiv f$, and, consequently, f is periodic, which is a contradiction to the far-field boundary conditions. ∎

4. Local stability of small amplitude traveling waves. In this section we prove dynamic stability of the small amplitude traveling waves constructed above. As mentioned in the introduction, the techniques we employ are commonly used for conservation laws regularized with diffusion terms. This motivates the following scaling: we write (1.1) in the traveling wave variable $\xi = (x - st)\varepsilon$, and introduce the *parabolic* scaling $t \to t/\varepsilon^2$, where ε is the amplitude of the wave. Then (1.1) reads

$$\varepsilon^2 \partial_t f + \varepsilon(v - s)\partial_\xi f = M(\rho_f) - f\,, \tag{4.1}$$

and we pose the same far-field boundary conditions as for the traveling wave:

$$f(t, \xi = \pm\infty, v) = M(\rho_\pm, v)\,.$$

Let us denote by ϕ a traveling wave solution as constructed in Theorem 3.9. By Lemma 3.10 its macroscopic profile is monotone, implying

$$\partial_\xi(a'(\rho_\phi)) \le 0\,. \tag{4.2}$$

Observe that formally the integral of $\rho_f - \rho_\phi$ is constant in t. This allows us to choose ϕ, by shifting in the ξ-direction if necessary, such that

$$\int_{\mathbb{R}} (\rho_f - \rho_\phi)\, d\xi = 0\,. \tag{4.3}$$

Condition (4.3) fixes the shift in ξ; we expect the solution f to approach this particular ϕ as $t \to \infty$.

Let us denote by G the deviation of f from ϕ, namely,

$$\varepsilon G = f - \phi\,. \tag{4.4}$$

Then G satisfies the equation

$$\varepsilon\partial_t G + (v - s)\partial_\xi G = \frac{1}{\varepsilon^2}\left[M(\rho_\phi + \varepsilon\rho_G) - M(\rho_\phi)\right] - \frac{1}{\varepsilon}G\,. \tag{4.5}$$

We recall that condition (4.3) allows us to deal with the *macroscopic* unknown $W = \int_{-\infty}^{\xi} \rho_G\, d\xi$; see (1.20).

We decompose G into a macroscopic part and into a microscopic part by simply using the natural *macroscopic* projection $f \to F\rho_f$; thus we write

$$G = \rho F + \varepsilon g, \quad \text{i.e.,} \quad \rho := \rho_G\,.$$

Analogously, we split (4.5) into its microscopic and macroscopic parts, i.e., we apply the macroscopic projection, and its complementary *microscopic* projection, which is in fact the operator $-\mathcal{L}$. Application of the macroscopic projection and division by F gives the equation

$$\partial_t \rho + \frac{1}{\varepsilon}(a'(\rho_-) - s)\partial_\xi \rho + \partial_\xi \int (v-s) g \, d\mu(v) = 0\,, \tag{4.6}$$

and application of $-\mathcal{L}$ gives

$$\varepsilon^2 \partial_t g + (v - a'(\rho_-))F\partial_\xi \rho - \varepsilon \partial_\xi \mathcal{L}((v-s)g) = R_2[\rho] - g\,, \tag{4.7}$$

with

$$R_2[\rho](t,\xi,v) = \frac{1}{\varepsilon^2}\left[M(\rho_\phi(\xi) + \varepsilon\rho(t,\xi), v) - M(\rho_\phi(\xi), v) - \varepsilon\rho(t,\xi)F(v)\right]\,. \tag{4.8}$$

Integrating (4.6) with respect to ξ gives, in terms of $W = \int_{-\infty}^{\xi} \rho(t,y)\,dy$,

$$\partial_t W - \sigma \partial_\xi W + \int (v-s) g \, d\mu(v) = 0\,. \tag{4.9}$$

Using (4.7) we compute

$$\begin{aligned}\int (v-s) g \, d\mu(v) = \int (v-s) R_2[\rho] \, d\mu(v) - \varepsilon^2 \int (v-s)\partial_t g \, d\mu(v) \\ + \varepsilon\, \partial_\xi \int (v-s)\mathcal{L}((v-s)g) d\mu(v) - D_0 \partial_\xi \rho\,.\end{aligned} \tag{4.10}$$

Substituting (4.10) into (4.9) and setting

$$r_2[\rho](t,\xi) := \int (v-s) R_2[\rho](t,\xi,v)\, d\mu(v)\,, \tag{4.11}$$

we arrive at the integrated *macroscopic* equation

$$\partial_t W - \sigma \partial_\xi W + r_2[\rho] - D_0 \partial_\xi^2 W = \varepsilon S[g]\,, \tag{4.12}$$

with

$$S[g] = \partial_\xi \int (v-s)\left(\varepsilon \partial_t g - \mathcal{L}((v-s)\partial_\xi g)\right)\, d\mu(v)\,. \tag{4.13}$$

Observe that (4.12) has the form of the perturbation equation for viscous shock profiles with a microscopic perturbation on the right-hand side and the nonlinearity

$$r_2[\rho] = \frac{1}{\varepsilon^2}(a(\rho_\phi + \varepsilon\rho) - a(\rho_\phi) - \varepsilon\rho a'(\rho_-))\,. \tag{4.14}$$

For controlling the nonlinear terms, a uniform (in ε) bound on the L^∞-norm of the density ρ is needed. For the macroscopic equation without the microscopic perturbation on the right-hand side, this is a consequence of the maximum principle. Here we shall employ bounds in H^1_ξ for the same purpose.

Assuming such a bound, we write $R_2[\rho]$ as

$$R_2[\rho] = \rho \int_0^1 \frac{M'(\rho_\phi + \varepsilon\rho\eta) - M'(\rho_-)}{\varepsilon}\, d\eta\,.$$

Then by differentiation with respect to ξ and by assumption (2.4),

$$\|R_2[\rho]\|_{H^k_\xi(L^2_v)} \le C\|\rho\|_{H^k_\xi} \quad \text{for } k=0,1,2\,, \tag{4.15}$$

and, consequently, by (2.5),

$$\|r_2[\rho]\|_{H^k_\xi} \le C\|\rho\|_{H^k_\xi} \quad \text{for } k=0,1,2\,, \tag{4.16}$$

where here and in the following the symbols C as well as C_j with various j denote constants depending on $\|\rho\|_{L^\infty_\xi}$ but independent of ε.

As the next step we derive integral estimates as one would do for the purely macroscopic case.

LEMMA 4.1. *Let W be a solution of* (4.12) *(for given g) and $\rho=\partial_\xi W$. Then the following estimates hold:*

$$\frac{1}{2}\frac{d}{dt}\|W\|^2_\xi + (D_0 - C_0\|W\|_{L^\infty_\xi})\|\partial_\xi W\|^2_\xi \le \varepsilon\int_{-\infty}^{\infty} W\,S[g]d\xi\,, \tag{4.17}$$

$$\frac{1}{2}\frac{d}{dt}\|\partial^k_\xi W\|^2_\xi + \frac{D_0}{2}\|\partial^{k+1}_\xi W\|^2_\xi - C_k\|\rho\|^2_{H^{k-1}_\xi} \le \varepsilon\int_{-\infty}^{\infty}\partial^k_\xi W\,S[\partial^k_\xi g]d\xi \tag{4.18}$$

for $k=1,2$. The constants C_0, C_1, C_2 depend on $\|\rho\|_{L^\infty_\xi}$.

Proof. For proving (4.17), we test (4.12) with W. Let us look at the term containing $r_2[\rho]$ by writing

$$r_2[\rho] = \frac{1}{\varepsilon}\left(a'(\rho_\phi) - a'(\rho_-)\right)\rho + \frac{1}{2}a''(\tilde\rho)\rho^2\,,$$

with $\tilde\rho$ between ρ_ϕ and $\rho_\phi+\varepsilon\rho$. Then we get

$$\begin{aligned}\int_{\mathbb{R}} r_2[\rho]\,W d\xi &= \frac{1}{2\varepsilon}\int_{\mathbb{R}}(a'(\rho_\phi)-a'(\rho_-))\partial_\xi(W^2)\,d\xi + \frac{1}{2}\int_{\mathbb{R}} a''(\tilde\rho)(\partial_\xi W)^2 W\,d\xi\\ &\ge -\frac{1}{2\varepsilon}\int_{\mathbb{R}}\partial_\xi(a'(\rho_\phi))W^2 d\xi - C_0\|W\|_{L^\infty_\xi}\|\partial_\xi W\|^2_\xi\,.\end{aligned} \tag{4.19}$$

The first term on the right-hand side of (4.19) is positive by (4.2), completing the proof of (4.17).

For $k=1,2$, the corresponding ξ-derivatives of (4.12) are tested with $\partial^k_\xi W$. As already mentioned in the introduction, no positive term can be expected to arise from the nonlinearity. Therefore we just estimate the corresponding terms using (4.16):

$$\left|\int_{\mathbb{R}}\partial^k_\xi\, r_2[\rho]\,\partial^k_\xi W d\xi\right| \le \int_{\mathbb{R}}|\partial^{k-1}_\xi r_2[\rho]\,\partial^{k+1}_\xi W|d\xi \le C_k\|\rho\|^2_{H^{k-1}_\xi} + \frac{D_0}{2}\|\partial^{k+1}_\xi W\|^2_\xi\,, \tag{4.20}$$

completing the proof of (4.18). ∎

Before deriving estimates for the microscopic contributions, we have to deal with the difficulty that the operator $S[g]$ describing the microscopic perturbation of the macroscopic equation contains the time derivative $\partial_t g$.

LEMMA 4.2. *Let W and g satisfy* (4.9)*, and let the operator S be defined by* (4.13)*. Then for $k=0,1,2$ the following holds:*

$$\begin{aligned}&\int_{-\infty}^{\infty}\partial^k_\xi W\,S[\partial^k_\xi g]d\xi\\ &\le \varepsilon\frac{d}{dt}\int_{-\infty}^{\infty}\langle F\partial^k_\xi W, (v-s)\partial^k_\xi g\rangle_{\xi,v} + C(\|\partial^{k+1}_\xi W\|^2_\xi + \|\partial^k_\xi g\|^2_{\xi,v})\,.\end{aligned} \tag{4.21}$$

Proof. A straightforward computation, using the kth order derivative of (4.9), gives

$$\int_{-\infty}^{\infty} \partial_\xi^k W\, S[\partial_\xi^k g] d\xi = \varepsilon \frac{d}{dt} \int_{-\infty}^{\infty} \partial_\xi^k W \int (v-s) \partial_\xi^k g d\mu(v)\, d\xi$$
$$- \varepsilon\sigma \int_{-\infty}^{\infty} \partial_\xi^{k+1} W \int (v-s)\partial_\xi^k g\, d\mu(v)\, d\xi + \int_{-\infty}^{\infty} \left(\int (v-s)\partial_\xi^k g\, d\mu(v) \right)^2 d\xi$$
$$+ \int_{-\infty}^{\infty} \partial_\xi^{k+1} W \int (v-s)\mathcal{L}((v-s)\partial_\xi^k g) d\mu(v)\, d\xi\,.$$

To estimate the last three terms, we use the Cauchy–Schwarz inequality and (2.5). ∎

To get control of the microscopic part, we derive entropy estimates from the full kinetic perturbation equation (4.5).

LEMMA 4.3. *Let $G = \rho F + \varepsilon g$ be a solution of* (4.5). *Then, for $k = 0, 1, 2$,*

$$\frac{d}{dt} \left(\|\partial_\xi^k \rho\|_\xi^2 + \varepsilon^2 \|\partial_\xi^k g\|_{\xi,v}^2 \right) + \|\partial_\xi^k g\|_{\xi,v}^2 \le C \|\rho\|_{H_\xi^k}^2\,. \tag{4.22}$$

Proof. Writing the right-hand side of (4.5) as $R_2[\rho] - g$ and taking the scalar product of its kth derivative with $\partial_\xi^k G = \partial_\xi^k \rho F + \partial_\xi^k g$, we get

$$\frac{1}{2}\frac{d}{dt} \|\partial_\xi^k G\|_{\xi,v}^2 + \|\partial_\xi^k g\|_{\xi,v}^2 = \langle \partial_\xi^k R_2[\rho], \partial_\xi^k g \rangle_{\xi,v}\,.$$

The result is a consequence of using $\|\partial_\xi^k G\|_{\xi,v}^2 = \|\partial_\xi^k \rho\|_\xi^2 + \varepsilon^2 \|\partial_\xi^k g\|_{\xi,v}^2$, and then applying the Cauchy–Schwarz inequality, the Young inequality, and (4.15) to the right-hand side. ∎

Now we are prepared for proving our stability result for small kinetic shock profiles.

THEOREM 4.4. *Let the assumptions of Theorem* 3.9 *hold and let ϕ be a traveling wave solution. Let $f_0(\xi, v)$ be an initial datum for* (4.1) *and let $W_0(\xi) = \frac{1}{\varepsilon}\int_{-\infty}^{\xi} (\rho_{f_0}(\eta) - \rho_\phi(\eta)) d\eta$. Let $f_0 - \phi \in H_\xi^2(L_v^2)$ (implying that f_0 satisfies the same far-field conditions as ϕ) and $W_0 \in L_\xi^2$ (implying $W_0(\infty) = \int_{-\infty}^{\infty} (\rho_{f_0}(\xi) - \rho_\phi(\xi)) d\xi = 0$). Let*

$$\|W_0\|_{L_\xi^2} + \frac{1}{\varepsilon} \|f_0 - \phi\|_{H_\xi^2(L_v^2)} \le \delta \tag{4.23}$$

for δ small enough, but independent from ε. Then (4.1) *subject to the initial condition $f(t=0) = f_0$ has a unique global solution and*

$$\lim_{t\to\infty} \int_t^\infty \|f(s,\cdot,\cdot) - \phi\|_{H_\xi^2(L_v^2)}^2\, ds = 0\,.$$

REMARK 4.5. *The shortcomings of Theorem* 4.4 *are that it gives only local stability (i.e., the perturbations have to be small enough) of small amplitude traveling waves. Also convergence as $t \to \infty$ holds only in a very weak sense. The smallness of the wave is the basic assumption of this work allowing a perturbative treatment close to macroscopic equations. The other shortcomings are already present in the underlying results for viscous shock profiles.*

Proof. The proof is based on the construction of a Lyapunov functional H, such that both H and $-dH/dt$ measure the size of the perturbation. However, it will be impossible to estimate H in terms of $-dH/dt$, which is the reason why we do not have a result on the time decay rate.

We start by defining a partial functional for each differentiation order k appearing in Lemmas 4.1 and 4.3. By adding the corresponding inequality from Lemma 4.1 and the product of εA_k with the corresponding inequality from Lemma 4.3, we produce an inequality for the time derivative of

$$\begin{aligned} H_k &= \frac{1}{2}\|\partial_\xi^k W\|_{L^2_\xi}^2 - \varepsilon^2\langle F\partial_\xi^k W, (v-s)\partial_\xi^k g\rangle_{\xi,v} + \varepsilon^3 A_k \|\partial_\xi^k g\|^2_{L^2_\xi(L^2_v)} + \varepsilon A_k \|\partial_\xi^{k+1} W\|^2_{L^2_\xi} \\ &\geq \kappa_k(\|\partial_\xi^k W\|^2_{L^2_\xi} + \varepsilon\|\partial_\xi^{k+1} W\|^2_{L^2_\xi} + \varepsilon^3\|\partial_\xi^k g\|^2_{L^2_\xi(L^2_v)})\,, \end{aligned}$$

where the last inequality holds with an ε-independent $\kappa_k > 0$ if $A_k > 0$ is chosen independently of ε (but otherwise arbitrarily). With two more positive constants γ_1 and γ_2, we define the Lyapunov functional as $H = H_0 + \gamma_1 H_1 + \gamma_2 H_2$ and observe that it can be bounded from above and below by

$$\|W\|^2_{H^2_\xi} + \varepsilon\|\partial_\xi^3 W\|^2_\xi + \varepsilon^3\|g\|^2_{H^2_\xi(L^2_v)}\,.$$

With the aid of the Lemmas 4.1, 4.2, and 4.3, it is now straightforward to obtain an inequality of the form

$$\frac{dH}{dt} \leq -C(\|\rho\|_{L^\infty_\xi})(1 - \|W\|_{L^\infty_\xi})(\|\rho\|^2_{H^2_\xi} + \varepsilon\|g\|^2_{H^2_\xi(L^2_v)})\,, \tag{4.24}$$

with an ε-independent positive $C(\|\rho\|_{L^\infty_\xi})$. Since (by one-dimensional Sobolev embedding) H controls the L^∞_ξ-norms of W and ρ, the right-hand side is negative at $t = 0$ and remains so if $H(0)$ is small enough. This in turn is guaranteed by assumption (4.23). Note that the L^2_ξ-norm of W (and, thus, H) is not controlled by the dissipation term in (4.24), which is the reason that we cannot obtain a decay rate. The proof is completed by integrating (4.24) with respect to t. ∎

REFERENCES

[1] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation laws*, J. Statist. Phys., 95 (1999), pp. 113–170.

[2] R. E. Caflisch and B. Nicolaenko, *Shock profile solutions of the Boltzmann equation*, Comm. Math. Phys., 86 (1982), pp. 161–194.

[3] C. M. Cuesta and C. Schmeiser, *Kinetic profiles for shock waves of scalar conservation laws*, Transport Theory Statist. Phys., to appear.

[4] F. Golse, *Shock profiles for the Perthame-Tadmor kinetic model*, Comm. Partial Differential Equations, 23 (1998), pp. 1857–1874.

[5] M. A. Katsoulakis and A. E. Tzavaras, *Contractive relaxation systems and the scalar multidimensional conservation law*, Comm. Partial Differential Equations, 22 (1997), pp. 195–233.

[6] S. Kawashima and A. Matsumura, *Asymptotic stability of traveling wave solutions of systems for one-dimensional gas motion*, Comm. Math. Phys., 101 (1985), pp. 97–127.

[7] T.-P. Liu and S.-H. Yu, *Boltzmann equation: Micro-macro decompositions and positivity of shock profiles*, Comm. Math. Phys., 246 (2004), pp. 133–179.

[8] R. L. Pego, *Remarks on the stability of shock profiles for conservation laws with dissipation*, Trans. Amer. Math. Soc., 291 (1985), pp. 353–361.

[9] B. Perthame and E. Tadmor, *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys., 136 (1991), pp. 501–517.

[10] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 657–680

RIGOROUS DERIVATION OF FÖPPL'S THEORY FOR CLAMPED ELASTIC MEMBRANES LEADS TO RELAXATION*

SERGIO CONTI†, FRANCESCO MAGGI†, AND STEFAN MÜLLER‡

Abstract. We consider the nonlinear elastic energy of a thin membrane whose boundary is kept fixed, and assume that the energy per unit volume scales as h^β, with h the film thickness and $\beta \in (0, 4)$. We derive, by means of Γ convergence, a limiting theory for the scaled displacements, which takes a form similar to the one proposed by Föppl in 1907. Our variational approach fully incorporates the possibility of buckling already observed during the derivation of the reduced two-dimensional theory. At variance with Föppl's, our limiting model is lower semicontinuous and has an energetics that vanishes on all contractions. Therefore buckling does not need to be explicitly resolved when computing with the reduced theory. If forces normal to the membrane are included, then our result predicts that the normal displacement scales as the cube root of the force. This scaling depends crucially on the clamped boundary conditions. Indeed, if the boundary is left free, then a much softer response is obtained, as was recently shown by Friesecke, James, and Müller [*Arch. Ration. Mech. Anal.*, 180 (2006), pp. 183–236].

Key words. Gamma convergence, thin-film elasticity, relaxation

AMS subject classifications. 74K15, 35J35, 49J45, 35B40, 74G65

DOI. 10.1137/050632567

1. Introduction.

Reduced theories for thin elastic bodies have been proposed and used since the early days of the theory of elasticity, but only in the last decade has it become possible to derive them rigorously from three-dimensional nonlinear elasticity. The convergence criterion which has been used for these problems is Γ-convergence (which is very closely related to convergence of minimizers), and the different physical regimes are reflected by different scalings of the applied forces, and hence of the elastic energy, and different topologies on the space of deformations [17, 18, 13, 12, 22, 23, 10] (we refer to [10] for a review of the recent mathematical literature and of the mechanical context).

One key property of the elasticity of thin bodies is that tangential displacements enter the strain to first order, but normal displacements only to second order (see Figure 1.1). Therefore linear theories are not usable, as they would describe all normal displacements as completely stress-free (soft). The first nonvanishing contribution of normal displacements to strain is quadratic, and correspondingly the leading energy contribution is of fourth order.

A generalization of the linear theory which incorporates the normal displacements to leading order was proposed by Föppl [9]. In a variational language, and for the special case of isotropic elastic moduli and zero Poisson's ratio, his model corresponds

*Received by the editors May 27, 2005; accepted for publication (in revised form) February 13, 2006; published electronically July 31, 2006. This work was partially supported by Deutsche Forschungsgemeinschaft through the Schwerpunktprogramm 1095 Analysis, Modeling and Simulation of Multiscale Problems.

http://www.siam.org/journals/sima/38-2/63256.html

†Fachbereich Mathematik, Universität Duisburg-Essen, Lotharstr. 65, 47057 Duisburg, Germany (sergio.conti@uni-due.de, maggi@math.unifi.it).

‡Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, 04103 Leipzig, Germany (sm@mis.mpg.de). This author was partially supported by the Marie Curie Research Training Network MULTIMAT, MRTN-CT-2004-505226.

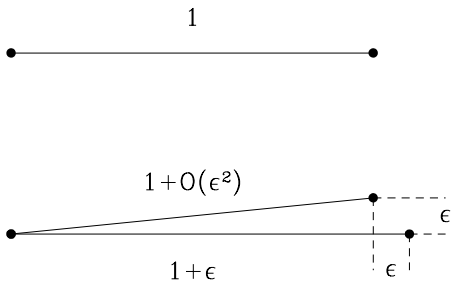


FIG. 1.1. *Consider a rod of unit length. If one endpoint is displaced tangentially by ϵ, the length also changes by ϵ. If instead the endpoint is displaced by ϵ in the normal direction, then the length changes only to order ϵ^2.*

to minimizing

$$\frac{1}{2}\int_S \left|\nabla u + \nabla u^T + \nabla v \otimes \nabla v\right|^2 + fv\, dx', \tag{1.1}$$

subject to appropriate boundary conditions. Here $S \subset \mathbb{R}^2$ represents the cross-section of the membrane, $u : S \to \mathbb{R}^2$ the tangential displacement, $v : S \to \mathbb{R}$ the normal displacement, and $f : S \to \mathbb{R}$ the applied normal force.

The functional (1.1) is not lower semicontinuous with respect to the weak $W^{1,2}(S,\mathbb{R}^2) \times W^{1,4}(S,\mathbb{R})$ topology. Physically, a sheet subject to moderate compression can relax its strain by forming fine-scale folds, which are not penalized by the functional (1.1) since it does not contain any curvature term. (We note in passing that even if bending energy is included compression is often still relaxed by fine-scale oscillations; see, e.g., [3, 7].)

It is therefore to be expected that a variational derivation will not lead to the functional (1.1), but to its relaxation. Indeed, we show here that under suitable scaling assumptions and with clamped boundary conditions three-dimensional elasticity reduces, in the sense of Γ-convergence, to a functional corresponding to the relaxation of (1.1), which, for the same special case, takes the form

$$\frac{1}{2}\int_S W_{\rm rel}\left(\nabla u + (\nabla u)^T + \nabla v \otimes \nabla v\right) dx' + fv\, dx', \tag{1.2}$$

where $W_{\rm rel}(F) = (\lambda_1^+(F))^2 + (\lambda_2^+(F))^2$, $\lambda_1(F)$ and $\lambda_2(F)$ are the eigenvalues of the symmetric matrix F, and $\lambda^+ = \max\{\lambda, 0\}$.

Our result, as will be explained in greater detail in the next section, has important consequences for the scaling behavior of the response of clamped membranes. Consider indeed application of a force $f_h(x') = h^\alpha f(x')$ normal to the membrane. If $\alpha \in (0,3)$, then our convergence result applied for $\beta = 4\alpha/3$ implies that the three-dimensional variational problems converge as $h \to 0$ to the relaxed problem $I_0(u,v) + \int_S fv dx'$, for I_0 as in (1.2). The tangential displacements scale as $h^{\beta/2} = h^{2\alpha/3}$, and the normal one as $h^{\beta/4} = h^{\alpha/3}$.

For $\alpha > 3$ one obtains a different limiting theory, which is quadratic and involves only bending energy (see, e.g., [10]). The limit functional takes the form $\int |\nabla^2 v|^2 + fv$. In this regime the out-of-plane displacement is linear in the applied force and thus scales like h^α.

Understanding the crossover from the linear to the sublinear scaling, which had also been observed experimentally, was an important motivation for the work of Föppl and von Kármán [27]. Indeed von Kármán points out that his theory interpolates between the linear (pure bending) theory and Föppl's theory [27, p. 350].[1] This crossover was analyzed quantitatively by Neubert and Sommer [20] and Way [28], finding good quantitative agreement with von Kármán's theory. Extensive experimental tests in the nonlinear regime appropriate to Föppl's theory were performed by Head and Sechler [14], finding very good agreement between Föppl's theory and the measured deflections with applied pressures varying over four orders of magnitude. While the von Kármán equations have received a lot of attention and there are several derivations available (with asymptotic expansions [5], with the implicit function theorem [19], and with Γ-convergence [10, 11]), we are not aware of any derivation of Föppl's geometrically linear membrane theory.

The scaling of the force appropriate for the present result is different than the one used in the derivation of the theory which includes bending [10]. Roughly speaking, one can say that the presence of the boundary conditions makes the membrane stiffer. In parallel, this additional stiffness of the membrane term makes the bending term irrelevant in the present scaling. The crucial difference between the present result and the hierarchy of models derived by Γ-convergence in [10] resides in the boundary conditions. In [10] natural boundary conditions are used, and as a consequence a variety of deformations $y : S \to \mathbb{R}^3$ are admissible which have zero membrane energy (isometric immersions, or developable maps). This leads to a rather soft response, dominated by the bending term. Here the boundary conditions rule out all nontrivial deformations with zero membrane energy, and as a consequence the response is much stiffer. The crucial dependence of the behavior on boundary conditions and in particular on the availability of (infinitesimally) isometric deformations was emphasized by Sanchez-Palencia [24]. In his terminology our situation corresponds to the case of ill-inhibited flexion, where the approach based on the spectral method is bound to fail. Our variational approach identifies the appropriate limiting theory, and shows that in this nonlinear regime the deformation scales as the cube root of the applied force.

Notation. The vectors e_1, e_2, and e_3 form an orthonormal basis of $\mathbb{R}^3$, and $\mathbb{R}^2$ is the space generated by e_1 and e_2. To every element $x = x_1e_1 + x_2e_2 + x_3e_3 \in \mathbb{R}^3$ we associate $x' := x_1e_1 + x_2e_2 \in \mathbb{R}^2$. Thus $x = x' + x_3e_3$.

The space of tensors generated by $\{e_i \otimes e_j\}_{i,j=1,2,3}$ is denoted by $\mathbb{R}^{3\times3}$, and $\mathbb{R}^{2\times2}$ is the subspace of $\mathbb{R}^{3\times3}$ generated by the tensors $\{e_i \otimes e_j\}_{i,j=1,2}$. To every $F = \sum_{i,j=1,2,3} F_{ij}\, e_i \otimes e_j \in \mathbb{R}^{3\times3}$ we associate $F' := \sum_{i,j=1,2} F_{ij}\, e_i \otimes e_j \in \mathbb{R}^{2\times2}$. By $\mathbb{R}^{n\times n}_{\mathrm{sym}}$ we denote the space of symmetric matrices, and by $\mathbb{R}^{n\times n}_{+}$ the subsets of positive semidefinite symmetric ones (i.e., $\{F \in \mathbb{R}^{n\times n}_{\mathrm{sym}} : F \geq 0\}$). Finally Id_n is the identity matrix in $\mathbb{R}^{n\times n}$. For scalar functions, we write for brevity $W^{1,2}_0(S) = W^{1,2}_0(S, \mathbb{R})$.

2. The relaxed Föppl functional. We consider the nonlinear elastic energy of a thin three-dimensional body $\Omega_h := S \times (-h/2, h/2)$, where $S \subset \mathbb{R}^2$ is the cross section and $h > 0$ the (small) thickness. The deformation is a map $w_h \in W^{1,2}(\Omega_h, \mathbb{R}^3)$,

[1] "In dieser Hinsicht liegt die wirkliche Platte zwischen den beiden Grenzfällen der *vollkommen steifen Platte* nach Gl. (27) und der *vollkommen biegsamen Platte*, deren Gleichungen sich aus dem System (29) mit $D = 0$ ergeben." Translation: In this regard the real plate lies in between the two limiting cases of the *completely stiff plate* according to Eq. (27) and the *completely flexible plate*, whose equations are obtained from the system (29) [i.e., the vK equations] with $D = 0$.

and its elastic energy per unit thickness is

$$E(w_h, \Omega_h) := \frac{1}{h} \int_{\Omega_h} W(\nabla w_h(x)) dx.$$

By general arguments it suffices to prove Γ-convergence without external forces [8]. The stored energy function W is assumed to satisfy the following:

(W1) $W : \mathbb{R}^{3\times 3} \to [0, \infty]$ is a Borel measurable function of class C^2 in an open neighborhood of $SO(3)$.

(W2) $W(RF) = W(F)$ for every $R \in SO(3)$ and every $F \in \mathbb{R}^{3\times 3}$; furthermore, $W(\mathrm{Id}_3) = 0$.

(W3) $W(F) \geq C \operatorname{dist}^2(F, SO(3))$ for every $F \in \mathbb{R}^{3\times 3}$.

We study the asymptotic behavior as $h \to 0$ of the minimization problems

$$\inf \left\{ \frac{E(w_h, \Omega_h)}{h^\beta} : w_h \in W^{1,2}(\Omega_h, \mathbb{R}^3),\ w_h(x) = x \text{ on } \partial S \times (-h/2, h/2) \right\}$$

in the range $\beta \in (0, 4)$, by means of Γ-convergence theory.

In order to define an appropriate convergence criterion for a sequence of deformations w_h, which are all defined on different domains, we rescale (following standard practice) to a unique domain. Precisely, for each $w_h \in W^{1,2}(\Omega_h, \mathbb{R}^3)$ we define $y_h \in W^{1,2}(\Omega_1, \mathbb{R}^3)$ by $y_h(x) = w_h(x' + hx_3 e_3)$. Then

$$E_h(w_h, \Omega_h) = \int_{\Omega_1} W(\nabla_h y_h(x)) dx,$$

where ∇_h is the operator $\nabla_h := \nabla' + (1/h)\partial_3 \otimes e_3$, i.e.,

$$\nabla_h y(x) = \partial_1 y(x) \otimes e_1 + \partial_2 y(x) \otimes e_2 + \frac{1}{h} \partial_3 y(x) \otimes e_3 \,.$$

In terms of the rescaled deformations, and including the constraint given by the boundary conditions, our problem corresponds to minimizing the functional $I_h : W^{1,2}(\Omega_1, \mathbb{R}^3) \to [0, \infty]$ given by

$$I_h(y) := \begin{cases} \displaystyle\int_{\Omega_1} W(\nabla_h y(x)) dx & \text{if } y(x) = x' + hx_3 e_3 \text{ for } x \in \partial S \times (-\tfrac{1}{2}, \tfrac{1}{2}), \\ +\infty & \text{else.} \end{cases}$$

Due to the boundary conditions and to the energy regime under consideration, the behavior of a low energy sequence y_h will be understood by considering the scaled displacements

$$u_h(x') := \frac{1}{h^{\beta/2}} \int_0^1 (y_h(x) - x)' \, dx_3, \tag{2.1}$$

$$v_h(x') := \frac{1}{h^{\beta/4}} \int_0^1 (y_h(x) - hx) \cdot e_3 \, dx_3. \tag{2.2}$$

Note that for every h we have $u_h \in W^{1,2}_0(S, \mathbb{R}^2)$ and $v_h \in W^{1,2}_0(S)$. However, for a sequence y_h such that $h^{-\beta} I_h(y_h)$ stays bounded, we shall prove that, up to extracting subsequences, (u_h, v_h) is weakly$*$ convergent only in the larger space $BD(S) \times W^{1,2}_0(S)$ (compare with part I of Theorem 2.1 below). We recall that $BD(S)$ denotes the space

of the deformations $u \in L^1(S, \mathbb{R}^2)$ such that the symmetric part of the *distributional gradient* $D'u$ is a Radon measure on S, namely

$$\operatorname{sym} D'u \in \mathcal{M}(S, \mathbb{R}^{2\times 2}_{\operatorname{sym}})$$

(the symbol $\mathcal{M}$ is used for spaces of Radon measures). The limit of the in-plane displacements u_h will take values in the smaller space

(2.3)
$$X(S) := \{u \in BD(S) : \exists M \in \mathcal{M}(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+) \text{ s.t. } \operatorname{sym} D'\overline{u} + M \in L^1(\mathbb{R}^2, \mathbb{R}^{2\times 2})\},$$

where $\overline{u} := u$ in S and $\overline{u} := 0$ in $\mathbb{R}^2 \setminus S$. This corresponds to requiring that the symmetrized distributional derivative is the sum of an L^1 term and a negative definite measure, singular with respect to Lebesgue measure. This sign condition does not bring any additional regularity, as $X(S)$ still contains elements that are not in $BV(S, \mathbb{R}^2)$. The formulation of (2.3) in terms of the extension $\bar{u}$ corresponds to a sign condition on the boundary values of u (in the sense of inner traces). Precisely, functions $u \in X(S)$ obey $\operatorname{tr}(u) = \lambda \nu_S$, where $\lambda \geq 0$ and ν_S is the outer normal. The structure of $X(S)$ is discussed in more detail in the appendix.

The main result of this paper is that for all $\beta \in (0, 4)$, as $h \to 0$ the functionals $h^{-\beta} I_h$ converge (in the sense of Γ-convergence) to the limit functional $I_0 : X(S) \times W^{1,2}_0(S) \to [0, \infty]$, defined as

$$I_0(u, v) := \inf \left\{ \frac{1}{2} \int_S Q_2 \left((\operatorname{sym} D'u + M)(x') + \frac{\nabla' v(x') \otimes \nabla' v(x')}{2} \right) dx' : \right.$$
$$\left. M \in \mathcal{M}(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+),\ \operatorname{sym} D'\overline{u} + M \in L^1(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+) \right\}.$$

Here $Q_2 : \mathbb{R}^{2\times 2} \to [0, \infty)$ is the quadratic form

$$Q_2(A) := \min \left\{ Q_3(\operatorname{sym} A + \operatorname{sym}(a \otimes e_3)) : a \in \mathbb{R}^3 \right\},$$

and $Q_3 : \mathbb{R}^{3\times 3} \to [0, \infty)$ is the Hessian of the energy at the identity, i.e.,

$$Q_3(F) := \nabla^2 W(\mathrm{Id}_3)[F, F].$$

By (W3) the quadratic forms Q_2 and Q_3 are positive definite on symmetric matrices. If $u \in W^{1,1}(S, \mathbb{R}^2)$ and $I_0(u, v) < \infty$, as one can see, the above expression for I_0 reduces to

$$I_0(u, v) = \frac{1}{2} \int_S W_{\mathrm{F\ddot{o}}}(\nabla' u(x'), \nabla' v(x')) dx', \tag{2.4}$$

where $W_{\mathrm{F\ddot{o}}} : \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \to [0, \infty)$ is defined by

$$W_{\mathrm{F\ddot{o}}}(A, b) := \min \left\{ Q_2 \left(\operatorname{sym} A + \frac{b \otimes b}{2} + M \right) : M = M^T,\ M \geq 0 \right\}.$$

We notice that $W_{\mathrm{F\ddot{o}}}$ is a convex function; see Lemma A.2 in the appendix. In the special case mentioned in the introduction, which corresponds to $Q_3(F) = |F|^2$, we get $Q_2(A) = |A|^2$, and $W_{\mathrm{F\ddot{o}}}(A, b)$ coincides, up to a normalization factor, with $W_{\mathrm{rel}}(A + b \otimes b)$ as given after (1.2).

The minimization over positive-definite matrices entering the definition of $W_{\mathrm{F\ddot{o}}}$ corresponds to the relaxation of compression by means of oscillations, and implies that $W_{\mathrm{F\ddot{o}}}$ vanishes on all contractions. This minimization was not present in the original theory by Föppl (i.e., he used $\tilde{W}_{\mathrm{F\ddot{o}}} = Q_2(\operatorname{sym} A + b \otimes b/2)$). This difference is the geometrically linear analogue of the difference between the membrane theory rigorously derived by Le Dret and Raoult [17, 18] and those that had been heuristically proposed before.

We now give a precise statement of our convergence result.

THEOREM 2.1. *Let $S \in \mathbb{R}^2$ be a bounded, strictly star-shaped, Lipschitz domain, and let W satisfy* (W1), (W2), (W3). *Then for every $\beta \in (0,4)$ the functionals $h^{-\beta} I_h$ Γ-converge (as $h \to 0$) to the relaxed Föppl functional I_0. More precisely we have the following:*

I. Compactness. For every sequence $h \to 0$ and every y_h such that

$$\limsup_{h\to 0} h^{-\beta} I_h(y_h) < \infty,$$

the sequences (u_h, v_h) defined by (2.1)–(2.2) *have a subsequence such that*

$$\begin{aligned} u_h &\rightharpoonup u \text{ weakly in } L^2(S,\mathbb{R}^2),\\ \operatorname{sym}\nabla' u_h &\overset{*}{\rightharpoonup} \operatorname{sym} D' u \text{ weakly* in } \mathcal{M}(S,\mathbb{R}^{2\times 2}),\\ v_h &\rightharpoonup v \text{ weakly in } W^{1,2}_0(S,\mathbb{R}^2) \end{aligned}$$

for some $u \in X(S)$ and $v \in W^{1,2}_0(S)$.

II. Lower bound. Under the same assumptions, and along the same subsequence,

$$\liminf_{h\to 0} \frac{I_h(y_h)}{h^\beta} \geq I_0(u,v).$$

III. Upper bound. For every pair of functions $u \in X(S)$ and $v \in W^{1,2}_0(S)$ and every sequence $h \to 0$ there exists a sequence of functions $y_h \in C^\infty(\Omega_1,\mathbb{R}^3)$ with $y_h(x) = x' + h x_3 e_3$ for $x \in \partial S \times (-1/2, 1/2)$ and such that the pair $(u_h, v_h) \in C^\infty_0(S,\mathbb{R}^2) \times C^\infty_0(S)$ defined via (2.1)–(2.2) *converges to (u,v) as above, and*

$$\lim_{h\to 0} \frac{I_h(y_h)}{h^\beta} = I_0(u,v).$$

By strictly star-shaped we mean that there is a point $x \in S$ such that for each $y \in \partial S$ the open segment (x,y) is contained in S. Parts I and II of the theorem hold for generic bounded Lipschitz domains.

We recall that such a Γ-convergence result implies convergence of minimizers, in the sense that Theorem 2.1 implies that the set of minima of I_0 coincides with the set of accumulation points of asymptotically minimizing sequences for $h^{-\beta} I_h$. Explicitly, (u,v) is a minimizer of I_0 if and only if there is a sequence y_h, converging to (u,v) as above, such that $h^{-\beta}[I_h(y_h) - \inf I_h] \to 0$.

Further, the same holds if a continuous perturbation, such as external forces, is included. In the relevant case of normal forces, this means that the sequence of functionals

$$h^{-\beta}\left[I_h(y_h) + \int_{\Omega_1} f_h(x')\,(y_h(x) - hx)\cdot e_3 dx\right]$$

Γ-converges to

$$I_0(u,v) + \int_S f(x')v(x')dx',$$

provided that $h^{3\beta/4}f_h(x')$ converges to f in $L^2(S)$.

We remark that the range of scalings covered by the present result ($\beta \in (0,4)$) is much broader than the one covered by the corresponding Γ-convergence results obtained without clamped boundary conditions. Indeed, without boundary conditions, different Γ-limits for $h^{-\beta}I_h$ have been determined for $\beta \in (0,5/3)$, $\beta = 2$, $\beta \in (2,4)$ (no result is yet known for $\beta \in [5/3,2)$). The two extreme cases $\beta = 0$ and $\beta = 4$ are special both in the presence or in the absence of clamped boundary conditions. We refer to [10] for a more complete presentation of these different regimes.

3. Proof of Theorem 2.1. We prove the three parts in sequence. We start from the argument for the compactness part, which is more specific to this situation where the energy has very little coercivity and different growth conditions in different variables. The form (1.1) shows that in this scaling regime one cannot expect to have a local coercivity. Compactness is gained by means of the boundary conditions. Indeed, the boundary values imply that ∇u_h has a zero average; hence the integral of $|\nabla v_h|^2$ is controlled by the energy. This gives control of ∇v_h in L^2, but of $\operatorname{sym}\nabla u_h$ only in L^1.

The lower bound is obtained by a standard argument exploiting the form of W close to the minimum, again with some subtleties arising from the weakness of the topologies.

Finally, in the upper bound an explicit construction is needed, which characterizes the folds which are used to reduce the energy of compressive deformations. In a first step we reduce to smooth displacements (u,v) with compact support, using the star-shapedness of S and the convexity of $W_{\mathrm{F\ddot{o}}}$. Then we provide a construction which reverses the relaxation. This is based on the explicit definition of oscillatory sequences which reduce the energy of compressive deformations. From the viewpoint of nonlinear elasticity the typical construction can be seen as a laminate between isometric deformations, whose average is, in general, a short deformation, i.e., a deformation whose gradient lies in the convex hull of the set of isometries $O(2,3)$.

Proof.

Part I: *Compactness.* We have a family of deformations y_h such that

$$y_h(x) = x' + hx_3e_3 \quad \forall x \in \partial S \times (-1/2, 1/2); \tag{3.1}$$

$$\int_{\Omega_1} W(\nabla_h y_h(x))dx \le Ch^\beta. \tag{3.2}$$

We now introduce new functions which characterize the deviation of the elastic deformation y_h from the identity $x' + hx_3e_3$. Since we are dealing with thin sheets it is natural to separate the tangential and the normal displacement. Therefore we consider $U_h \in W^{1,2}(\Omega_1, \mathbb{R}^2)$ and $V_h \in W^{1,2}(\Omega_1)$ defined by

$$y_h(x) = x' + hx_3e_3 + U_h(x) + V_h(x)e_3.$$

Equivalently,

$$U_h(x) := (y_h(x) - x)', \quad V_h(x) := (y_h(x) - hx) \cdot e_3.$$

The gradients are related by

$$\nabla_h y_h(x) = \mathrm{Id}_3 + \nabla' U_h(x) + e_3 \otimes \nabla' V_h(x) + \frac{1}{h}(\partial_3 U_h(x) + \partial_3 V_h(x) e_3) \otimes e_3.$$

The tangential nonlinear strain takes the form

$$\big[(\nabla_h y_h)^T \nabla_h y_h - \mathrm{Id}_3\big]' = 2\mathrm{sym}\,\nabla' U_h + (\nabla' U_h)^T(\nabla' U_h) + \nabla' V_h \otimes \nabla' V_h \tag{3.3}$$

(recall that F' denotes projection of F onto $\mathbb{R}^{2\times 2}$, and that $(\mathrm{Id}_3 + F)^T(\mathrm{Id}_3 + F) = \mathrm{Id}_3 + 2\mathrm{sym}\,F + F^T F$).

Integrating (3.3) over $x' \in S$ the first term cancels, since $\int_S \nabla' U(x)dx' = 0$ by (3.1). Taking the trace and integrating over $x_3 \in (-1/2, 1/2)$ leads to

$$\int_{\Omega_1} |\nabla' U_h(x)|^2 + |\nabla' V_h(x)|^2 dx = \mathrm{Tr} \int_{\Omega_1} \big[(\nabla_h y_h)^T \nabla_h y_h - \mathrm{Id}_3\big]' dx \le C h^{\beta/2}. \tag{3.4}$$

In the last step we used $|F^T F - \mathrm{Id}| \le C\mathrm{dist}(F, SO(3)) + C\mathrm{dist}^2(F, SO(3))$, (W3) and (3.2). Plugging this information back into (3.3) gives an analogous bound for $\mathrm{sym}\,\nabla' U_h$ in $L^1(\Omega_1; \mathbb{R}^{2\times 2}_{\mathrm{sym}})$. Summarizing we have

$$\int_{\Omega_1} |\mathrm{sym}\,\nabla' U_h(x)| + |\nabla' U_h(x)|^2 + |\nabla' V_h(x)|^2 dx \le C h^{\beta/2}. \tag{3.5}$$

Therefore it is natural to rescale the tangential displacement U_h by $h^{\beta/2}$, and the normal one V_h by $h^{\beta/4}$.

Taking averages over x_3, we define the rescaled displacements $u_h \in W^{1,2}_0(S, \mathbb{R}^2)$ and $v_h \in W^{1,2}_0(S)$ by

$$u_h(x') := \frac{1}{h^{\beta/2}} \int_{-1/2}^{1/2} U_h(x', x_3) dx_3, \quad v_h(x') := \frac{1}{h^{\beta/4}} \int_{-1/2}^{1/2} V_h(x', x_3) dx_3.$$

This definition is equivalent to (2.1) and (2.2) above.

By (3.5) the sequence $\nabla' v_h$ is bounded in $L^2(S, \mathbb{R}^2)$; hence there is a subsequence such that

$$v_h \rightharpoonup v \text{ weakly in } W^{1,2}_0(S). \tag{3.6}$$

By (3.5) the sequence $\mathrm{sym}\,\nabla' u_h$ is bounded in $L^1(S, \mathbb{R}^{2\times 2}_{\mathrm{sym}})$, and since $u_h \in W^{1,2}_0$ we can apply the Poincaré–Korn inequality [25] (see also [15, 16] and [26, sect. II.1]) to find

$$\|u_h\|_{L^2(S,\mathbb{R}^2)} \le C \|\mathrm{sym}\,\nabla' u_h\|_{L^1(S, \mathbb{R}^{2\times 2}_{\mathrm{sym}})} \le C.$$

In particular there is a subsequence and $u \in L^2$ such that

$$u_h \rightharpoonup u \qquad \text{weakly in } L^2(S, \mathbb{R}^2)\,. \tag{3.7}$$

Further, $\nabla' u_h$ converges to $D'u$ in the sense of distributions, and by (3.5)

$$\mathrm{sym}\,\nabla' u_h(x')dx' \overset{*}{\rightharpoonup} \mathrm{sym}\,D'u \qquad \text{weakly* in } \mathcal{M}(S, \mathbb{R}^{2\times 2}_{\mathrm{sym}})\,. \tag{3.8}$$

This is the compactness entailed in the functionals under considerations. We now use this information to obtain a lower bound that in turn will also allow us to prove that $u \in X(S)$.

Part II: *Lower bound*. The first part of the argument is along the lines of [12], and in a sense it constitutes the "generic" lower bound argument used in the regime $I_h(y_h) \to 0$, i.e., for $\nabla_h y_h$ close to $SO(3)$. In this range it is natural to "normalize" the deformation gradients $\nabla_h y_h$ in order to use the structure of W near $SO(3)$. This amounts to considering a field of rotations $R_h : \Omega_1 \to SO(3)$ such that

$$|\nabla_h y_h(x) - R_h(x)| = \operatorname{dist}(\nabla_h y_h(x), SO(3)).$$

The function R_h can be chosen to be measurable (see Lemma A.6 in the appendix), and hence in $L^\infty(\Omega_1, \mathbb{R}^{3\times 3})$. We also note (see Lemma A.1 in the appendix) that

$$R_h(x)^T \nabla_h y_h(x) \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}.$$

Consider now

$$G_h := \frac{R_h^T \nabla_h y_h - \mathrm{Id}_3}{h^{\beta/2}}. \tag{3.9}$$

Since $|G_h| = \operatorname{dist}(\nabla_h y_h, SO(3))/h^{\beta/2}$, from (W3) and (3.2) we get that G_h is uniformly bounded in L^2, and taking a subsequence,

$$G_h \rightharpoonup G \quad \text{weakly in } L^2(\Omega_1, \mathbb{R}^{3\times 3}).$$

We now use Taylor's formula to obtain a lower bound in terms of the second derivatives of W at the identity. Precisely, by (W1) and (W2) there is $\rho : \mathbb{R}_+ \to \mathbb{R}$ such that $\lim_{t\to 0} \rho(t)/t^2 = 0$ and

$$\begin{aligned} W(\nabla_h y_h) &= W(\mathrm{Id}_3 + R_h^T \nabla_h y_h - \mathrm{Id}_3) \\ &\geq \frac{1}{2} Q_3(R_h^T \nabla_h y_h - \mathrm{Id}_3) - \rho\left(|R_h^T \nabla_h y_h - \mathrm{Id}_3|\right). \end{aligned}$$

It is convenient to consider separately the part of the domain where $\nabla_h y_h$ is close to a rotation, which is large, and the small exceptional set. To do this, let

$$\omega_h = \{x \in \Omega_1 : \operatorname{dist}(\nabla_h y_h(x), SO(3)) \leq h^{\beta/4}\}.$$

Let χ_h be the characteristic function of ω_h. By (W3) and (3.2) we get $|\omega_h| \to |\Omega_1|$. Restricting the integration to ω_h we get

$$\begin{aligned} \frac{I_h(y_h)}{h^\beta} \geq{}& \frac{1}{2} \int_{\Omega_1} \chi_h(x) Q_3\left(\frac{R_h(x)^T \nabla_h y_h(x) - \mathrm{Id}_3}{h^{\beta/2}}\right) dx \\ &- \frac{1}{h^\beta} \int_{\Omega_1} \chi_h(x) \rho\left(\operatorname{dist}(\nabla_h y_h(x), SO(3))\right) dx. \end{aligned} \tag{3.10}$$

The second term goes to zero as $h \to 0$, for it is equal to the integral of

$$\frac{\chi_h \ \rho\left(\operatorname{dist}(\nabla_h y_h, SO(3))\right)}{\operatorname{dist}^2(\nabla_h y_h, SO(3))} \cdot \frac{\operatorname{dist}^2(\nabla_h y_h, SO(3))}{h^\beta}.$$

By the definition of ω_h the first fraction converges uniformly to zero as $h \to 0$, at the same time the second one is uniformly bounded in L^1 by (3.2).

As $\chi_h(x) \in \{0,1\}$ we also have $\chi_h Q_3(G_h) = Q_3(\chi_h G_h)$, and since $\chi_h G_h \rightharpoonup G$ weakly in $L^2(\Omega_1, \mathbb{R}^{3\times 3})$ we easily conclude from (3.10) that

$$\liminf_{h\to 0} \frac{I_h(y_h)}{h^\beta} \geq \frac{1}{2}\int_{\Omega_1} Q_3(G(x))dx.$$

Note that G is symmetric, as was G_h.

In order to extract further information on G it is useful to express it as a limit of a sequence not involving R_h. Since $\nabla_h y_h = R_h(\mathrm{Id}_3 + h^{\beta/2}G_h)$ we get

$$(\nabla_h y_h)^T(\nabla_h y_h) = \mathrm{Id}_3 + 2h^{\beta/2}G_h + h^\beta G_h^T G_h$$

and thus

(3.11)
$$G_h - \frac{(\nabla_h y_h)^T(\nabla_h y_h) - \mathrm{Id}_3}{2h^{\beta/2}} = -\frac{h^{\beta/2}}{2}G_h^T G_h \to 0 \quad \text{strongly in } L^1(\Omega_1, \mathbb{R}^{3\times 3}).$$

In particular

$$\frac{(\nabla_h y_h)^T(\nabla_h y_h) - \mathrm{Id}_3}{2h^{\beta/2}} \rightharpoonup G \quad \text{weakly in } L^1(\Omega_1, \mathbb{R}^{3\times 3}). \tag{3.12}$$

As $G(x)$ is symmetric we have $Q_3(G(x)) \geq Q_2(G(x)')$. Furthermore, as Q_2 is convex, we can apply Jensen's inequality in the x_3 direction and find

$$\liminf_{h\to 0} \frac{I_h(y_h)}{h^\beta} \geq \frac{1}{2}\int_S Q_2(A(x'))dx',$$

where

$$A(x') = \int_{-1/2}^{1/2} G(x' + x_3 e_3)' dx_3 \quad \forall x' \in S\,.$$

It remains to relate A to u and v. To do this, we consider the integral over $x_3 \in (-1/2, 1/2)$ of the nonlinear strain,

$$A_h(x') := \int_{-1/2}^{1/2} \frac{\left[(\nabla_h y_h)^T(\nabla_h y_h) - \mathrm{Id}_3\right]'}{2h^{\beta/2}} dx_3.$$

By (3.12) we have

$$A_h \rightharpoonup A \text{ weakly in } L^1(S, \mathbb{R}^{2\times 2})\,. \tag{3.13}$$

At the same time, dividing (3.3) by $2h^{\beta/2}$ and integrating over x_3 gives

$$A_h(x') = \frac{1}{h^{\beta/2}}\int_{-1/2}^{1/2} \mathrm{sym}\,\nabla' U_h(x) + \frac{\nabla' V_h(x)\otimes\nabla' V_h(x)}{2} + \frac{\nabla' U_h(x)^T\nabla' U_h(x)}{2} dx_3.$$

The first term equals $\mathrm{sym}\,\nabla' u_h(x')$, and the other two can be bounded via Jensen's inequality, leading to

$$A_h(x') \geq \mathrm{sym}\,\nabla' u_h(x') + \frac{\nabla' v_h(x')\otimes\nabla' v_h(x')}{2} + h^{\beta/2}\nabla' u_h(x')^T\nabla' u_h(x').$$

As v_h is bounded in $W^{1,2}(S)$ we have that $\nabla v_h \otimes \nabla v_h$ converges weakly* to a measure $\mu \in \mathcal{M}(S, \mathbb{R}^{2\times 2})$, and by a standard lower semicontinuity argument $\mu \geq \nabla v \otimes \nabla v$. Using (3.8) and the fact that the third term on the right-hand side is positive semidefinite, we conclude that

$$A(x')\,dx' \geq \operatorname{sym} D'u + \frac{\nabla' v(x') \otimes \nabla' v(x')}{2}\,dx'. \tag{3.14}$$

The difference of the two sides of this inequality defines a Radon measure on S with values in $\mathbb{R}^{2\times 2}_+$ that we denote by M. In particular $\operatorname{sym} D'u + M$ is absolutely continuous with respect to the Lebesgue measure as

$$\operatorname{sym} D'u + M = \left\{ A(x') - \frac{\nabla' v(x') \otimes \nabla' v(x')}{2} \right\} dx'.$$

Finally,

$$\liminf_{h\to 0} \frac{I_h(y_h)}{h^\beta} \geq \frac{1}{2} \inf \left\{ \int_S Q_2 \left((\operatorname{sym} D'u + M)(x') + \frac{\nabla' v(x') \otimes \nabla' v(x')}{2} \right) dx' \right\},$$

where the infimum runs over all $M \in \mathcal{M}(S, \mathbb{R}^{2\times 2}_+)$ such that $\operatorname{sym} D'u + M \in L^1(S, \mathbb{R}^{2\times 2}_{\operatorname{sym}})$.

Finally, we repeat the argument for $\overline{y}_h(x) := y_h(x)$ if $x \in S \times (-h/2, h/2)$, $\overline{y}_h(x) := x' + hx_3 e_3$ if $x \in (\mathbb{R}^2 \setminus S) \times (-h/2, h/2)$. As $W(\mathrm{Id}_3) = 0$ and $Q_3(0) = Q_2(0) = 0$ the above argument can be repeated without any change, and we find that there exists a measure $M \in \mathcal{M}(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+)$ such that $\operatorname{sym} D'\overline{u} + M \in L^1(\mathbb{R}^2, \mathbb{R}^{2\times 2}_{\operatorname{sym}})$. Thus $u \in X(S)$.

Part III: *Upper bound.* We are given $u \in X(S)$ and $v \in W^{1,2}_0(S)$ with $I_0(u, v) < \infty$ (otherwise there is nothing to prove), and we have to construct a recovery sequence. We shall now use the star-shapedness of S to show that it suffices to consider u and v with compact support in S; then we use convexity of $W_{\mathrm{F\ddot{o}}}$ to show that it suffices to consider smooth u and v, and finally we provide an explicit construction.

After a translation we can assume that S is star-shaped with respect to the origin. Fix $\varepsilon > 0$ and consider the functions

$$u_\varepsilon(x') = \frac{1}{1+\varepsilon}\,\bar{u}((1+\varepsilon)x')\,, \qquad v_\varepsilon(x') = \frac{1}{1+\varepsilon}\,\bar{v}((1+\varepsilon)x')\,.$$

As above, we denote by a bar extension by zero outside S, so that, e.g., $\bar{u} = u$ on S and $\bar{u} = 0$ in $\mathbb{R}^2 \setminus S$. It is clear that u_ε and v_ε are supported on $S/(1+\varepsilon) \subset\subset S$. At the same time $u_\varepsilon \in X(S)$ (as $u \in X(S)$), $v_\varepsilon \in W^{1,2}_0(S)$, and, as $\varepsilon \to 0$,

$$(u_\varepsilon, v_\varepsilon) \rightharpoonup (u, v) \qquad \text{weakly}* \text{ in } X(S) \times W^{1,2}(S),$$

(i.e., in the convergence stated in Part I). Now we remark that

$$I_0(u_\varepsilon, v_\varepsilon) \leq (1+\varepsilon)^{-2} I_0(u, v). \tag{3.15}$$

This follows from a change of variables, once one has proven that $\nabla' v_\varepsilon(x') = \nabla' v((1+\varepsilon)x')$ and that for any $M \in \mathcal{M}(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+)$ such that $\operatorname{sym} D'\overline{u} + M \in L^1(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+)$ we can find $M_\varepsilon \in \mathcal{M}(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+)$ such that

$$\operatorname{sym} D'\overline{u_\varepsilon} + M_\varepsilon = (\operatorname{sym} D'\overline{u} + M)((1+\varepsilon)x')dx'.$$

We now show how to construct M_ε. Since

$$\operatorname{sym} D'\overline{u_\varepsilon} = \frac{1}{(1+\varepsilon)^2}\left[\frac{1}{1+\varepsilon}\mathrm{Id}_2 \# \operatorname{sym} D'\overline{u}\right]$$

(where # stands for the push-forward of measures, that is, $f\#\mu(E) := \mu(f^{-1}(E))$), it suffices to choose

$$M_\varepsilon := \frac{1}{(1+\varepsilon)^2}\left[\frac{1}{1+\varepsilon}\mathrm{Id}_2 \# M\right].$$

This concludes the proof of (3.15). From now on we assume that (u,v) is supported on $S_0 \subset\subset S$.

To show that (u,v) can be assumed to be smooth, fix $\delta < \operatorname{dist}(S_0, \partial S)$, and set

$$u_\delta(x') = \int_{S_0} \rho_\delta(x'-y')u(y')dy', \qquad v_\delta(x') = \int_{S_0} \rho_\delta(x'-y')v(y')dy',$$

where ρ_δ is a standard mollification kernel on the scale δ, i.e., $\rho_\delta(x') = \delta^{-2}\rho(x'/\delta)$ for $\rho \in C_c^\infty(B^2)$, $\int_{\mathbb{R}^2}\rho = 1$. Then automatically $(u_\delta, v_\delta) \in C_c^\infty(S,\mathbb{R}^2)\times C_c^\infty(S)$, and as $\delta \to 0$ we have $(u_\delta, v_\delta) \to (u,v)$ weakly in $X(S)\times W^{1,2}(S)$. It remains to show that $\limsup_{\delta\to 0} I_0(u_\delta, v_\delta) \le I_0(u,v)$. To see this let $M \in \mathcal{M}(\mathbb{R}^2, \mathbb{R}^{2\times 2}_+)$ be such that $f = \operatorname{sym} D'\overline{u} + M \in L^1(\mathbb{R}^2, \mathbb{R}^{2\times 2}_{\mathrm{sym}})$, and

$$I_0(u,v) \le \frac{1}{2}\int_S Q_2\left((\operatorname{sym} D'u + M)(x') + \frac{\nabla v(x')\otimes\nabla v(x')}{2}\right)dx' + \delta$$

(M and f will depend on δ). Then

$$\nabla' u_\delta(x') = \int_S \rho_\delta(x'-y')f(y')dy' - \int_S \rho_\delta(x'-y')dM(y'),$$

where the second integral takes values in the (convex) set $\mathbb{R}^{2\times 2}_+$. We now use that $W_{\mathrm{F\ddot{o}}}$ is nondecreasing in its (matrix-valued) first argument and that it is convex to obtain

$$\begin{aligned}\int_S W_{\mathrm{F\ddot{o}}}(\nabla' u_\delta, \nabla' v_\delta)dx' &\le \int_S W_{\mathrm{F\ddot{o}}}(\rho_\delta * f, \rho_\delta * \nabla' v)dx' \\ &\le \int_S W_{\mathrm{F\ddot{o}}}(f, \nabla' v)dx'.\end{aligned}$$

On the smooth functions (u_δ, v_δ) we can use (2.4), and since $W_{\mathrm{F\ddot{o}}} \le Q_2$, we get

$$I_0(u_\delta, v_\delta) \le I_0(u,v) + \delta.$$

It remains to prove the thesis for the case $u \in C_c^\infty(S,\mathbb{R}^2)$, $v \in C_c^\infty(S)$. We first show that for every $j \in \mathbb{N}$ we can find $M_j \in L^\infty(S, \mathbb{R}^{2\times 2}_+)$ and $a_j \in C_c^\infty(S,\mathbb{R}^3)$ such that

$$\frac{1}{2}\int_S Q_3\left(\operatorname{sym}\,(\nabla' u + a_j\otimes e_3) + \frac{\nabla' v\otimes\nabla' v}{2} + M_j\right)dx' \le I_0(u,v) + \frac{C}{j}, \tag{3.16}$$

with M_j taking only a finite number of values, each of them on a Lipschitz subset of S. To see this, consider a subdivision of S into small squares, say of side l_j. The

oscillation of the smooth fields ∇u and ∇v on each square is uniformly small; hence—provided l_j is small enough—on each square we can pick one value of a and one value of M so that

$$Q_3\left(\operatorname{sym}\,(\nabla' u + a_j \otimes e_3) + \frac{\nabla' v \otimes \nabla' v}{2} + M_j\right) \le W_{\mathrm{F\ddot{o}}}(\nabla' u, \nabla' v) + \frac{1}{j}\,.$$

Further, on the squares intersecting ∂S we can choose $a = 0$, since u and v have zero boundary values. This defines piecewise constant fields a_j and M_j with the required property. Smoothing a_j concludes the proof of (3.16).

Claim. Given $u \in C_0^\infty(S, \mathbb{R}^2)$, $v \in C_0^\infty(S)$, $a \in C_0^\infty(S, \mathbb{R}^3)$, and $M \in L^\infty(S, \mathbb{R}^{2\times 2}_+)$, taking finitely many values on Lipschitz subsets of S, *there exists* a sequence $y_h \in C^\infty(\Omega_1, \mathbb{R}^3)$ such that $y_h(x) = x' + hx_3 e_3$ for $x \in \partial S \times (-1/2, 1/2)$; the functions u_h and v_h defined as in (2.1) and (2.2) satisfy (3.6), (3.7), and (3.8); the scaled nonlinear strain

$$F_h := \frac{(\nabla_h y_h)^T(\nabla_h y_h) - \mathrm{Id}_3}{2h^{\beta/2}}$$

converges to

$$F_h \to \operatorname{sym}\,(\nabla' u + a \otimes e_3) + \frac{\nabla' v \otimes \nabla' v}{2} + M \quad \text{strongly in } L^2(\Omega_1, \mathbb{R}^{3\times 3}); \tag{3.17}$$

and such that there is a field of rotations $R_h \in L^\infty(\Omega_1, SO(3))$ such that

$$\|R_h^T \nabla_h y_h - \mathrm{Id}_3\|_{L^\infty(S, \mathbb{R}^{3\times 3})} \le C h^{\beta/2} \tag{3.18}$$

for some constant C which does not depend on h.

Assume for the moment that this can be done. By (W1) and (W2) we get

$$W(\nabla_h y_h) = W(R_h^T \nabla_h y_h) = \frac{1}{2} Q_3\left(R_h^T \nabla_h y_h - \mathrm{Id}_3\right) + o(|R_h^T \nabla_h y_h - \mathrm{Id}_3|^2),$$

so that by (3.18) it follows that

$$\lim_{h\to 0} \frac{1}{h^\beta} \int_{\Omega_1} W(\nabla_h y_h) dx = \lim_{h\to 0} \frac{1}{2} \int_{\Omega_1} Q_3\left(G_h\right) dx < \infty,$$

where $G_h := h^{-\beta/2}(R_h^T \nabla_h y_h - \mathrm{Id}_3)$. By (3.18) G_h is bounded in L^∞. Then $F_h - G_h = 2^{-1} h^{\beta/2} G_h^T G_h$ (compare with (3.11)) converges strongly to zero in L^∞, while by (3.17) F_h itself has a strong limit in L^2. Therefore G_h converges strongly in L^2 to the same limit as F_h, and this limit is

$$G(x) := \operatorname{sym}\,(\nabla' u(x') + a(x') \otimes e_3) + \frac{\nabla' v(x') \otimes \nabla' v(x')}{2} + M(x')\,.$$

This expression does not depend on x_3, and recalling (3.16) we get

$$\lim_{h\to 0} \frac{1}{h^\beta} \int_{\Omega_1} W(\nabla_h y_h) dx = \int_{\Omega_1} Q_3\left(G(x)\right) dx = \int_S Q_3\left(G(x')\right) dx' = I_0(u, v) + \frac{C}{j},$$

which is the thesis.

Now we prove the claim. Let us define

$$y_h(x) := x' + hx_3e_3 + h^{\beta/2}(u(x') + \xi_h(x')) + h^{\beta/4}(v(x') + \varphi_h(x'))e_3 \\ + h\,x_3\left(h^{\beta/4}b_h(x') + h^{\beta/2}s_h(x')e_3 + h^{\beta/2}a(x')\right),$$

where $b_h \in C_0^\infty(S,\mathbb{R}^2)$, $s_h \in C_0^\infty(S)$, $\xi_h \in C_0^\infty(S,\mathbb{R}^2)$, and $\varphi_h \in C_0^\infty(S)$ have to be chosen properly. The choice of these spaces ensures that the boundary condition $y_h(x) = x' + hx_3e_3$ for $x \in \partial S \times (-1/2, 1/2)$ is satisfied. Further, we shall choose all those functions to be uniformly Lipschitz (i.e., their gradients are bounded by a constant which can depend on M, u, and v, but not on h).

The linear term in x_3 cancels under integration over $x_3 \in (-1/2, 1/2)$; the sequences u_h and v_h defined via (2.1) and (2.2) satisfy

$$u_h = u + \xi_h, \quad v_h = v + \varphi_h.$$

We shall choose $\xi_h \in C_0^\infty(S,\mathbb{R}^2)$ and $\varphi_h \in C_0^\infty(S)$ in such a way that

$$\xi_h \rightharpoonup 0 \qquad \text{weakly in } W^{1,2}(S,\mathbb{R}^2), \tag{3.19}$$

$$\varphi_h \rightharpoonup 0 \qquad \text{weakly in } W^{1,4}(S), \tag{3.20}$$

$$\|(\nabla')^2\varphi_h\|_{L^\infty(S,\mathbb{R}^{2\times 2})} \le \frac{C}{\varepsilon_h} \tag{3.21}$$

for a suitable sequence $\varepsilon_h \to 0$ as $h \to 0$. Note that (3.19) and (3.20) ensure the convergence properties (3.6), (3.7), and (3.8).

Let us now note that we have

$$\nabla_h y_h = \mathrm{Id}_3 + h^{\beta/4}H_1 + h^{\beta/2}H_2 + h^{1+\beta/4}H_3 + h^{1+\beta/2}H_4, \tag{3.22}$$

where

$$H_1 := e_3 \otimes \nabla' v_h + b_h \otimes e_3,$$
$$H_2 := \nabla' u_h + a \otimes e_3 + s_h e_3 \otimes e_3,$$
$$H_3 := x_3 \nabla' b_h,$$
$$H_4 := x_3(\nabla' s_h + \nabla' a).$$

Expanding the nonlinear strain $(\nabla_h y_h)^T(\nabla_h y_h)$ via the rule $(\mathrm{Id}_3 + F)^T(\mathrm{Id}_3 + F) = \mathrm{Id}_3 + 2\mathrm{sym}\,F + F^TF$ we get

$$(\nabla_h y_h)^T(\nabla_h y_h) - \mathrm{Id}_3 = 2h^{\beta/4}\mathrm{sym}\,H_1 + h^{\beta/2}(2\mathrm{sym}\,H_2 + H_1^TH_1) + o(h^{\beta/2})J_h$$

for a suitable tensor field J_h that we shall consider again later on. In order to obtain a strain of order $h^{\beta/2}$ we need to render H_1 antisymmetric, which can be done by choosing

$$b_h := -\nabla' v_h. \tag{3.23}$$

In this way we find

$$F_h = \mathrm{sym}\,H_2 + \frac{H_1^TH_1}{2} + \frac{o(h^{\beta/2})}{2h^{\beta/2}}\,J_h \\ = \mathrm{sym}\,(\nabla' u_h + a \otimes e_3) + \left(s_h - \frac{|\nabla' v_h|^2}{2}\right)e_3 \otimes e_3 + \frac{\nabla' v_h \otimes \nabla' v_h}{2} + \frac{o(h^{\beta/2})}{2h^{\beta/2}}\,J_h.$$

Since we are looking for (3.17) we choose

$$s_h := -\frac{|\nabla' v_h|^2}{2}, \tag{3.24}$$

and then, in order to have (3.17), it remains to show that (i) ξ_h and φ_h can be chosen in such a way that (3.19), (3.20) hold and

$$\operatorname{sym}(\nabla'\xi_h + \nabla' v \otimes \nabla'\varphi_h) + \frac{\nabla'\varphi_h \otimes \nabla'\varphi_h}{2} \to M \quad \text{strongly in } L^2(S, \mathbb{R}^{2\times 2}); \tag{3.25}$$

and that (ii) the resulting tensor field J_h satisfies

$$\frac{o(h^{\beta/2})}{h^{\beta/2}} J_h \to 0 \quad \text{strongly in } L^2(S, \mathbb{R}^{2\times 2}). \tag{3.26}$$

This can be done as follows. Let us define

$$\xi_h = \psi_h - \varphi_h \nabla' v$$

for some $\psi_h \in W_0^{1,\infty}(S, \mathbb{R}^2)$ to be chosen later. Then we find

$$\begin{aligned}&\operatorname{sym}(\nabla'\xi_h + \nabla' v \otimes \nabla'\varphi_h) + \frac{\nabla'\varphi_h \otimes \nabla'\varphi_h}{2}\\ &= \operatorname{sym}\nabla'\psi_h + \frac{\nabla'\varphi_h \otimes \nabla'\varphi_h}{2} - \varphi_h(\nabla')^2 v.\end{aligned}$$

According to Lemma A.4 we can find $\psi_h \in C_0^\infty(S, \mathbb{R}^2)$ and $\varphi_h \in C_0^\infty(S)$ uniformly Lipschitz and such that (3.20) and (3.21) hold (with an ε_h that we can choose arbitrarily, provided it goes to zero), with

$$\operatorname{sym}\nabla'\psi_h + \frac{\nabla'\varphi_h \otimes \nabla'\varphi_h}{2} \to M$$

strongly in $L^2(S, \mathbb{R}^{2\times 2})$ and $\psi_h \rightharpoonup 0$ weakly in $W_0^{1,2}(S, \mathbb{R}^2)$. As a consequence the resulting sequence ξ_h will satisfy (3.19), and also (3.25) will hold true. We now prove that (3.26) is also true and (3.17) will be established. To this end let us note that, with the above choices of b_h, s_h, ξ_h and φ_h, we have that, for every h,

$$\begin{aligned}&\|H_1\|_{L^\infty(S,\mathbb{R}^{2\times 2})} + \|H_2\|_{L^\infty(S,\mathbb{R}^{2\times 2})} \le C,\\ &\|H_3\|_{L^\infty(S,\mathbb{R}^{2\times 2})} + \|H_4\|_{L^\infty(S,\mathbb{R}^{2\times 2})} \le C(1 + \|(\nabla')^2\varphi_h\|_{L^\infty(S,\mathbb{R}^{2\times 2})}).\end{aligned}$$

Then

$$\frac{o(h^{\beta/2})}{h^{\beta/2}}|J_h| \le C\left(h^{\beta/4} + h^{1-\beta/4}|(\nabla')^2\varphi_h|\right) \le C\left(h^{\beta/4} + \frac{h^{1-\beta/4}}{\varepsilon_h}\right).$$

Since we are working in the regime $0 < \beta < 4$, it suffices to choose $\varepsilon_h = h^{(1-\beta/4)/2}$.

In the end we prove (3.18). First of all let us note that for every $F \in \mathbb{R}^{3\times 3}$ we have

$$\operatorname{dist}(F, SO(3)) \le |\operatorname{sym} F - \mathrm{Id}_3| + C|F - \mathrm{Id}|^2,$$

an inequality that reflects the fact that the tangent space of $SO(3)$ at Id_3 is the space of antisymmetric matrices. Next we consider a measurable field $\mathbb{R}_h : S \to SO(3)$ such that $\operatorname{dist}(\nabla_h y_h, SO(3)) = |R_h - \nabla_h y_h|$. From (3.22) we deduce that $\|\nabla_h y_h - \mathrm{Id}_3\|_{L^\infty(S,\mathbb{R}^{3\times 3})} \le Ch^{\beta/4}$ (in particular R_h is uniquely defined) and that $\|\operatorname{sym}\nabla_h y_h - \mathrm{Id}_3\|_{L^\infty(S,\mathbb{R}^{3\times 3})} \le Ch^{\beta/2}$, as $\operatorname{sym} H_1 = 0$. Thus from the inequality we pointed out above we have $|R_h - \nabla_h y_h| \le Ch^{\beta/2}$, from which (3.18) immediately follows. ∎

4. Membranes with an applied normal force. The Γ-convergence result of Theorem 2.1 permits one to study the behavior of minimizing sequences for the physically relevant case where external forces are included without any a priori assumption on the scaling of the energy. The following discussion clarifies in particular that the scaling of the energy is dictated by the scaling of the applied forces. Precisely, we consider the extended functional

$$F_h(w_h,\Omega_h) = E(w_h,\Omega_h) - \frac{1}{h}\int_{\Omega_h} f_h(x')(w_h(x)-x)\cdot e_3\,dx\,. \tag{4.1}$$

Here and below $w_h \in \mathcal{A}_h$, where

$$\mathcal{A}_h := \left\{ w \in W^{1,2}(\Omega_h;\mathbb{R}^3) : \ \ w(x) = x \text{ on } \partial S \times \left(-\frac{h}{2},\frac{h}{2}\right)\right\}.$$

The normal force f_h is in $H^{-1}(S)$ (the dual space to $H^1_0(S) := W^{1,2}_0(S)$).

For any h and any f_h, one has $F_h(\mathrm{Id},\Omega_h) = 0$, where $\mathrm{Id}(x) = x$ is the identical map. Hence the infimum of F_h is always less than or equal to zero, and if it equals zero it is attained. Therefore we can construct sequences of almost-minimizers, i.e., sequences of deformations $w_h \in \mathcal{A}_h$ such that

$$\lim_{h\to 0} \frac{F_h(w_h,\Omega_h)}{\inf_{\mathcal{A}_h}\{F_h\}} = 1\,, \tag{4.2}$$

with the convention that $0/0 = 1$, and $t/0 = \infty$ for $t > 0$. We are interested in the behavior of such almost-minimizers of F_h for small h, assuming that the external forces f_h scale as h^α.

COROLLARY 4.1. *Let $f_h \in H^{-1}(S)$ be such that*

$$\frac{1}{h^\alpha} f_h \to f \qquad \text{in } H^{-1}(S) \tag{4.3}$$

for some $f \in H^{-1}(S)$ and some $\alpha \in (0,3)$. Let $w_h \in \mathcal{A}_h$ be a sequence of almost-minimizers of F_h, in the sense of (4.2)*. Then there exists a constant C (depending on the sequences $\{f_h\}$ and $\{w_h\}$) such that, for $\beta = 4\alpha/3$,*

$$-Ch^\beta \le \inf_{\mathcal{A}_h}\{F_h\} \le 0\,. \tag{4.4}$$

Moreover, there is a subsequence which converges to a limit $(\bar u,\bar v)$ in the sense specified in Theorem 2.1*. The limit $(\bar u,\bar v)$ minimizes the functional*

$$I_0(u,v) - \int_S f\,v\,dx', \tag{4.5}$$

and

$$\lim_{h\to 0}\frac{1}{h^\beta}\inf_{\mathcal{A}_h}\{F_h\} = \lim_{h\to 0}\frac{1}{h^\beta}F_h(w_h,\Omega_h) = I_0(\bar u,\bar v) - \int_S f\,\bar v\,dx'. \tag{4.6}$$

Finally, let $(\tilde u,\tilde v)$ be any minimizer of $I_0(u,v) - \int_S f v dx'$. Then there is a sequence $w_h \in \mathcal{A}_h$, converging in the sense specified in Theorem 2.1 *to $(\tilde u,\tilde v)$, which fulfills* (4.2) *and* (4.6)*.*

Proof. We define

$$E_h := E(w_h, \Omega_h) = \frac{1}{h}\int_{\Omega_h} W(\nabla w_h(x))dx = \int_{\Omega_1} W(\nabla_y y_h(x))dx \tag{4.7}$$

and first claim that $E_h \to 0$. Indeed, for sufficiently small h, one has $\|f_h\|_{H^{-1}} \le h^\alpha(\|f\|_{H^{-1}} + 1)$ and

$$\begin{aligned}\frac{1}{h}\int_{\Omega_h} f_h(x')(w_h(x) - x)\cdot e_3\, dx &= \int_S f_h(x')\int_{-h/2}^{h/2}(w_h(x) - x)\cdot e_3\, dx_3\, dx' \\ &\le Ch^\alpha\left(\|f\|_{H^{-1}(S)} + 1\right)\|\nabla'\bar v_h\|_{L^2(S)}\,,\end{aligned}$$

where

$$\bar v_h(x') := \int_{-h/2}^{h/2}(w_h(x) - x)\cdot e_3\, dx_3\,.$$

In turn, by assumption (W3) we have

$$|\nabla' w_h|^2 \le C\,(W(\nabla w_h) + 1)\,,$$

and therefore

$$\|\nabla'\bar v_h\|^2_{L^2(S)} \le \frac{1}{h}\|\nabla' w_h\|^2_{L^2(\Omega_h)} \le C(1 + E_h)\,.$$

We conclude that

$$F_h(w_h, \Omega_h) \ge E_h - Ch^\alpha\left(\|f\|_{H^{-1}(S)} + 1\right)(1 + E_h)^{1/2}\,.$$

By (4.2) and the fact that $\inf\{F_h\} \le F_h(\mathrm{Id}, \Omega_h) = 0$, it follows that for sufficiently small h we have

$$F_h(w_h, \Omega_h) \le 0\,. \tag{4.8}$$

Therefore $E_h \to 0$.

The proof of part I of Theorem 2.1 can now be repeated, replacing h^β by E_h in all statements, and no other change. Precisely, (4.7) replaces (3.2), in both (3.4) and (3.5) the estimate is in terms of $CE_h^{1/2}$, and the definitions of u_h and v_h are replaced by

$$\tilde u_h(x') := \frac{1}{E_h^{1/2}}\int_{-1/2}^{1/2} U_h(x', x_3)dx_3\,, \qquad \tilde v_h(x') := \frac{1}{E_h^{1/4}}\int_{-1/2}^{1/2} V_h(x', x_3)dx_3$$

(recall that $y_h(x) = w_h(x' + hx_3e_3) = x' + hx_3e_3 + U_h(x) + V_h(x)e_3$). We also obtain convergence as in (3.6), (3.7), and (3.8) for $\tilde u_h$ and $\tilde v_h$, and to some limit $\tilde u$ and $\tilde v$.

Consider now the forcing term. Arguing as above, we write

$$\frac{1}{h}\int_{\Omega_h} f_h(x')(w_h(x) - x)\cdot e_3\, dx = \int_{\Omega_1} f_h\, V_h\, dx = E_h^{1/4}\int_S f_h\, v_h\, dx'.$$

Since v_h is bounded in $W^{1,2}_0$ and $h^{-\alpha}f_h$ is bounded in the dual space H^{-1}, we obtain

$$F_h(w_h, \Omega_h) \ge E_h - CE_h^{1/4}h^\alpha\,. \tag{4.9}$$

Here and in the following C is a constant which depends on the sequences $\{f_h\}$ and $\{w_h\}$. Recalling (4.8), we see that, for sufficiently small h, $E_h \le CE_h^{1/4}h^\alpha$ and hence

$$E_h \le Ch^{4\alpha/3}. \tag{4.10}$$

In turn, this implies

$$F_h(w_h, \Omega_h) \ge -CE_h^{1/4}h^\alpha \ge -Ch^{4\alpha/3},$$

and (4.4) is proven.

By (4.10) we can apply Theorem 2.1 with $\beta = 4\alpha/3$. Let $\bar u$, $\bar v$ be the limit deformations in the sense of part I of Theorem 2.1, and fix that subsequence. Using part II of Theorem 2.1 and (4.3) we conclude that, along the same subsequence,

$$\liminf_{h\to 0} \inf_{\mathcal{A}_h} \frac{1}{h^{4\alpha/3}} F_h = \liminf_{h\to 0} \frac{1}{h^{4\alpha/3}} F_h(w_h, \Omega_h) \ge I_0(\bar u, \bar v) - \int_S f\,\bar v\,dx'. \tag{4.11}$$

At the same time, from part III of Theorem 2.1 for every $(\tilde u, \tilde v)$ there is a sequence $\tilde w_h$ converging to $(\tilde u, \tilde v)$ and such that

$$\lim_{h\to 0} \frac{1}{h^{4\alpha/3}} F_h(\tilde w_h, \Omega_h) = I_0(\tilde u, \tilde v) - \int_S f\,\tilde v\,dx'.$$

This shows in particular that

$$\liminf_{h\to 0} \inf_{\mathcal{A}_h} \frac{1}{h^{4\alpha/3}} F_h \le I_0(\tilde u, \tilde v) - \int_S f\,\tilde v\,dx';$$

hence the pair (u, v) is a minimizer of the limiting functional, and (4.11) is an equality. This concludes the proof of (4.6). Finally, taking $(\tilde u, \tilde v)$ to be any minimizer of the limiting functional we conclude the proof of the corollary. □

Appendix. We start by briefly analyzing the properties of the space $u \in X(S)$, the convex cone in $BD(S)$ that was introduced in (2.3) and that arises naturally in the determination of the domain of the Γ-limit I_0. General references for the space of functions of bounded deformation $BD(S)$ are, for example, the monograph by Temam [26] and the paper by Ambrosio, Coscia, and Dal Maso [2].

Let us recall that if $u \in BD(S)$, then

$$\operatorname{sym} D'u = f_u(x')dx' + \mu_u,$$

where $\mu_u \in \mathcal{M}(S, \mathbb{R}^{2\times 2}_{\rm sym})$ is singular with respect to the Lebesgue measure on S, and $f_u \in L^1(S, \mathbb{R}^{2\times 2})$ is the density of $\operatorname{sym} D'u$ with respect to the Lebesgue measure. Then $u \in X(S)$ if and only if $\mu_{\overline u} \le 0$, where $\overline u = u$ in S and $\overline u = 0$ in $\mathbb{R}^2 \setminus S$.

The structure of the singular part of the strain $\mu_{\overline u}$ can be further analyzed: indeed, it turns out that there is a rectifiable set J_u in S and that, once we have fixed an orientation of it in $\nu_u \in L^\infty(\mathcal{H}^1 \lfloor J_u, S^1)$, there are functions $u^+, u^- \in L^1(\mathcal{H}^1 \lfloor J_u, \mathbb{R}^2)$, and a measure $(\operatorname{sym} D'u)^c$ singular with respect to both dx' and $\mathcal{H}^1$, such that

$$\mu_{\overline u} = (\operatorname{sym} D'u)^c + \operatorname{sym}((u^+ - u^-) \otimes \nu_u) d\mathcal{H}^1 \lfloor J_u + \operatorname{sym}(-\operatorname{tr}(u) \otimes \nu_S) d\mathcal{H}^1 \lfloor \partial S,$$

where $\operatorname{tr}(u) \in L^1(\mathcal{H}^1 \lfloor \partial S, \mathbb{R}^2)$ is the trace of u on ∂S and ν_S is the outer normal to S. In particular the condition $\mu_{\overline u} \le 0$ implies the compatibility condition

$$u^+(x') - u^-(x') = -\lambda(x')\nu_u(x') \quad \text{for } \mathcal{H}^1\text{-a.e. } x' \in J_u$$

for a suitable $\lambda \in L^1(\mathcal{H}^1 \lfloor J_u, [0,\infty))$ (we recall that $\operatorname{sym} a \otimes b \leq 0$, with $b \neq 0$, if and only if $a = -\lambda b$). The sign condition on the boundary term gives analogously

$$\operatorname{tr}(u)(x') = \lambda(x')\nu_S(x') \quad \text{for } \mathcal{H}^1 \text{ a.e. } x' \in \partial S$$

for a function $\lambda \in L^1(\partial S, [0,\infty))$. The geometric meaning of the condition $(\operatorname{sym} D'u)^c \leq 0$ for the Cantor part of $\operatorname{sym} D'u$ is instead less clear as the validity of the "rank-one property" (established in the space BV by Alberti [1]) in BD is at present unknown.

One could ask if the sign condition $\mu_{\overline{u}} \leq 0$ is sufficient to gain more regularity for the distributional gradient $D'u$. It turns out that this is not the case, in the sense that there are functions in $X(S)$ that are not in $BV(S, \mathbb{R}^2)$. For example, let $S = (-1,1)^2$, and for $i > 2$ let $Q_i = (2^{-i}, 2^{-i+1})^2$. In each Q_i by [21, Theorem 1] (see also [6, Theorem 1]) there is $u_i \in C_0^\infty(Q_i, \mathbb{R}^2)$ such that

$$\int_{Q_i} |\operatorname{sym} \nabla' u_i| dx' \leq 2|Q_i|, \qquad \int_{Q_i} |\nabla' u_i| dx' \geq 2^i.$$

We set $u = u_i$ in Q_i, $u = 0$ on $S \setminus \cup Q_i$. It is clear that u is in $BD(S)$ but not in $BV(S; \mathbb{R}^2)$, and that it has zero trace on ∂S. To show that it is in X, it suffices to check that the symmetric part of the distributional gradient is absolutely continuous with respect to the Lebesgue measure. Since $u \in C^1(S \setminus \{0\}, \mathbb{R}^2)$, it suffices to check that the n-dimensional density of $\operatorname{sym} D'u$ at zero is finite. To this end let ρB^2 be the ball of radius ρ and center in the origin; then

$$|\operatorname{sym} D'u|(\rho B^2) \leq \sum_{\{i : Q_i \cap \rho B^2 \neq \emptyset\}} |\operatorname{sym} D'u|(Q_i) \leq 4|\rho B^2|\,.$$

This concludes the proof.

It is not clear if for the u constructed above we can find a $v \in W_0^{1,2}(S)$ such that $I_0(u,v) < \infty$. In other words, the question of whether the space $\{u \in X(S) : I_0(u,v) < \infty \text{ for some } v \in W_0^{1,2}(S)\}$ is contained in $BV(S, \mathbb{R}^2)$ remains open. It is, however, clear that this space is not more regular than BV. Indeed, let $f : (0,1) \to (0,1)$ be a generic monotonic BV function, and extend it to $\mathbb{R}$ by $f(t) = t$. Then set $u(x) = -(f(x_1) - x_1, 0)$, $v = 0$, $S = (-2,2)^2$. Then $I_0(u,v) < \infty$. This construction provides an example of where the jump and Cantor part of Du are nonzero.

The rest of the appendix is devoted to the statement and proof of some lemmas that were used in the proof of the upper bound. Of particular relevance in the description of the relaxation process of compressive deformations are Lemmas A.3 and A.4.

LEMMA A.1. *Let $F \in \mathbb{R}^{n\times n}$. Then there is $R \in SO(n)$ such that $\operatorname{dist}(F, SO(n)) = |R^T F - \mathrm{Id}_n|$. For all such R, the product $R^T F$ is symmetric.*

Proof. This is well known. We recall the argument for the convenience of the reader. Existence is clear. To show symmetry, observe that by replacing F by $\tilde{F} = R^T F$ one can reduce to the case $R = \mathrm{Id}_n$, i.e., it suffices to show that $\operatorname{dist}(F, SO(n)) = |F - \mathrm{Id}_n|$ implies that F is symmetric. Consider the function

$$f(Q) = |F - Q|^2 = |F|^2 - 2F : Q + |Q|^2$$

(we write $F : G = \operatorname{Tr} F^T G = \sum F_{ij} G_{ij}$). The first and last term are constant (for $Q \in SO(n)$) and can be ignored. That $Q = \mathrm{Id}$ is a local minimum among all $Q \in SO(n)$ implies that the gradient of the linear term $-2F : Q$, i.e., $-2F$, is normal

to the constraint $SO(n)$ at the identity. The tangent space to $SO(n)$ at the identity is the space of skew-symmetric matrices; hence this requirement corresponds to $-2F$ being symmetric. □

LEMMA A.2. *$W_{F\ddot{o}}$ is convex.*

Proof. Choose $\lambda \in (0,1)$, A, $A' \in \mathbb{R}^{2\times 2}_{\mathrm{sym}}$, and b, $b' \in \mathbb{R}^2$, and set

$$A_\lambda = \lambda A + (1-\lambda)A', \qquad b_\lambda = \lambda b + (1-\lambda)b'.$$

We have to show that

$$W_{\mathrm{F\ddot{o}}}(A_\lambda, b_\lambda) \le \lambda W_{\mathrm{F\ddot{o}}}(A,b) + (1-\lambda) W_{\mathrm{F\ddot{o}}}(A',b')\,.$$

The key observation is that

$$b_\lambda \otimes b_\lambda = \lambda b\otimes b + (1-\lambda) b\otimes b - \lambda(1-\lambda)(b-b')\otimes(b-b')\,.$$

Therefore for any $M_\lambda \in \mathbb{R}^{2\times 2}_+$ we have

$$\begin{aligned} W_{\mathrm{F\ddot{o}}}(A_\lambda, b_\lambda) &\le Q_2\left(\operatorname{sym} A_\lambda + \frac{1}{2} b_\lambda \otimes b_\lambda + M_\lambda\right) \\ &= Q_2\left(\lambda\left[\operatorname{sym} A + \frac{b\otimes b}{2}\right] + (1-\lambda)\left[\operatorname{sym} A' + \frac{b'\otimes b'}{2}\right] - M_b + M_\lambda\right), \end{aligned}$$

where $M_b = \lambda(1-\lambda)(b-b')\otimes(b-b') \in \mathbb{R}^{2\times 2}_+$.

Choose $M, M' \in \mathbb{R}^{2\times 2}_+$ so that

$$W_{\mathrm{F\ddot{o}}}(A,b) = Q_2\left(\operatorname{sym} A + \frac{1}{2} b\otimes b + M\right),$$

and the same condition holds for A', b', and M', and set $M_\lambda = \lambda M + (1-\lambda)M' + M_b \in \mathbb{R}^{2\times 2}_+$. Then the previous expression takes the form

$$Q_2\left(\lambda\left[\operatorname{sym} A + \frac{b\otimes b}{2} + M\right] + (1-\lambda)\left[\operatorname{sym} A' + \frac{b'\otimes b'}{2} + M'\right]\right),$$

and the convexity of Q_2 concludes the proof. □

LEMMA A.3. *For each $M \in \mathbb{R}^{2\times 2}_+$ there are $\psi_\delta \in W^{1,\infty}(\mathbb{R}^2,\mathbb{R}^2)$ and $\varphi_\delta \in W^{1,\infty}(\mathbb{R}^2)$ such that*

$$\begin{aligned} &\psi_\delta \overset{*}{\rightharpoonup} 0 \quad \textit{weakly* in } W^{1,\infty}(\mathbb{R}^2,\mathbb{R}^2), \\ &\varphi_\delta \overset{*}{\rightharpoonup} 0 \quad \textit{weakly* in } W^{1,\infty}(\mathbb{R}^2) \end{aligned}$$

as $\delta \to 0$,

$$\operatorname{sym}\nabla'\psi_\delta(x') + \frac{\nabla'\varphi_\delta(x')\otimes\nabla'\varphi_\delta(x')}{2} = M$$

for a.e. $x' \in \mathbb{R}^2$ and $\|\psi_\delta\|_{W^{1,\infty}} + \|\varphi_\delta\|_{W^{1,\infty}} \le C(|M|+1)$.

Proof. Let $\zeta(t)$ be defined as t if $0 < t < 1/2$, as $(1-t)$ if $1/2 < t < 1$ and extended periodically on the rest of $\mathbb{R}$. Let $\zeta_\delta(t) := \delta\zeta(t/\delta)$ for every $\delta > 0$ so that $\zeta_\delta \rightharpoonup^* 0$ weakly* in $W^{1,\infty}(\mathbb{R})$ as $\delta \to 0$.

We can write $M = \lambda_1 a_1 \otimes a_1 + \lambda_2 a_2 \otimes a_2$ for $a_1, a_2 \in S^1$ and $\lambda_1, \lambda_2 \geq 0$. We define

$$\psi_\delta(x') := (\sqrt{\lambda}_1 a_1 - \sqrt{\lambda}_2 a_2)\zeta_\delta((\sqrt{\lambda}_1 a_1 + \sqrt{\lambda}_2 a_2) \cdot x')$$

so that

$$\nabla' \psi_\delta(x') := \zeta'_\delta((\sqrt{\lambda}_1 a_1 + \sqrt{\lambda}_2 a_2) \cdot x')(\sqrt{\lambda}_1 a_1 - \sqrt{\lambda}_2 a_2) \otimes (\sqrt{\lambda}_1 a_1 + \sqrt{\lambda}_2 a_2).$$

In particular

$$\operatorname{sym} \nabla' \psi_\delta(x') = \begin{cases} \lambda_1 a_1 \otimes a_1 - \lambda_2 a_2 \otimes a_2 & \text{on } S^+_\delta, \\ \lambda_2 a_2 \otimes a_2 - \lambda_1 a_1 \otimes a_1 & \text{on } S^-_\delta, \end{cases}$$

where we have put $S^-_\delta = \mathbb{R}^2 \setminus S^+_\delta$ and

$$S^+_\delta := \left\{ x' \in \mathbb{R}^2 : \text{for a } k \in \mathbb{N} \text{ we have } (\sqrt{\lambda}_1 a_1 + \sqrt{\lambda}_2 a_2) \cdot x' \in \left(k\delta, k\delta + \tfrac{1}{2}\delta\right) \right\}.$$

Correspondingly we define

$$\varphi_\delta(x') := \begin{cases} \zeta_\delta(2\sqrt{\lambda}_2\, a_2 \cdot x') & \text{if } x' \in S^+_\delta, \\ \zeta_\delta(-2\sqrt{\lambda}_1\, a_1 \cdot x') & \text{if } x' \in S^-_\delta. \end{cases}$$

Note that $\varphi_\delta \in W^{1,\infty}(\mathbb{R}^2)$. Indeed if $x' \in \overline{S^+_\delta} \cap \overline{S^-_\delta}$ we have that for some $j \in \mathbb{N}$

$$j\delta = 2(\sqrt{\lambda}_1 a_1 + \sqrt{\lambda}_2 a_2) \cdot x',$$

and since ζ_δ is δ-periodic we deduce that φ_δ is continuous on the interfaces, and thus Lipschitz on $\mathbb{R}^2$. On the other hand we have that

$$\nabla' \varphi_\delta(x') = \begin{cases} 2\sqrt{\lambda}_2 \zeta'_\delta(2\sqrt{\lambda}_2\, a_2 \cdot x') a_2 & \text{if } x' \in S^+_\delta, \\ -2\sqrt{\lambda}_1 \zeta'_\delta(-2\sqrt{\lambda}_1\, a_1 \cdot x') a_1 & \text{if } x' \in S^-_\delta, \end{cases}$$

and since $\zeta'_\delta = \pm 1$ a.e. we get

$$\nabla' \varphi_\delta(x') \otimes \nabla' \varphi_\delta(x') = \begin{cases} 4\lambda_2 a_2 \otimes a_2 & \text{if } x' \in S^+_\delta, \\ 4\lambda_1 a_1 \otimes a_1 & \text{if } x' \in S^-_\delta. \end{cases}$$

The thesis follows. □

LEMMA A.4. *Let $M \in L^\infty(S, \mathbb{R}^{2\times 2}_+)$ be constant on each of finitely many Lipschitz subsets S_j covering S, and let $\varepsilon_h \to 0$, $\varepsilon_h > 0$. Then there are $\psi_h \in C^\infty_0(S, \mathbb{R}^2)$ and $\varphi_h \in C^\infty_0(S)$ such that*

$$\begin{aligned} \psi_h &\rightharpoonup 0 \quad \textit{weakly in } W^{1,2}(S, \mathbb{R}^2), \\ \varphi_h &\rightharpoonup 0 \quad \textit{weakly in } W^{1,4}(S), \\ \operatorname{sym} \nabla' \psi_h(x') + \frac{\nabla' \varphi_h(x') \otimes \nabla' \varphi_h(x')}{2} &\to M \quad \textit{strongly in } L^2(S, \mathbb{R}^{2\times 2}), \end{aligned}$$

and

$$\begin{aligned} \varepsilon_h \|(\nabla')^2 \varphi_h\|_{L^\infty(S,\mathbb{R}^{2\times 2})} &\leq 1\,, \\ \|\psi_h\|_{W^{1,\infty}(S,\mathbb{R}^2)} + \|\varphi_h\|_{W^{1,\infty}(S)} &\leq C(\|M\|_{L^\infty(S,\mathbb{R}^{2\times 2})} + 1)\,. \end{aligned}$$

Proof. We can without loss of generality assume that M is constant on the entire S (if not, we perform the construction independently on each S_j).

Let $\tilde{\psi}_\delta$, $\tilde{\varphi}_\delta$ be the functions given by Lemma A.3, let $S_\rho = \{x' \in S : \mathrm{dist}(x', \partial S) > \rho\}$, and let $\eta_\rho \in C_0^\infty(B_\rho, \mathbb{R})$ be a mollification kernel on the scale ρ, i.e., such that

$$\int_{\mathbb{R}^2} \eta_\rho(x')dx' = 1\,, \qquad \int_{\mathbb{R}^2} \rho|\nabla'\eta_\rho(x')| + \rho^2|(\nabla')^2\eta_\rho(x')|dx' \le C\,.$$

We set

$$\psi_{\delta,\rho}(x') = \int_{S_\rho} \tilde{\psi}_\delta(y')\eta_\rho(x'-y')dy'$$

and analogously $\varphi_{\delta,\rho}$. Clearly $\psi_{\delta,\rho} \in C_0^\infty(S, \mathbb{R}^2)$, $\varphi_{\delta,\rho} \in C_c^\infty(S)$, and as $\rho \to 0$

$$\psi_{\delta,\rho} \to \tilde{\psi}_\delta\,, \quad \varphi_{\delta,\rho} \to \tilde{\varphi}_\delta \qquad \text{strongly in } W^{1,2}(S,\mathbb{R}^2), \text{ resp. } W^{1,4}(S).$$

It remains to take a suitable diagonal subsequence. Indeed, for each δ we can choose $\rho(\delta)$ such that

$$\|\psi_{\delta,\rho(\delta)} - \tilde{\psi}_\delta\|_{W^{1,2}(S,\mathbb{R}^2)} + \|\varphi_{\delta,\rho(\delta)} - \tilde{\varphi}_\delta\|_{W^{1,4}(S)} \le \delta\,.$$

This ensures all desired convergence properties as $\delta \to 0$. To include the bound on the second gradient it suffices to choose $\delta(h)$ as the smallest δ for which

$$\varepsilon_h\|(\nabla')^2\varphi_{\delta,\rho(\delta)}\|_{L^\infty(S,\mathbb{R}^{2\times 2})} \le 1.$$

This is possible since $\varepsilon_h \to 0$, and for the same reason $\delta(h) \to 0$. Finally, we set $\psi_h = \psi_{\delta(h),\rho(\delta(h))}$ and define φ_h likewise. ∎

In the proof of Theorem 2.1 we have stated the existence of certain measurable functions. This can be proved by a rather standard application of the measurable selections principle, which is, however, typically disregarded in the literature. We therefore choose to provide here the simple details for the case of interest here.

The basic tool is the following slight simplification of Theorem III.6 in [4].

LEMMA A.5. *Let X be a set with a σ-algebra $\mathcal{F}$, let Y be a complete, separable metric space, and for every $x \in X$ let a nonempty subset $F(x)$ of Y be given in such a way that*

$$\{x \in X : F(x) \cap U \neq \emptyset\} \in \mathcal{F} \tag{A.1}$$

for every open set U in Y.

Then a measurable map $f : X \to Y$ can be defined in such a way that $f(x) \in F(x)$ for every $x \in X$.

For the convenience of the reader we recall the brief proof.

Proof. Let $\{y_k\}_k$ be a countable and dense subset of Y and let $f_0 : X \to Y$ be defined by

$$f_0(x) := y_{k_0(x)},$$
$$k_0(x) := \min\{k \in \mathbb{N} : F(x) \cap B(y_k, 2^0) \neq \emptyset\}.$$

Note that f_0 is measurable as it takes values in $\{y_k\}_k$ and as $(f_0)^{-1}(y_k)$ is measurable for every k, by (A.1). Assume that a measurable $f_j : X \to Y$ has been defined in

such a way that $f_j(x) = y_{k_j(x)}$, for $k_j(x)$ such that $F(x) \cap B(y_{k_j(x)}, 2^{-j}) \neq \emptyset$. Then we define $f_{j+1}(x)$ as

$$f_{j+1}(x) := y_{k_{j+1}(x)},$$
$$k_{j+1}(x) := \min\{k \in \mathbb{N} : F(x) \cap B(y_{k_j(x)}, 2^{-j}) \cap B(y_k, 2^{-j-1}) \neq \emptyset\}.$$

Once again f_{j+1} is measurable by (A.1). Furthermore we have easily that

$$\operatorname{dist}(f_j(x), F(x)) \leq 2^{-j}, \quad \operatorname{dist}(f_j(x), f_{j+1}(x)) \leq 2^{-j+1},$$

so that $\operatorname{dist}(f_j(x), f_{j+h}(x)) \to 0$ as $j \to \infty$ for every h. Since Y is complete for every $x \in X$ we find $f(x) \in F(x)$ such that $f_j(x) \to f(x)$, and in particular the map $f : X \to Y$ is measurable. This completes the proof of the lemma. □

We then state and prove some consequences of this lemma that we have used in the proof of Theorem 2.1.

LEMMA A.6. *Let $M : \Omega \to \mathbb{R}^{n\times n}$ be measurable. Then there is a measurable $R : \Omega \to SO(n)$ such that*

$$|M(x) - R(x)| = \operatorname{dist}(M(x), SO(n)) \quad \forall x \in \Omega.$$

Proof. We apply Lemma A.5 with $X = \Omega$, $\mathcal{F}$ the σ-algebra of the Lebesgue measurable sets of Ω, $Y = SO(n)$, and $F(x) = \{Q \in SO(n) : |Q - M(x)| = \operatorname{dist}(M(x), SO(n))$. Let U be an open set of $SO(3)$ and let U_k be an increasing sequence of compact sets exhausting U. Then

$$\begin{aligned}\{x \in X : F(x) \cap U \neq \emptyset\} &= \{x \in \Omega : \exists\, Q \in U,\ |Q - M(x)| = \operatorname{dist}(M(x), SO(n))\} \\ &= \bigcup_{k\in\mathbb{N}} \{x \in \Omega : \operatorname{dist}(M(x), U_k) = \operatorname{dist}(M(x), SO(n))\}\end{aligned}$$

and each set in this countable union is measurable as it is the coincidence set of two measurable functions. □

Acknowledgment. This research was carried out mainly while all authors were at the Max-Planck Institute for Mathematics in the Sciences, Leipzig, Germany.

REFERENCES

[1] G. ALBERTI, *Rank one property for derivatives of functions with bounded variation*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 239–274.
[2] L. AMBROSIO, A. COSCIA, AND G. DAL MASO, *Fine properties of functions with bounded deformation*, Arch. Ration. Mech. Anal., 139 (1997), pp. 201–238.
[3] H. BEN BELGACEM, S. CONTI, A. DESIMONE, AND S. MÜLLER, *Energy scaling of compressed elastic films—three-dimensional elasticity and reduced theories*, Arch. Ration. Mech. Anal., 164 (2002), pp. 1–37.
[4] C. CASTAIGN AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Springer-Verlag, Berlin, 1977.
[5] P. G. CIARLET, *Mathematical Elasticity*. Vol. II. *Theory of Plates*, Elsevier, Amsterdam, 1997.
[6] S. CONTI, D. FARACO, AND F. MAGGI, *A new approach to counterexamples to L^1 estimates: Korn's inequality, geometric rigidity, and regularity for gradients of separately convex functions*, Arch. Ration. Mech. Anal., 175 (2005), pp. 287–300.
[7] S. CONTI AND F. MAGGI, *Confining Thin Elastic Sheets and Folding Paper*, preprint, 2005.
[8] G. DAL MASO, *An Introduction to Γ-Convergence*, Birkhäuser, Boston, 1993.
[9] A. FÖPPL, *Vorlesungen über technische Mechanik*, Vol. 5, 1907, pp. 132–139.

[10] G. Friesecke, R. D. James, and S. Müller, *A hierarchy of plate models derived from nonlinear elasticity by Gamma-convergence*, Arch. Ration. Mech. Anal., 180 (2006), pp. 183–236.
[11] G. Friesecke, R. D. James, and S. Müller, *The Föppl-von Kármán plate theory as a low energy Γ-limit of nonlinear elasticity*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 201–206.
[12] G. Friesecke, R. D. James, and S. Müller, *A theorem on geometric rigidity and the derivation of nonlinear plate theory from three dimensional elasticity*, Comm. Pure Appl. Math., 55 (2002), pp. 1461–1506.
[13] G. Friesecke, S. Müller, and R. D. James, *Rigorous derivation of nonlinear plate theory and geometric rigidity*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 173–178.
[14] R. M. Head and E. E. Sechler, *Normal Pressure Tests on Unstiffened Flat Plates*, NACA Technical Note 943, 1944.
[15] R. V. Kohn, *New Estimates for Deformations in Terms of Their Strains*, Ph.D. thesis, Princeton University, Princeton, NJ, 1979.
[16] R. V. Kohn, *New integral estimates for deformations in terms of their nonlinear strains*, Arch. Ration. Mech. Anal., 78 (1982), pp. 131–172.
[17] H. Le Dret and A. Raoult, *Le modéle de membrane nonlinéaire comme limite variationelle de l'élasticité non linéaire tridimensionelle*, C. R. Math. Acad. Sci. Paris, 317 (1993), pp. 221–226.
[18] H. Le Dret and A. Raoult, *The nonlinear membrane model as a variational limit of nonlinear three-dimensional elasticity*, J. Math. Pures Appl., 73 (1995), pp. 549–578.
[19] R. Monneau, *Justification of the nonlinear Kirchhoff-Love theory of plates as the application of a new singular inverse method*, Arch. Ration. Mech. Anal., 169 (2003), pp. 1–34.
[20] M. Neubert and A. Sommer, *Rechteckige Blechhaut unter gleichmäßig verteiltem Flüssigkeitsdruck*, Luftfahrtforschung, 17 (1940), pp. 207–210, English translation available as NACA Technical Memorandum 965, 1940.
[21] D. A. Ornstein, *Non-inequality for differential operators in the L^1-norm*, Arch. Ration. Mech. Anal., 11 (1962), pp. 40–49.
[22] O. Pantz, *Une justification partielle du modèle de plaque en flexion par Γ-convergence*, C. R. Math. Acad. Sci. Paris Série I, 332 (2001), pp. 587–592.
[23] O. Pantz, *On the justification of the nonlinear inextensional plate model*, Arch. Ration. Mech. Anal., 167 (2003), pp. 179–209.
[24] E. Sanchez-Palencia, *Statique et dynamique des coques minces.* II. *Cas de flexion pure inhibée. Approximation membranaire*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 531–537.
[25] M. J. Strauss, *Variations of Korn's and Sobolev's inequalities*, in Partial Differential Equations (Univ. California, Berkeley, CA, 1971), Proc. Sympos. Pure Math. 23, AMS, Providence, RI, 1973, pp. 207–214.
[26] R. Temam, *Mathematical Problems in Plasticity*, Bordas, Paris, 1985.
[27] T. von Kármán, *Festigkeitsprobleme im Maschinenbau*, in Encyclopädie der Mathematischen Wissenschaften, Vol. IV/4, Leipzig, 1910, pp. 311–385.
[28] S. Way, *Bending of circular plates with large deflection*, Trans. Am. Soc. Mech. Eng., 56 (1934), pp. 627–632.

SIAM J. MATH. ANAL.
Vol. 38, No. 2, pp. 681–692

A RANGE DESCRIPTION FOR THE PLANAR CIRCULAR RADON TRANSFORM*

GAIK AMBARTSOUMIAN† AND PETER KUCHMENT†

Abstract. The transform considered in the paper integrates a function supported in the unit disk on the plane over all circles centered at the boundary of this disk. Such a circular Radon transform arises in several contemporary imaging techniques, as well as in other applications. As is common for transforms of Radon type, its range has infinite codimension in standard function spaces. Range descriptions for such transforms are known to be very important for computed tomography—for instance, when dealing with incomplete data, error correction, and other issues. A complete range description for the circular Radon transform is obtained. Range conditions include the recently found set of moment-type conditions, which happens to be incomplete, as well as other conditions that have less standard form. In order to explain the procedure better, a similar (nonstandard) treatment of the range conditions is described first for the usual Radon transform on the plane.

Key words. thermoacoustic tomography, Radon transform, range description

AMS subject classifications. 44A12, 92C55, 65R32

DOI. 10.1137/050637492

1. Introduction.

The following "circular" Radon transform, which is the main object of study in this article, arises in several applications, including the newly developing thermoacoustic tomography and its sibling, optoacoustic tomography (e.g., [5, 16, 29, 56, 67, 68, 69, 70]), as well as radar, sonar, and other applications [40, 48, 51, 52]. It has also been considered in relation to some problems of approximation theory, mathematical physics, and other areas [1, 2, 10, 13, 28, 31, 38, 39].

Let $f(x)$ be a continuous function on $\mathbb{R}^d$, $d \geq 2$.

DEFINITION 1. *The circular Radon transform of f is defined as*

$$Rf(p,\rho) = \int_{|y-p|=\rho} f(y)d\sigma(y),$$

where $d\sigma(y)$ is the surface area on the sphere $|y-p| = \rho$ centered at $p \in \mathbb{R}^d$.

In this definition we do not restrict the set of centers p or radii r. It is clear, however, that this mapping is overdetermined, since the dimension of pairs (p, r) is $d+1$, while the function f depends on d variables only. This (as well as the tomographic motivation) suggests that we restrict the set of centers to a set (hypersurface) $S \subset \mathbb{R}^d$, while not imposing any restrictions on the radii. This restricted transform is denoted by R_S:

$$R_S f(p,\rho) = Rf(p,\rho)|_{p\in S}.$$

In this paper we will be dealing with the planar case only, i.e., the dimension d will be equal to 2. Due to tomographic applications, where S is the set of locations of

*Received by the editors August 3, 2005; accepted for publication (in revised form) February 16, 2006; published electronically July 31, 2006. This work was supported in part by NSF grants DMS 9971674 and 0002195. The result of the paper was first presented at the Fully Three-Dimensional Image Reconstruction Meeting in Radiology and Nuclear Medicine, Salt Lake City, UT, 2005.
http://www.siam.org/journals/sima/38-2/63749.html
†Texas A&M University, College Station, TX 77843-3368 (haik@tamu.edu, kuchment@math.tamu.edu).

transducers [29, 67, 68, 70], from now on we will be looking at the specific case when S is the unit circle $|x| = 1$ in the plane.

There are many questions one can ask concerning the circular transform R_S, e.g., concerning its injectivity, inversion formulas, stability of inversion, and range description. Experience in computerized tomography shows (e.g., [44, 46]) that all these questions are important. Although none of them has been resolved completely for R_S, significant developments have occurred recently (see, e.g., [1, 2, 5, 6, 12, 14, 16, 20, 35, 46, 47, 51, 52, 55, 59, 60, 67, 68, 69]). The goal of this article is to describe the range of R_S in the two-dimensional case, with S being the unit circle. Moreover, we will be dealing with functions supported inside the circle S only. The properties of the operator R_S (e.g., stability of the inversion, its Fourier integral operator (FIO) properties, etc.) deteriorate on functions with supports extending outside S (e.g., [2, 16, 70] and remarks in the last section of this article). However, in tomographic applications one normally deals with functions supported inside S only [29, 56, 67, 70].

As was already mentioned, the range of R_S has infinite codimension (e.g., in spaces of smooth functions, see details below) and thus infinitely many range conditions appear. It seems to be a rather standard situation for various types of Radon transforms that range conditions split into two types, one of which is usually easier to discover, while the other "half" is harder to come by. For instance, it took about a decade to find the complete range description for the so-called exponential Radon transform arising in SPECT (single photon emission computed tomography) [3, 4, 33, 34, 66]. For a more general attenuated transform arising in SPECT, it took twice as much time to move from a partial set of range conditions [44, 45] to the complete set [53]. In the circular case, a partial set of such conditions was discovered recently [56]. It happens to be incomplete, and the goal of this text is to find the complete one.

One might ask why is it important to know the range conditions. Such conditions have been used extensively in tomography (as well as in radiation therapy planning, e.g., [8, 9, 30]) for various purposes: completing incomplete data, detecting and correcting measurement errors and hardware imperfections, recovering unknown attenuation, etc. [26, 41, 42, 43, 44, 49, 50, 57, 64, 65]. Thus, as soon as a new Radon-type transform arises in an application, a quest for the range description begins.

In order to explain our approach, we start in the next section with treating a toy example of the standard Radon transform on the plane, where the range conditions are well known (e.g., [13, 17, 18, 19, 25, 44, 46]). Our approach, however, is different from the standard ones and naturally leads to the considerations of the circular transform in the rest of the paper.

2. The case of the planar Radon transform. In this section we will approach in a somewhat nonstandard way the issue of the range description for the standard Radon transform on the plane. Consider a compactly supported smooth function $f(x)$ on the plane and its Radon transform

$$(\mathcal{R}f)(\omega, s) = g(\omega, s) := \int_{x\cdot\omega=s} f(x)dl, \tag{1}$$

where $s \in \mathbb{R}$, $\omega \in S^1$ is a unit vector in $\mathbb{R}^2$, and dl is the arc length measure on the line $x \cdot \omega = s$. We want to describe the range of this transform, say on the space $C_0^\infty(\mathbb{R}^2)$. Such a description is well known (see, e.g., [13, 17, 18, 19, 25, 44, 46], or any other book or survey on Radon transforms or computed tomography).

THEOREM 2. *A function g belongs to the range of the Radon transform on C_0^∞ if and only if the following conditions are satisfied:*

1. $g \in C_0^\infty(S^1 \times \mathbb{R})$;
2. *for any $k \in \mathbb{Z}^+$ the kth moment $G_k(\omega) = \int_{-\infty}^{\infty} s^k g(\omega, s) ds$ is the restriction to the unit circle S^1 of a homogeneous polynomial of ω of degree k;*
3. $g(\omega, s) = g(-\omega, -s)$.

We would like to look at this result from a slighty different perspective, which will allow us to do a similar thing in the case of the circular Radon transform.

In order to do so, let us expand $g(\omega, s)$ into the Fourier series with respect to the polar angle ψ (i.e., $\omega = (\cos\psi, \sin\psi)$):

$$g(\omega, s) = \sum_{n=-\infty}^{\infty} g_n(s) e^{in\psi}. \tag{2}$$

We can now reformulate the last theorem in the following, somewhat unusual way.

THEOREM 3. *A function g belongs to the range of the Radon transform on C_0^∞ if and only if the following conditions are satisfied:*

1. $g \in C_0^\infty(S^1 \times \mathbb{R})$;
2. *for any n, the Mellin transform $Mg_n(\sigma) = \int_0^\infty s^{\sigma-1} g_n(s) ds$ of the nth Fourier coefficient g_n of g vanishes at any pole σ of the function $\Gamma(\frac{\sigma+1-|n|}{2})$;*
3. $g(\omega, s) = g(-\omega, -s)$.

Since the only difference in the statements of these two theorems is in conditions 2, let us check that these conditions mean the same thing in both cases. Indeed, let us expand $g(\omega, s)$ into Fourier series (2) with respect to ψ. Representing $e^{in\psi}$ as the homogeneous polynomial $(\omega_1 + i(sign\, n)\omega_2)^{|n|}$ of ω of degree $|n|$, and noticing that $\omega_1^2 + \omega_2^2 = 1$ on the unit circle, one easily concludes that condition 2 in Theorem 2 is equivalent to the following: the kth moment $\int_{\mathbb{R}} s^k g_n(s) ds$ of the nth Fourier coefficient vanishes for integers $0 \le k < |n|$ such that $k - n$ is even.

Let us now look at condition 2 in Theorem 3, still using the same Fourier expansion. Notice that when $k - |n|$ is a negative even integer, $Mg_n(\sigma)$ is one-half of the moment of order $k = \sigma - 1$ of g_n. Taking into account that $\Gamma(\frac{\sigma+1-|n|}{2}) = \Gamma(\frac{k+2-|n|}{2})$ has poles exactly when $k - |n|$ is a negative even integer, we see that conditions 2 in both theorems are in fact saying the same thing.

One can now ask the question, why should one disguise in the statement of Theorem 3 negative integers as poles of Gamma-function and usual moments as values of Mellin transforms? The answer is that in the less invariant and thus more complex situation of the circular Radon transform, one can formulate a range description in the spirit of Theorem 3, although it is unclear how to get an analogue of the version given in Theorem 2.

As a warm-up, let us derive condition 2 in Theorem 3 directly, without relying on the version given in the preceding theorem. This is in fact an easy by-product of Cormack's inversion procedure; see, e.g., [46, section II.2]. Indeed, if we write down the original function $f(x)$ in polar coordinates $r(\cos\phi, \sin\phi)$ and expand into the Fourier series with respect to the polar angle ϕ,

$$f(r(\cos\phi, \sin\phi)) = \sum_{n=-\infty}^{\infty} f_n(r) e^{in\phi}, \tag{3}$$

then the Fourier coefficients f_n and g_n of the original and of its Radon transform are

related as follows [44, formula (2.17) and further]:

$$M(rf_n(r))(s) = \frac{(Mg_n)(s)}{B_n(s)}, \tag{4}$$

where

$$B_n(s) = \text{const}\frac{\Gamma(s)2^{-s}}{\Gamma((s+1+|n|)/2)\Gamma((s+1-|n|)/2)}. \tag{5}$$

Thus, condition 2 of Theorem 3 guarantees that the function $M(rf_n(r))(s)$ does not develop singularities (which it cannot do for a C_0^∞-function f) at zeros of $B_n(s)$. It is not that hard now to prove also sufficiency in the theorem, applying Cormack's inversion procedure to g satisfying conditions 1–3. However, we are not going to do so, since in the next sections we will devote ourselves to doing a similar thing in the more complicated situation of the circular Radon transform.

3. The circular Radon transform. Formulation of the main result. Let us recall the notion of the Hankel transform (e.g., [11]). For a function $h(r)$ on $\mathbb{R}^+$, one defines its Hankel transform of an integer order n as follows:

$$(\mathcal{H}_n h)(\sigma) = \int_0^\infty J_n(\sigma r)h(r)r\,dr, \tag{6}$$

where the standard notation J_n is used for Bessel functions of the first kind.

As in the introduction, let R_S be the circular Radon transform on the plane that integrates functions compactly supported inside the unit disk D over all circles $|x-p| = \rho$ with centers p located on the unit circle $S = \{p\,|\,|p| = 1\}$. Since this transform commutes with rotations about the origin, the Fourier series expansion with respect to the polar angle partially diagonalizes the operator, and thus the nth Fourier coefficient $g_n(\rho)$ of $g = R_S f$ will depend on the nth coefficient f_n of the original f only. It was shown in [51] that the following relation between these coefficients holds:

$$g_n(\rho) = 2\pi\rho\mathcal{H}_0\{J_n\mathcal{H}_n\{f_n\}\}. \tag{7}$$

For the reader's convenience, we will provide the brief derivation from [51]. Considering a single harmonic $f = f_n(r)e^{in\phi}$ and using polar coordinates, one obtains

$$g_n(\rho) = \int_0^\infty rf_n(r)dr\int_0^{2\pi}\delta\left[(r^2+1-2r\cos\phi)^{1/2}-\rho\right]e^{-in\phi}d\phi. \tag{8}$$

Thus, the computation boils down to evaluating the integral

$$I = \int_0^{2\pi}\delta\left[(r^2+1-2r\cos\phi)^{1/2}-\rho\right]e^{-in\phi}d\phi.$$

Using the standard identity

$$\delta(\rho'-\rho) = \rho\int_0^\infty J_0(\rho' z)J_0(\rho z)z dz$$

and the identity that is easy to obtain from one of the addition formulas, e.g., from [7, formula (4.10.6)]

$$2\pi J_n(az)J_n(bz) = \int_0^{2\pi} J_0[z(a^2+b^2-2ab\cos\phi)^{1/2}]e^{-in\phi}d\phi,$$

one goes from (8) to (7).

Since Hankel transforms are involutive, it is easy to invert (7) and get Norton's inversion formulas [51]

$$f_n = \frac{1}{2\pi}\mathcal{H}_n\left\{\frac{\mathcal{H}_0\{g_n(\rho)/\rho\}}{J_n}\right\}. \tag{9}$$

Now one can clearly see analogies to the case of the Radon transform, where zeros of Bessel functions should probably introduce some range conditions. This happens to be correct and leads to the main result of this article, as follows.

THEOREM 4. *In order for the function $g(p,\rho)$ on $S^1\times\mathbb{R}$ to be representable as $R_S f$ with $f\in C_0^\infty(D)$, it is necessary and sufficient that the following conditions be satisfied:*

1. $g\in C_0^\infty(S^1\times(0,2))$.
2. *For any n, the $2k$th moment $\int_0^\infty \rho^{2k} g_n(\rho)d\rho$ of the nth Fourier coefficient of g vanishes for integers $0\le k<|n|$. (Equivalently, the $2k$th moment $\int_0^\infty \rho^{2k} g(p,\rho)d\rho$ is the restriction to the unit circle S of a (nonhomogeneous) polynomial of p of degree at most k.)*
3. *For any $n\in\mathbb{Z}$, function $\mathcal{H}_0\{g_n(\rho)/\rho\}(\sigma)=\int_0^\infty J_0(\sigma\rho)g_n(\rho)d\rho$ vanishes at any zero $\sigma\neq 0$ of Bessel function J_n. (Equivalently, the nth Fourier coefficient with respect to $p\in S^1$ of the "Bessel moment" $G_\sigma(p)=\int_0^\infty J_0(\sigma\rho)g(p,\rho)d\rho$ vanishes if $\sigma\neq 0$ is a zero of Bessel function J_n.)*

4. Proof of the main result. Let us start by proving necessity, which is rather straightforward. Indeed, the necessity of condition 1 is obvious. Let us prove the second condition. In fact, it has already been established in [56]. Let us repeat for completeness its simple proof. Let k be an integer. Consider the moment of order $2k$ of g:

$$\int_0^\infty \rho^{2k} g(p,\rho)d\rho = \int_{\mathbb{R}^2} |x-p|^{2k} f(x)dx = \int_{\mathbb{R}^2} (|x|^2 - 2x\cdot p + 1)^k f(x)dx \tag{10}$$

(we have taken into account that $|p|=1$). We see that the resulting expression is the restriction to S^1 of a (nonhomogeneous) polynomial of degree k in variable p. Expanding into Fourier series with respect to the polar angle of p, we see that the nth harmonic g_n contributes the following homogeneous polynomial of degree $|n|$ in the variable p:

$$\left(\int_0^\infty \rho^{2k} g_n(\rho)d\rho\right) e^{in\psi}.$$

Here as before $p=(\cos\psi,\sin\psi)$. Thus, for $|n|>k$, this term must vanish, which gives necessity of condition 2. We will return to a discussion of this condition below to add a new twist to it.

Necessity of condition 3 follows immediately from Norton's formula (9), which implies in particular that

$$\mathcal{H}_0\{g_n(\rho)/\rho\} = 2\pi J_n \mathcal{H}_n\{f_n\}.$$

Since both functions J_n and $\mathcal{H}_n\{f_n\}$ are entire, $\mathcal{H}_0\{g_n(\rho)/\rho\}$ vanishes whenever J_n does.

Remark 5. The reader might ask why in the third condition of the theorem we do not take into account the zero root of J_n, which in fact has order n, while nonzero

roots are all simple. The reason is that condition 2 already guarantees that $\sigma = 0$ is zero of order $2n$ of $\mathcal{H}_0\{g_n(\rho)/\rho\}$ (twice higher than that of J_n). Indeed, due to evenness of J_0, function $\mathcal{H}_0\{g_n(\rho)/\rho\}(\sigma)$ is also even. Thus, all odd order derivatives at $\sigma = 0$ vanish. The known Taylor expansion of J_0 at zero leads to the formula

$$\mathcal{H}_0\{g_n(\rho)/\rho\}(\sigma) = \sum_m \frac{(-1)^m}{(m!)^2} \left(\frac{\sigma}{2}\right)^{2m} \int_0^\infty r^{2m} g_n(r) dr.$$

We see now that the moment condition 2 guarantees that $\sigma = 0$ is zero of order $2n$ of $\mathcal{H}_0\{g_n/\rho\}(\sigma)$.

Let us move to the harder part, proving sufficiency. Assume a function g satisfies conditions of the theorem and is supported in $S \times (\epsilon, 2-\epsilon)$ for some positive ϵ. We will show that then $g = R_S f$ for some $f \in C_0^\infty(D_\epsilon)$, where D_ϵ is the disk $|x| < 1 - \epsilon$ in the plane.

Due to Norton's formulas, it is natural to expect the proof to go along the following lines: expand g into the Fourier series $g = \sum_m g_m(\rho) e^{im\psi}$ with respect to the angle variable ψ, then use (9) to construct a function f and then show that f is of an appropriate function class and that its circular Radon transform is equal to g. This is what we are going to do, with a small caveat: instead of constructing f itself, we will construct its two-dimensional Fourier transform. Besides, we will start considering the partial sums of the series $h_n = \sum_{|m| \le n} g_m(\rho) e^{im\phi}$. But first we need to get some simple estimates from below for the Bessel function of the first kind J_n.

LEMMA 6. *On the entire complex plane, except for a disk S_0 centered at the origin and a countable number of disks S_k of radii $\pi/6$ centered at points $\pi(k + \frac{2n+3}{4})$, one has*

$$|J_n(z)| \ge \frac{C e^{|\operatorname{Im} z|}}{\sqrt{|z|}}, \qquad C > 0. \tag{11}$$

Proof. Let us split the complex plane into three parts by a circle S_0 of a radius R (to be chosen later) centered at the origin and a planar strip $\{z = x + iy \,|\, |y| < a\}$, as follows: part I consists of points z satisfying $|z| \ge R$ and $|\operatorname{Im} z| \ge a$; part II consists of points such that $|z| \ge R$ and $|\operatorname{Im} z| < a$; part III is the interior of S_0, i.e., $|z| < R$. It is clearly sufficient to prove the estimate (11) in the first two parts: outside and inside the strip. Using the parity property of J_n, it suffices to consider only the right half plane $\operatorname{Re} z \ge 0$.

The Bessel function of the first kind $J_n(z)$ has the following known asymptotic representation in the sector $|\arg z| \le \pi - \delta$ (e.g., [7, formula (4.8.5)] or [36, formula (5.11.6)]):

$$\begin{aligned} J_n(z) = {} & \sqrt{\frac{2}{\pi z}} \cos\left(z - \frac{\pi n}{2} - \frac{\pi}{4}\right) (1 + O(|z|^{-2})) \\ & - \sqrt{\frac{2}{\pi z}} \sin\left(z - \frac{\pi n}{2} - \frac{\pi}{4}\right) \left(\frac{4n^2 - 1}{8z} + O(|z|^{-3})\right). \end{aligned} \tag{12}$$

Let us start estimating in the first part of the complex plane, i.e., where $|\operatorname{Im} z| > a$ and $|z| > R$ for sufficiently large a and R (and, as we have agreed, $\operatorname{Re} z \ge 0$). There, due to boundedness of $\tan z$ in this region, one concludes that $\frac{\sin z}{z} = \cos z\, (O(|z|^{-1}))$, and thus (12) implies

$$J_n(z) = \sqrt{\frac{2}{\pi z}} \cos\left(z - \frac{\pi n}{2} - \frac{\pi}{4}\right) (1 + O(|z|^{-1})),$$

which in turn for sufficiently large a, R leads to

$$|J_n(z)| \geq \frac{Ce^{|\operatorname{Im} z|}}{\sqrt{|z|}}. \tag{13}$$

In the second part of the plane (right half of the strip), due to boundedness of $\sin z$ we have

$$J_n(z) = \sqrt{\frac{2}{\pi z}} \left[\cos\left(z - \frac{\pi n}{2} - \frac{\pi}{4}\right)(1 + O(|z|^{-2})) + O(|z|^{-1})\right].$$

Consider the system of nonintersecting circles S_k with centers at $z_k = \frac{\pi}{2} + k\pi + \frac{\pi n}{2} + \frac{\pi}{4}$ and radii equal to $\frac{\pi}{6}$. Then outside these circles $|\cos(z - \frac{\pi n}{2} - \frac{\pi}{4})| \geq C$ and

$$|J_n(z)| \geq \frac{C}{\sqrt{|z|}}(1 + O(|z|^{-1})).$$

This implies that for a suitably chosen and sufficiently large R, inside of the strip and outside the circles S_k, we have

$$|J_n(z)| \geq \frac{Ce^{|\operatorname{Im} z|}}{\sqrt{|z|}} \tag{14}$$

for $|z| > R$. This proves the statement of the lemma. □

Let us now return to our task: consider the function g and the partial sums h_n of its Fourier series.

LEMMA 7.

1. *If $g(\phi, \rho) = \sum_m g_m(\rho)e^{im\psi}$ satisfies conditions of Theorem 4 and is supported in $S \times (\epsilon, 2 - \epsilon)$, then each partial sum $h_n = \sum_{|m|<n} g_m(\rho)e^{im\psi}$ does so.*
2. *For any n, h_n is representable as $R_S f_n$ for a function $f_n \in C_0^\infty(D_\epsilon)$.*

Proof. The first statement of the lemma is obvious.

Thus, it is sufficient to prove the second statement for a single term $g = g_n(\rho)e^{in\psi}$. As was just mentioned, we will reconstruct the Fourier transform F of the function f. In order to do this, we will use the standard relation between Fourier and Hankel transforms. As before, let $f(x) = f_n(r)e^{in\phi}$, where $r = |x|$ and ϕ are polar coordinates on $\mathbb{R}^2$. Then the Fourier transform $F(\xi)$ of f at points of the form $\xi = \sigma\omega$, where $\sigma \in \mathbb{C}$ and $\omega = (\cos\psi, \sin\psi) \in \mathbb{R}^2$ can be written up to a constant factor as follows:

$$F(\sigma\omega) = \mathcal{H}_n(f_n)(\sigma)e^{in\psi} \tag{15}$$

(e.g., [11, end of section 14.1]). If we knew that $g = R_S f$, then according to (7) this would mean that

$$F(\sigma\omega) = F(\sigma)e^{in\psi} = \frac{1}{2\pi}\frac{\mathcal{H}_0(g_n(\rho)/\rho)(\sigma)}{J_n(\sigma)}e^{in\psi}. \tag{16}$$

Let us now take (16) as the definition of $F(\sigma\omega)$. Due to the standard parity property of Bessel functions, such an F is a correctly defined function of $\sigma\omega$ for $\sigma \neq 0$ (i.e., $F(\sigma\omega) = F((-\sigma)(-\omega))$). We would like to show that it is the Fourier transform of a function $f \in C_0^\infty(D_\epsilon)$. Let us prove first that F belongs to the Schwartz space $\mathcal{S}(\mathbb{R}^2)$. In order to do so, we need to show its smoothness with respect to the angular variable ψ; smoothness and fast decay with all derivatives in the radial variable σ; as

well as that no singularity arises at the origin, which in principle could, due to usage of polar coordinates. Smoothness with respect to the angular variable is obvious, due to (16). Let us deal with the more complex issue of smoothness and decay with respect to σ. First of all, taking into account that $g_n(\rho)$ is supported inside $(0, 2)$, and due to the standard two-dimensional Paley–Wiener theorem, we conclude that $u(\sigma) = \mathcal{H}_0(g_n(\rho)/\rho)$ is an entire function that satisfies for any N the estimate

$$|u(\sigma)| \le C_N(1 + |\sigma|)^{-N} e^{(2-\epsilon)|\mathrm{Im}\,\sigma|}. \tag{17}$$

According to the range conditions 2 and 3 of the theorem, this function vanishes at all zeros of Bessel function $J_n(\sigma)$ at least to the order of the corresponding zero. This means that function $F(\sigma) = \frac{u(\sigma)}{2\pi J_n(\sigma)}$ is entire. Let us show that it belongs to a Paley–Wiener class.

Indeed, $\mathcal{H}(g_n(\rho)/\rho)$ is an entire function with Paley–Wiener estimate (17). Due to the estimate from below for J_n (11) given in Lemma 6, we conclude that $F(\sigma\omega)$ is an entire function of Paley–Wiener class in the radial directions, uniformly with respect to the polar angle. Namely,

$$|F(\sigma)| \le C_N(1 + |\sigma|)^{-N} e^{(1-\epsilon)|\mathrm{Im}\,\sigma|}. \tag{18}$$

Indeed, outside the family of circles S_k the estimate (11) together with (17) gives the Paley–Wiener estimate (18) we need. Inside these circles, application of the maximum principle finishes the job. Smoothness with respect to the polar angle is obvious. Thus, the only thing one needs to establish to verify that F belongs to the Schwartz class is that F is smooth at the origin. This, however, is the standard question in the Radon transform theory, the answer to which is well known (e.g., [17, pp. 108–109], [18, 19], [25, Chap. 1, proof of Theorem 2.4]). Namely, one needs to establish that for any nonnegative integer k, the kth radial (i.e., with respect to σ) derivative of $F(\sigma\omega)$ at the origin is a homogeneous polynomial of order k with respect to ω. So, let us check that this condition is satisfied in our situation. First of all, the parity of the function F is the same as that of n. Thus, we do not need to worry about the derivatives $F_\sigma^{(k)}|_{\sigma=0}$ with $k-n$ odd, since they are zero automatically. Due to the special single-harmonic form of F, we only need to check that $F_\sigma^{(k)}|_{\sigma=0} = 0$ for $k < |n|$ with $k - n$ even. This, however, as we have already discussed in Remark 5, follows from the moment conditions 2 of the theorem.

Due to the smoothness that we have just established and Paley–Wiener estimates, $F \in \mathcal{S}(\mathbb{R}^2)$. Thus, $F = \hat{f}$ for some $f \in \mathcal{S}(\mathbb{R}^2)$. It remains to show that f is supported inside the disk D_ϵ. Consider the usual Radon transform $\mathcal{R}f(s, \phi)$ of f. According to the standard Fourier-slice theorem [13, 17, 18, 19, 25, 44], the one-dimensional Fourier transform (denoted by a "hat") from the variable s to σ gives (up to a fixed constant factor) the values $\widehat{\mathcal{R}f}(\sigma, \psi) = F(\sigma\omega)$ if, as before, $\omega = (\cos\psi, \sin\psi)$. Here $\mathcal{R}$, as before, denotes the standard Radon transform in the plane. Since functions $F(\sigma\omega)$ of σ, as we have just discussed, are uniformly with respect to ω of a Paley–Wiener class, this implies that $\mathcal{R}f(s, \omega)$ has uniformly with respect to ω bounded support in $|s| < 1 - \epsilon$. Now the "hole theorem" [25, 44] (which is applicable to functions of the Schwartz class), implies that f is supported in D_ϵ.

The last step is to show that $R_S f = g = g_n(r)e^{in\phi}$. This, however, immediately follows from comparing formulas (16) and (7), which finishes the proof of the main lemma, Lemma 7. □

Let us now return to the proof of Theorem 4. We have proven so far that any partial sum h_n of the Fourier series for g belongs to the range of the operator R_S

acting on smooth functions supported inside the disk D_ϵ. The function g itself is the limit of h_n in $C_0^\infty(S\times(\epsilon,2-\epsilon))$. The only thing that remains to be proven is that the range is closed in an appropriate topology. Microlocal analysis can help with this.

Consider R_S as an operator acting from functions defined on the open unit disk D to functions defined on the open cylinder $\Omega = S\times(0,2)$. As such, it is an FIO [21, 23, 58]. If R_S^t is the dual operator, then $E = R_S^t R_S$ is an elliptic pseudodifferential operator of order -1 [21, Theorem 1], [22].[1]

LEMMA 8. *The continuous linear operator $E : H_0^2(D_\epsilon)\mapsto H_{loc}^3(D)$ has zero kernel and closed image.*

Proof. Since $E = R_S^t R_S$, the kernel of this operator coincides with the kernel of R_S acting on $H_0^2(D_\epsilon)$. Since S is closed, it is known that R_S has no compactly supported functions in its kernel [1, 2] (this also follows from analytic ellipticity of E and Theorem 8.5.6 of [27]; see also Lemma 4.4 in [2]). Thus, the statement about the kernel is proven and we only need to prove the closedness of the range.

Let P be a properly supported pseudodifferential parametrix of order 1 for E [63]. Then $PE = I + B$, where B is an infinitely smoothing operator on D. Consider the operator Π that acts as the composition of restriction to D_ϵ and then orthogonal projection onto $H_0^2(D_\epsilon)$ in $H^2(D_\epsilon)$. On $H_0^2(D_\epsilon)$ one has $\Pi PE = I + K$, where K is a compact operator on $H_0^2(D_\epsilon)$. Notice that the operator ΠP is continuous from the Fréchet space $H_{loc}^3(D)$ to $H_0^2(D_\epsilon)$. Due to the Fredholm structure of the operator $\Pi PE = I + K$ acting on $H_0^2(D_\epsilon)$, its kernel is finite-dimensional. Let $M\subset H_0^2(D_\epsilon)$ be a closed subspace of finite codimension complementary to the kernel, so $I + K$ is injective on M and has closed range. Then one can find a bounded operator A in $H_0^2(D_\epsilon)$ such that $A(I + K)$ acts as identity on M. Thus, the operator $A\Pi P$ provides a continuous left inverse to $E : M\mapsto H_{loc}^3(D)$. This shows that the range of E on M is closed in $H_{loc}^3(D)$. On the other hand, the total range of E differs only by a finite dimension from the one on M. Thus, it is also closed. ∎

We can now finish the proof of the theorem. Indeed, the last lemma shows that the function $R_S^t g$, being in the closure of the range, is in fact in the range, and thus can be represented as Ef with some $f\in H_0^2(D_\epsilon)$. In other words, $R_S^t(R_S f - g) = 0$. Since the kernel of R_S^t on compactly supported functions is orthogonal to the range of R_S, we conclude that $R_S f - g = 0$. Since $Ef = R_S^t g$ is smooth, due to ellipticity of E we conclude that f is smooth as well. This concludes the proof of the theorem. ∎

We would like to finish with some remarks.

- It should be possible to prove that the operator R_S in the situation considered in the text is semi-Fredholm between appropriate Sobolev spaces (analogously to the properties of the standard and attenuated Radon transforms; see, e.g., [24, 44]). This would eliminate the necessity of the closedness of the range discussion in the end of the proof of Theorem 4.
  Such a statement could probably be proven either by using FIO techniques, or by controlling dependence on n of the constant C and of the radius of the circle S_0 in Lemma 6. The former approach would be better, being more general.
- Proving compactness of support of function f in Lemma 7, we used the standard Radon transform and the "hole theorem." Instead, one could probably use the fact that Fourier transform of f is, by construction, a Paley–Wiener class CR-function on the three-dimensional variety of points $\sigma\omega$ in $\mathbb{C}^2$ and then use an appropriate mandatory analytic extension theorem in the spirit of [54].

[1] Bolker's injective immersion condition [21, 22], which is needed for validity of this result, is satisfied here, as shown in the proof of Lemma 4.3 in [2].

- We considered the situation most natural for tomographic imaging, when the functions to reconstruct are supported inside the aperture curve S. What happens when the supports of functions extend outside the circle S? It is known that compactly supported functions [2] (or even those belonging to L_p with sufficiently small p [1]) can still be uniquely reconstructed. Necessity of the range conditions we derived apparently still holds, and they are still sufficient for finite Fourier series. However, many things do go wrong in this case. Our proof of the closedness of the range fails (in particular, since Bolker's condition for the corresponding FIO does not hold anymore, which was also the main hurdle in proving the results of [2]). Moreover, the range will no longer be closed. Indeed, reconstruction will become unstable, since due to standard microlocal reasons [32, 35, 40, 61, 70], some parts of the wave front set of the function outside S will not be stably recoverable. This means, in particular, that nonsmooth functions can have smooth circular Radon images. This, in turn implies that the range is not closed in the spaces under consideration, and so sufficiency of the range conditions should fail. We are not sure what kind of range description, if any, could work in this situation. By the way, the nice backprojection-type inversion formulas available in odd dimensions [16] also fail for such functions.
- It would be interesting to understand range conditions in the case of a closed curve S different from a circle. Since our method uses rotational invariance, it is not directly applicable to this situation.
- Our result is stated and proven in two dimensions only. It is possible that a similar approach might work in higher dimensions. As we have been notified by D. Finch, he and Rakesh have recently obtained by different methods some range descriptions in three dimensions [15].

Acknowledgment. The authors thank the NSF for its support. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] M. Agranovsky, C. A. Berenstein, and P. Kuchment, *Approximation by spherical waves in L^p-spaces*, J. Geom. Anal., 6 (1996), pp. 365–383.
[2] M. L. Agranovsky and E. T. Quinto, *Injectivity sets for the Radon transform over circles and complete systems of radial functions*, J. Funct. Anal., 139 (1996), pp. 383–413.
[3] V. Aguilar, L. Ehrenpreis, and P. Kuchment, *Range conditions for the exponential Radon transform*, J. Anal. Math., 68 (1996), pp. 1–13.
[4] V. Aguilar and P. Kuchment, *Range conditions for the multidimensional exponential X-ray transform*, Inverse Problems, 11 (1995), pp. 977–982.
[5] G. Ambartsoumian and P. Kuchment, *On the injectivity of the circular Radon transform arising in thermoacoustic tomography*, Inverse Problems, 21 (2005), pp. 473–485.
[6] L.-E. Andersson, *On the determination of a function from spherical averages*, SIAM J. Math. Anal., 19 (1988), pp. 214–232.
[7] G. E. Andrews, R. Askey, and R. Roy, *Special Functions*, Encyclopedia Math. Appl. 71, Cambridge University Press, Cambridge, UK, 1999.
[8] A. Cormack and E. T. Quinto, *A problem in radiotherapy: Questions of nonnegativity*, Internat. J. Imaging Systems Technology, 1 (1989), pp. 120–124.
[9] A. Cormack and E. T. Quinto, *The mathematics and physics of radiation dose planning*, Contemp. Math., 113 (1990), pp. 41–55.
[10] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Volume* II: *Partial Differential Equations*, Interscience, New York, 1962.
[11] B. Davies, *Integral Transforms and Their Applications*, Springer-Verlag, New York, 2001.

[12] A. Denisjuk, *Integral geometry on the family of semi-spheres*, Fract. Calc. Appl. Anal., 2 (1999), pp. 31–46.
[13] L. Ehrenpreis, *The Universality of the Radon Transform*, Oxford University Press, New York, 2003.
[14] J. A. Fawcett, *Inversion of n-dimensional spherical averages*, SIAM J. Appl. Math., 45 (1985), pp. 336–341.
[15] D. Finch and Rakesh, *The range of the spherical mean value operator for functions supported in a ball*, Inverse Problems, 22 (2006), pp. 923–938.
[16] D. Finch, S. K. Patch, and Rakesh, *Determining a function from its mean values over a family of spheres*, SIAM J. Math. Anal., 35 (2004), pp. 1213–1240.
[17] I. Gelfand, S. Gindikin, and M. Graev, *Integral geometry in affine and projective spaces*, J. Sov. Math., 18 (1980), pp. 39–167.
[18] I. Gelfand, S. Gindikin, and M. Graev, *Selected Topics in Integral Geometry*, Transl. Math. Monogr. 220, AMS, Providence, RI, 2003.
[19] I. Gelfand, M. Graev, and N. Vilenkin, *Generalized Functions, Volume* 5: *Integral Geometry and Representation Theory*, Academic Press, New York, 1965.
[20] S. Gindikin, *Integral geometry on real quadrics*, in Lie Groups and Lie Algebras: E. B. Dynkin's Seminar, Amer. Math. Soc. Transl. Ser. 2, 169, AMS, Providence, RI, 1995, pp. 23–31.
[21] V. Guillemin, *On some results of Gelfand in integral geometry*, in Pseudodifferential Operators and Applications, Proc. Sympos. Pure Math. 43, AMS, Providence, RI, 1985, pp. 149–155.
[22] V. Guillemin and S. Sternberg, *Geometric Asymptotics*, AMS, Providence, RI, 1977.
[23] V. Guillemin and S. Sterenberg, *Some problems of integral geometry and some related problems in microlocal analysis*, Amer. J. Math., 101 (1979), pp. 915–955.
[24] U. Heike, *Single-photon emission computed tomography by inverting the attenuated Radon transform with least-squares collocation*, Inverse Problems, 2 (1986), pp. 307–330.
[25] S. Helgason, *The Radon Transform*, Birkhäuser, Basel, 1980.
[26] A. Hertle, *The identification problem for the constantly attenuated Radon transform*, Math. Z., 197 (1988), pp. 13–19.
[27] L. Hörmander, *The Analysis of Linear Partial Differential Operators, Vol.* 1, Springer-Verlag, New York, 1983.
[28] F. John, *Plane Waves and Spherical Means, Applied to Partial Differential Equations*, Dover, Mineola, NY, 1971.
[29] R. A. Kruger, P. Liu, Y. R. Fang, and C. R. Appledorn, *Photoacoustic ultrasound (PAUS) reconstruction tomography*, Med. Phys., 22 (1995), pp. 1605–1609.
[30] P. Kuchment, *On positivity problems for the Radon transform and some related transforms*, Contemp. Math., 140 (1993), pp. 87–95.
[31] P. Kuchment, *Generalized transforms of Radon type and their applications*, in The Radon Transform, Inverse Problems, and Tomography, G. Olafsson and E. T. Quinto, eds., Proc. Sympos. Appl. Math. 63, AMS, Providence, RI, 2006, pp. 67–91.
[32] P. Kuchment, K. Lancaster, and L. Mogilevskaya, *On the local tomography*, Inverse Problems, 11 (1995), pp. 571–589.
[33] P. Kuchment and S. Lvin, *Paley–Wiener theorem for exponential Radon transform*, Acta Appl. Math., 18 (1990), pp. 251–260.
[34] P. Kuchment and S. Lvin, *The range of the exponential Radon transform*, Soviet Math. Dokl., 42 (1991), pp. 183–184.
[35] P. Kuchment and E. T. Quinto, *Some problems of integral geometry arising in tomography*, in The Universality of the Radon Transform, Oxford University Press, New York, 2003, pp. 681–722.
[36] N. N. Lebedev, *Special Functions and Their Applications*, Dover, New York, 1972.
[37] B. Levin, *Distribution of Zeros of Entire Functions*, Transl. Math. Monogr. 5, AMS, Providence, RI, 1964.
[38] V. Ya. Lin and A. Pinkus, *Fundamentality of ridge functions*, J. Approx. Theory, 75 (1993), pp. 295–311.
[39] V. Ya. Lin and A. Pinkus, *Approximation of multivariable functions*, in Advances in Computational Mathematics, H. P. Dikshit and C. A. Micchelli, eds., World Scientific, River Edge, NJ, 1994, pp. 1–9.
[40] A. K. Louis and E. T. Quinto, *Local tomographic methods in sonar*, in Surveys on Solution Methods for Inverse Problems, Springer-Verlag, Vienna, 2000, pp. 147–154.
[41] S. Lvin, *Data correction and restoration in emission tomography*, in Tomography, Impedance Imaging, and Integral Geometry, Lectures in Appl. Math. 30, AMS, Providence, RI, 1994, pp. 149–155.
[42] C. Mennesier, F. Noo, R. Clackdoyle, G. Bal, and L. Desbat, *Attenuation correction in*

SPECT using consistency conditions for the exponential ray transform, Phys. Med. Biol., 44 (1999), pp. 2483–2510.

[43] F. Natterer, *Exploiting the range of Radon transform in tomography*, in Numerical Treatment of Inverse Problems in Differential and Integral Equations, P. Deuflhard and E. Hairer, eds., Birkhäuser-Verlag, Basel, 1983, pp. 290–303.

[44] F. Natterer, *The Mathematics of Computerized Tomography*, Wiley, New York, 1986; reprinted by SIAM, Philadelphia, 2001.

[45] F. Natterer, *Inversion of the attenuated Radon transform*, Inverse Problems, 17 (2001), pp. 113–119.

[46] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*, SIAM Monogr. Math. Model. Comput. 5, SIAM, Philadelphia, 2001.

[47] S. Nilsson, *Application of Fast Backprojection Techniques for Some Inverse Problems of Integral Geometry*, Linkoeping Studies in Science and Technology, Dissertation 499, Department of Mathematics, Linkoeping University, Linkoeping, Sweden, 1997.

[48] C. J. Nolan and M. Cheney, *Synthetic aperture inversion*, Inverse Problems, 18 (2002), pp. 221–235.

[49] F. Noo, R. Clackdoyle, and J.-M. Wagner, *Inversion of the* 3*D exponential X-ray transform for a half equatorial band and other semi-circular geometries*, Phys. Med. Biol., 47 (2002), pp. 2727–35.

[50] F. Noo and J.-M. Wagner, *Image reconstruction in* 2*D SPECT with* 180° *acquisition*, Inverse Problems, 17 (2001), pp. 1357–1371.

[51] S. J. Norton, *Reconstruction of a two-dimensional reflecting medium over a circular domain: Exact solution*, J. Acoust. Soc. Amer., 67 (1980), pp. 1266–1273.

[52] S. J. Norton and M. Linzer, *Ultrasonic reflectivity imaging in three dimensions: Exact inverse scattering solutions for plane, cylindrical, and spherical apertures*, IEEE Trans, Biomed. Engrg., 28 (1981), pp. 200–202.

[53] R. Novikov, *On the range characterization for the two-dimensional attenuated* X*-ray transform*, Inverse Problems, 18 (2002), pp. 677–700.

[54] O. Öktem, *Extension of separately analytic functions and applications to range characterization of the exponential Radon transform*, in Complex Analysis and Applications (Warsaw, 1997), Ann. Polon. Math., 70 (1998), pp. 195–213.

[55] V. P. Palamodov, *Reconstruction from limited data of arc means*, J. Fourier Anal. Appl., 6 (2000), pp. 25–42.

[56] S. K. Patch, *Thermoacoustic tomography—consistency conditions and the partial scan problem*, Phys. Med. Biol., 49 (2004), pp. 1–11.

[57] I. Ponomarev, *Correction of emission tomography data. Effects of detector displacement and nonconstant sensitivity*, Inverse Problems, 10 (1995), pp. 1–8.

[58] E. T. Quinto, *The dependence of the generalized Radon transform on defining measures*, Trans. Amer. Math. Soc., 257 (1980), pp. 331–346.

[59] E. T. Quinto, *The invertibility of rotation invariant Radon transforms*, J. Math. Anal. Appl., 91 (1983), pp. 510–522.

[60] E. T. Quinto, *Null spaces and ranges for the classical and spherical Radon transforms*, J. Math. Anal. Appl., 90 (1982), pp. 408–420.

[61] E. T. Quinto, *Singularities of the X-ray transform and limited data tomography in* $\mathbb{R}^2$ *and* $\mathbb{R}^3$, SIAM J. Math. Anal., 24 (1993), pp. 1215–1225.

[62] E. T. Quinto, M. Cheney, and P. Kuchment, eds., *Tomography, Impedance Imaging, and Integral Geometry*, Lectures in Appl. Math. 30, AMS, Providence, RI, 1994.

[63] M. Shubin, *Pseudodifferential Operators and Spectral Theory*, 2nd ed., Springer-Verlag, Berlin, 2001.

[64] D. Solmon, *Two inverse problems for the exponential Radon transform*, in Inverse Problems in Action, P. S. Sabatier, ed., Springer-Verlag, Berlin, 1990, pp. 46–53.

[65] D. Solmon, *The identification problem for the exponential Radon transform*, Math. Methods Appl. Sci., 18 (1995), pp. 687–695.

[66] O. Tretiak and C. Metz, *The exponential Radon transform*, SIAM J. Appl. Math., 39 (1980), pp. 341–354.

[67] M. Xu and L.-H. V. Wang, *Time-domain reconstruction for thermoacoustic tomography in a spherical geometry*, IEEE Trans. Med. Imag., 21 (2002), pp. 814–822.

[68] Y. Xu, D. Feng, and L.-H. V. Wang, *Exact frequency-domain reconstruction for thermoacoustic tomography* I: *Planar geometry*, IEEE Trans. Med. Imag., 21 (2002), pp. 823–828.

[69] Y. Xu, M. Xu, and L.-H. V. Wang, *Exact frequency-domain reconstruction for thermoacoustic tomography* II: *Cylindrical geometry*, IEEE Trans. Med. Imag., 21 (2002), pp. 829–833.

[70] Y. Xu, L. Wang, G. Ambartsoumian, and P. Kuchment, *Reconstructions in limited view thermoacoustic tomography*, Med. Phys., 31 (2004), pp. 724–733.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 693–716

PERIODIC ROTATING WAVES IN AN UNDULATING ANNULUS AND THEIR HOMOGENIZATION LIMIT*

BENDONG LOU†

Abstract. We study a mean curvature flow equation in an annulus with periodically undulating boundaries and consider the homogenization limit problem as the period of the boundary undulation tends to zero. We first establish a necessary and sufficient condition for the existence of periodic rotating waves. Then we study how the average rotating speed of the periodic rotating wave depends on the geometry of the boundaries. Our results show that boundary undulation always lowers the speed of a rotating wave. We also determine the homogenization limit of the average rotating speed. Quite surprisingly, this homogenized speed depends only on the maximum opening angles of the domain boundaries.

Key words. mean curvature flow equations, periodic rotating waves, undulating annulus, homogenization limit, average rotating speed

AMS subject classifications. 35K55, 35B10

DOI. 10.1137/05062545X

1. Introduction.

We study a curvature-driven motion of curves in an annulus Ω_m ($m \in \mathbf{N}$), whose boundaries undulate periodically with period $\frac{2\pi}{m}$. The motion of the curves is a mean curvature flow equation,

$$V = A + \kappa, \tag{1.1}$$

where V denotes the normal velocity of plane curves, κ denotes the curvature, and $A > 0$ is a constant. Domain Ω_m is defined as follows: Let $\tilde{g}(s)$, $\tilde{h}(s)$ be 2π-periodic smooth functions satisfying

$$\begin{aligned} &\tilde{h}(0) = 0, \quad \tilde{h}(s) \geq 0, \quad \max_s \tilde{h}'(s) = \tan\alpha_1, \quad \min_s \tilde{h}'(s) = -\tan\beta_1, \\ &\tilde{g}(0) = 0, \quad \tilde{g}(s) \geq 0, \quad \max_s \tilde{g}'(s) = \tan\alpha_2, \quad \min_s \tilde{g}'(s) = -\tan\beta_2 \end{aligned}$$

for some $\alpha_i,\ \beta_i \in (0, \frac{\pi}{2})$ $(i = 1, 2)$. For $m \in \mathbf{N}$, define

$$h(s) := \frac{H}{m}\,\tilde{h}(ms), \qquad g(s) := \frac{G}{m}\,\tilde{g}(ms),$$

where $G > H > 0$ are fixed. In what follows, we use polar coordinates (r, θ) to express points in the plane. Define

$$\Omega_m := \left\{(r,\theta) \;\middle|\; H - h(\theta) < r < G + g(\theta), \quad \theta \in [0, 2\pi]\right\}.$$

Denote the outer (resp., inner) boundary of Ω_m by $\partial_2\Omega_m$ (resp., $\partial_1\Omega_m$) and call α_2 (resp., α_1) its *maximum opening angle* (see Figure 1.1).

*Received by the editors February 28, 2005; accepted for publication (in revised form) February 6, 2006; published electronically July 31, 2006. Part of this paper was published in Proceedings of the 5th East Asia PDE Conference, Gakutosho, Japan, 2005. This work is partially supported by VBL program in Hokkaido University (Japan) and by SRF for ROCS, SEM (China).
http://www.siam.org/journals/sima/38-3/62545.html
†Department of Applied Mathematics, Tongji University, Shanghai, 200092, China (blou@mail.tongji.edu.cn).

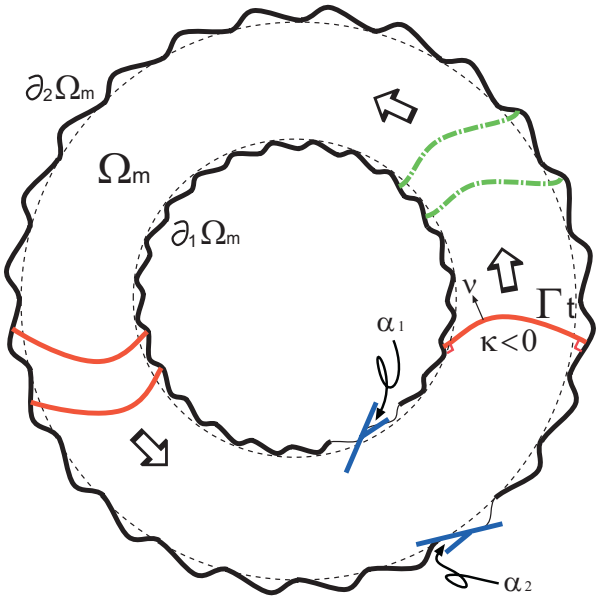


FIG. 1.1. *Undulating annulus and periodic rotating waves.*

By a solution of (1.1) we mean a time-dependent simple curve Γ_t in Ω_m which satisfies (1.1) and contacts both $\partial_1\Omega_m$ and $\partial_2\Omega_m$ perpendicularly. In this paper we are interested in those curves rotating counterclockwise along Ω_m periodically, as well as their average rotating speeds.

In the following we state the motivation and background of (i) the study of mean curvature flow equations; (ii) the study of propagation in annuli; and (iii) the study of propagation in a domain with undulating boundaries.

(i) In 1951 [3] proposed that, in the growth of crystals, the evolution of a crystal surface is governed by a curvature flow equation like (1.1). In 1956 [17] also proposed that the motion of idealized grain boundaries is governed by its curvature. Later *mean curvature flow equations* were used in the study of propagation of wave fronts in an excitable medium, in flame front propagation, in crystal growth, and in many other fields. From a mathematical point of view, it is found that sharp internal layers (or interfaces) appear in singular limit problems of reaction diffusion equations, and the motion of such layers is mean curvature flow equations. For example, a successful model used in the study of Belousov–Zhabotinsky (BZ) reactions are so-called FitzHugh–Nagumo (FHN) equations; the propagation of a pulse solution of FHN equations depends on the mean curvature of the pulse front (e.g., [11], [12], [22]), that is, the motion is a mean curvature flow equation like (1.1).

Mean curvature flow equations are also interesting in their own right from a geometrical point of view. Recently, many geometricians have studied the asymptotic behavior of mean curvature flows ([1], [2], [5], [8], [10], to name only a few). However, as far as we know, very little is known about (periodic) traveling/rotating waves of mean curvature flow equations, though periodic traveling wave solutions of reaction diffusion equations have been studied a lot (cf. [16], [23] and the references therein).

(ii) In the past twenty years many scientists have been interested in BZ reactions in annular gel (for example, [6], [18]). As we said above, one can use FHN equations

to model BZ reactions, and FHN equations can be reduced to a mean curvature flow equation. So the study of mean curvature flow equations in annuli is very helpful to the study of propagation of pulse solutions in BZ reactions in annuli. Some other motivation and studies of rotating waves in annuli can also be found in [9], [20], [22].

(iii) In some cases the domains are not necessarily equipped with flat boundaries. When we consider a propagation in a space with reticulated structure (cf. [4]), or in a media mixed with impurity which blocks the propagation, we may get a domain with undulating boundaries. Such a domain may be an undulating annulus like Ω_m provided the reticulated structure or the masses of impurity are arranged in circles. The propagation of pulses in such an annulus reduces to our problem.

In this paper, we study periodic rotating waves of mean curvature flow (1.1) in annulus Ω_m. We believe that the results in this paper can be extended to more general mean curvature flows.

To avoid sign confusion, the normal vector ν to Γ_t will always be chosen to be counterclockwise; the sign of the normal velocity V and the curvature κ will be understood in accordance with this normal direction (see details below).

We will consider the case that each curve Γ_t is expressed as a graph of a $C^{2,1}$ function $\theta = u(r,t)$ with $(r, u(r,t)) \in \Omega_m$. Let $\eta_1(t)$, $\eta_2(t)$ be the r-coordinates of the end points of Γ_t lying on $\partial_1\Omega_m$, $\partial_2\Omega_m$, respectively. In other words, $\eta_1(t) = H - h(u(\eta_1(t),t))$, $\eta_2(t) = G + g(u(\eta_2(t),t))$. Write the orthogonal coordinates as $(r\cos\theta, r\sin\theta)_\perp$ and denote the unit tangent vector of Γ_t by $\mathbf{T}$ (in the positive direction of r); then $\mathbf{T} = (\cos u - r\sin u \cdot u_r,\ \sin u + r\cos u(r,t)\cdot u_r)_\perp/\sqrt{1+r^2u_r^2}$, $\nu = (-\sin u - r\cos u\cdot u_r,\ \cos u - r\sin u(r,t)\cdot u_r)_\perp/\sqrt{1+r^2u_r^2}$ and

$$V = (-r\sin u\cdot u_t,\ r\cos u\cdot u_t)_\perp\cdot\nu = \frac{ru_t}{\sqrt{1+r^2u_r^2}}, \qquad \kappa = \frac{ru_{rr}+2u_r+r^2u_r^3}{(1+r^2u_r^2)^{3/2}}.$$

Hence (1.1) is equivalent to

$$u_t = \frac{u_{rr}}{1+r^2u_r^2} + \frac{2u_r+r^2u_r^3}{r(1+r^2u_r^2)} + A\frac{\sqrt{1+r^2u_r^2}}{r}, \qquad \eta_1(t) < r < \eta_2(t),\ t>0. \tag{1.2}$$

Denote the *clockwise* unit tangent vector of $r = H$ (resp., $r = G$, $\partial_1\Omega_m$, $\partial_2\Omega_m$) by $\mathbf{T}_1^0$ (resp., $\mathbf{T}_2^0$, $\mathbf{T}_1$, $\mathbf{T}_2$). In what follows, we say that the curve Γ_t contacts $r = H$ (resp., $r = G$, $\partial_1\Omega_m$, $\partial_2\Omega_m$) with angle γ in the sense that $\mathbf{T}\cdot\mathbf{T}_1^0 = \cos\gamma$ (resp., $\mathbf{T}\cdot\mathbf{T}_2^0 = -\cos\gamma$, $\mathbf{T}\cdot\mathbf{T}_1 = \cos\gamma$, $\mathbf{T}\cdot\mathbf{T}_2 = -\cos\gamma$). Our boundary conditions are $\mathbf{T}\cdot\mathbf{T}_i = 0$ on $\partial_i\Omega_m$ $(i=1,2)$, which are expressed as

$$u_r(\eta_1(t),t) = \frac{h'(u)}{(\eta_1(t))^2}, \qquad u_r(\eta_2(t),t) = \frac{-g'(u)}{(\eta_2(t))^2}. \tag{1.3}$$

Let $\Omega_0 = \{(r,\theta) \mid H < r < G,\ \theta\in[0,2\pi]\}$ be the trivial annulus which is *formally* a limit of Ω_m as $m\to\infty$. Problem (1.2) in Ω_0 with boundary conditions

$$u_r(H,t) = u_r(G,t) = 0 \tag{1.4}$$

is quite simple. In fact, as is shown in subsection 2.1.3, when $A > 0$, there exists a unique $\omega_0 > 0$ such that (1.2), (1.4) has a unique *rotating wave* $u(r,t) = \varphi(r) + \omega_0 t$, which has a certain nonplanar profile and rotating speed ω_0. Relevant study in trivial annulus can also be found in [9], [20].

On the other hand, in Ω_m, a rotating wave with a certain profile does not exist. In fact, as Γ_t propagates, its shape and speed fluctuate along with the undulation of the domain Ω_m. In such a situation, we adopt a generalized notion of rotating waves. A solution $U_m(r,t)$ of (1.2)–(1.3) is called a *periodic rotating wave* if it satisfies

$$U_m(r,t+T_m) = U_m(r,t) + \frac{2\pi}{m}$$

for some $T_m > 0$. Clearly, a periodic rotating wave changes its profile periodically in time (see Figure 1.1). The *average rotating speed* of a periodic rotating wave is

$$\omega_m = \frac{2\pi}{mT_m}.$$

In what follows we concentrate our attention on periodic rotating waves with average rotating speed $\omega_m = O(1)$ as $m \to \infty$.

Before stating the main results, we give some assumptions on the boundaries:

$$\alpha_1 + \beta_1 < \frac{\pi}{2}, \qquad \alpha_2 + \beta_2 < \frac{\pi}{2}, \tag{1.5}$$

$$|h'| \cdot \overline{M} < 1, \qquad |g'| \cdot \overline{M} < 1, \tag{1.6}$$

where $\overline{M} = \max\{M_1, M_2\} + 1$ with

$$M_1 := \max\left\{\frac{\tan\alpha_1}{H}, \frac{\tan\beta_1}{H}, \frac{\tan\alpha_2}{G}, \frac{\tan\beta_2}{G}\right\}, \tag{1.7}$$

and $M_2 > 0$ is such that

$$w(2 + 5H^2w^2 + H^4w^4) > A(1 + H^2w^2)^{3/2} \quad \text{for } w > M_2. \tag{1.8}$$

Roughly speaking these conditions require that the undulation of the boundaries be gradual. Assumption (1.5) excludes the possible singularity that the curve touches $\partial\Omega_m$ at some points besides the two end points. Assumption (1.6) guarantees the boundedness of $|u_r|$ on the boundaries (see Appendix A), and it also ensures that we can convert $u(r,t)$ into another unknown $v(z,t)$ defined on $z \in [0,1]$; we can then simply carry out a rigorous proof (see (A.9) in Appendix A). It should be pointed out that these conditions are not necessary ones and can be weakened in special cases.

About the existence we have the following.

THEOREM 1.1. *Assume* (1.5) *and* (1.6) *hold. Then for large* m, (1.2)–(1.3) *has a periodic rotating wave* $U_m(r,t)$ *if and only if*

$$A > \frac{2H\sin\alpha_1 + 2G\sin\alpha_2}{G^2 - H^2}. \tag{1.9}$$

Moreover, a periodic rotating wave is unique up to a time-shift when it exists.

In fact, (1.9) is a necessary and sufficient condition for the existence of rotating *lower* solutions (see subsection 2.1.3).

A more important aim in this paper is to study how the periodic rotating wave and its average speed depend on the shape of the boundaries. In chemical, physical, or biological experiments, traveling/rotating waves can be observed directly. In this sense, people concern themselves with the traveling/rotating speed rather than the

existence, though the latter is important from a mathematical point of view. Generally, traveling/rotating waves in a trivial domain with flat boundaries can be studied by converting the problem to an ODE. However, the propagation in a domain with undulating boundaries (a domain with reticulated structure) cannot be converted to an ODE. We have to deal with a PDE directly. Thus the study for the speed of traveling/rotating waves in undulating domains is important and difficult. Very little is known so far.

We estimate the average rotating speed ω_m and determine its homogenization limit as well as the homogenization limit of $U_m(r,t)$ as $m \to \infty$.

THEOREM 1.2. *Assume* (1.5), (1.6), *and* (1.9) *hold. Then for large* m,

(i) *there exists* $C = C(h', g', H, G, A) > 0$ *such that* ω_m *satisfies*

$$\omega^* - \frac{C}{m} \ < \ \omega_m \ < \ \omega^* + \frac{C}{\sqrt{m}} < \omega_0, \tag{1.10}$$

where $\omega^* = \omega^*(\alpha_1, \alpha_2, H, G, A)$ *is given by the unique solution* $(\omega^*, \varphi^*(r;\omega^*))$ *of*

$$\begin{cases} \omega = \dfrac{\varphi''}{1+r^2\varphi'^2} + \dfrac{2\varphi' + r^2\varphi'^3}{r(1+r^2\varphi'^2)} + A\dfrac{\sqrt{1+r^2\varphi'^2}}{r}, & H < r < G, \\ \varphi'(H) = \dfrac{\tan\alpha_1}{H}, \qquad \varphi'(G) = \dfrac{-\tan\alpha_2}{G}, \end{cases} \tag{1.11}$$

where ω_0 *is given by the unique rotating wave solution of* (1.2), (1.4) *in* Ω_0;

(ii) *as* $m \to \infty$, $U_m(r,t) \to \varphi^*(r;\omega^*) + \omega^* t + C$ *in* $C^{2,1}([H,G] \times [-T,T])$ *for any* $T > 0$, *where* C *is a constant independent of* T.

$\omega_m < \omega_0$ in (1.10) implies that boundary undulation always lowers the speed of the rotating wave, and $\omega^* < \omega_0$ implies that the effect of spatial inhomogeneity of Ω_m is left to the homogenization limit. Moreover, the fact that homogenized speed ω^* depends only on α_1, α_2 (besides H, G, A) is a surprising result.

In [15], we studied periodic traveling waves of (1.1) in an undulating band domain, obtaining results similar to those above. The problem in that paper is different from the present one on several points. First, since the boundaries of an annulus have period 2π anyway, Theorem 1.1 remains valid even if the smallest periods of h and g are 2π, provided the undulation of h and g is gradual. (We omit the detail in this paper.) Second, the backgrounds are different. Mean curvature flows in an *unbounded* band domain are reduced from a traveling front or a traveling pulse, but in a *bounded* annulus can be reduced only from a rotating pulse. Third, the symmetries of a band domain and an annulus are different. This can be seen from the following fact. A flat band domain is quite simple, in which we have a planar traveling wave with speed A, while a trivial annulus is not symmetric in the direction of radius, in which the rotating wave of $V = \kappa + A$ is $\varphi(r) + \omega_0 t$; its profile is not planar. Especially, in case $G - H$ is large, the graph of $\varphi(r)$ may be a spiral which turns around the origin for several rounds (cf. [9], [20], [22]). Fourth, the boundaries of the band domain in [15] are symmetric and hence the boundary conditions are symmetric. But in this paper, outer and inner boundaries of Ω_m are given by different functions h, g, and the boundary conditions (1.3) are also not symmetric.

In section 2, we prove Theorem 1.1. First, we give a *global solution* of an initial-boundary value problem for appropriate initial data. Next, we use the global solution to construct an *entire solution* by using the *renormalization method*. Then we prove the uniqueness (up to a time-shift) of the entire solution; this immediately implies the

existence and uniqueness of *periodic rotating wave*. The necessity of (1.9) is explained in section 2.5. Finally, we state without proof the stability of periodic rotating waves.

In section 3 we prove Theorem 1.2: we estimate the average rotating speed by constructing a precise upper solution. We point out that our construction for an upper solution is a peculiar method, since the upper solution is larger than the solution not only on one period and not for all time, but just in time-interval $[0,1]$. However, this is good enough to give the upper bound of the average rotating speed.

2. Existence of periodic rotating waves.

2.1. Global solutions of the initial-boundary value problem. The proof for the existence of global solutions is divided into several steps. In subsection 2.1.1 we state the comparison principle as a preliminary. In subsection 2.1.2 we construct appropriate initial data $u_0(r)$. In subsection 2.1.3 we study rotating waves in trivial annuli, and select two of them as the lower solution and the upper solution. Using them we give the estimate of $|u|$ in finite time-interval $(0,T]$. In subsection 2.1.4 we give the a priori estimates for $|u_r|$, $|u_r|_\mu$ and for $|u|_\mu$ by converting the problem of $u(r,t)$ to a problem of a new unknown $v(z,t)$ $(z \in (0,1))$. Also we show the global existence, smoothness, and $C^{2+\mu,1+\frac{\mu}{2}}$ bound of u. The proofs in subsection 2.1.4 are long and the idea is similar to that of [15]; we move the details to the appendices.

In what follows, we also write $\frac{1}{m} = \varepsilon$ for convenience.

2.1.1. Comparison principle.

DEFINITION 2.1. *Let $u_1(r,t)$, $u_2(r,t)$ $(t \ge 0)$ be two functions satisfying $(r, u_i(r,t)) \in \overline{\Omega}_m$ $(i = 1,2)$. Then u_1 is called a lower solution of* (1.2)–(1.3) *if*

$$u_{1t} \le \frac{u_{1rr}}{1 + r^2 u_{1r}^2} + \frac{2u_{1r} + r^2 u_{1r}^3}{r(1 + r^2 u_{1r}^2)} + A\frac{\sqrt{1 + r^2 u_{1r}^2}}{r} \tag{2.1}$$

for $t \ge 0$ and r with $(r, u_1(r,t)) \in \Omega_m$,

$$\begin{aligned} &u_{1r}(r,t) \ge \frac{h'(u_1)}{r^2} \qquad \text{for } t \ge 0 \quad \text{and} \quad r \text{ with } (r, u_1(r,t)) \in \partial_1\Omega_m, \\ &u_{1r}(r,t) \le \frac{-g'(u_1)}{r^2} \qquad \text{for } t \ge 0 \quad \text{and} \quad r \text{ with } (r, u_1(r,t)) \in \partial_2\Omega_m. \end{aligned} \tag{2.2}$$

u_2 is said to be an upper solution of (1.2)–(1.3) *if the opposite inequalities hold.*

LEMMA 2.2. *Assume $u_1(r,t)$ and $u_2(r,t)$ are the lower solution and the upper solution of* (1.2)–(1.3) *for $0 \le t < t_1$, respectively. If $u_1(r,0) \le u_2(r,0)$, then $u_1(r,t) \le u_2(r,t)$ for $0 \le t < t_1$ and r with $(r, u_i(r,t)) \in \overline{\Omega}_m$. If $u_1(r,0) \le u_2(r,0)$ and $u_1(r,0) \not\equiv u_2(r,0)$, then $u_1(r,t) < u_2(r,t)$ for $0 < t < t_1$ and r with $(r, u_i(r,t)) \in \overline{\Omega}_m$.*

This lemma follows from the maximum principle.

2.1.2. Appropriate initial data. Assume $\theta_1 \in (0, 2\pi\varepsilon)$ such that $h'(\theta_1) = 0$ and $h'(\theta) \le 0$ for $\theta_1 \le \theta < \theta_1 + \delta_1$ for small $\delta_1 > 0$. Since the period of the boundaries is $2\pi\varepsilon$, there exists $\theta_2 \in (-2\pi\varepsilon + \theta_1, 2\pi\varepsilon + \theta_1)$ such that $g'(\theta_2) = 0$ and $g'(\theta) \le 0$ for $\theta_2 \le \theta < \theta_2 + \delta_1$, provided $\delta_1 > 0$ is small. Denote $P_1 = (H - h(\theta_1), \theta_1)$ and $P_2 = (G + g(\theta_2), \theta_2)$. It is easy to connect P_1 and P_2 by a smooth curve Γ_0 (denote its function by $u_0(r)$) such that the parts of Γ_0 near the boundaries are straight line segments, it contacts $\partial_1\Omega_m$ (resp., $\partial_2\Omega_m$) at P_1 (resp., P_2) vertically, and

$$|u_0(r)| = O(\varepsilon), \quad |u_{0r}(r)| = O(\varepsilon), \quad |u_{0rr}(r)| = O(\varepsilon).$$

Choose $\delta_2 > 0$ small. Then function $u_0(r)+\delta_2 t$ satisfies (2.1). Moreover, for small $\sigma_1 > 0$ and $t \in [0, \sigma_1]$, this function also satisfies (2.2), which is equivalent to saying that the graph of $u_0(r) + \delta_2 t$ contacts $\partial_i \Omega_m$ with angles not less than $\frac{\pi}{2}$. Therefore, $u_0(r) + \delta_2 t$ is a lower solution of (1.2)–(1.3) on $t \in [0, \sigma_1)$.

Assume $u(r, t)$ is a solution of (1.2)–(1.3) with initial data $u_0(r)$ on time-interval $0 \le t < \sigma_2$. Denote $\sigma_3 := \min\{\sigma_1, \sigma_2\}$. Then using the comparison principle to solution $u(r, t)$ and lower solution $u_0(r) + \delta_2 t$ we have

$$u(r, t) \ge u_0(r) + \delta_2 t, \qquad 0 \le t < \sigma_3.$$

For any given $t_1, t_2 \in [0, \sigma_3)$ with $t_2 > t_1$, the above inequality implies that

$$u(r, t_2 - t_1) \ge u_0(r) + \delta_2(t_2 - t_1) > u_0(r).$$

Note that a solution starting at $u(r, t_2 - t_1)$ is nothing but $u(r, t + t_2 - t_1)$. Applying again the comparison principle again to $u(r, t + t_2 - t_1)$ and to solution $u(r, t)$ (which starts at $u_0(r)$), we have

$$u(r, t + t_2 - t_1) > u(r, t), \qquad 0 \le t < \sigma_2 - (t_2 - t_1).$$

Especially, at $t = t_1$, it is $u(r, t_2) > u(r, t_1)$. Since $t_1, t_2 \in (0, \sigma_3)$ with $t_2 > t_1$ can be chosen arbitrarily, we have

$$u_t(r, t) \ge 0, \qquad 0 \le t < \sigma_3.$$

Finally, it is easily seen that this inequality holds indeed on $[0, \sigma_2)$.

2.1.3. Rotating waves in trivial annuli. In this part we study rotating waves in trivial annuli. We shall select two such rotating waves as the lower solution and the upper solution, and then we can give an a priori estimate of $|u|$ for finite time T by using the lower and upper solutions.

Let $\widetilde{H}$ and $\widetilde{G}$ be constants satisfying $\widetilde{H} = H + O(\varepsilon), \widetilde{G} = G + O(\varepsilon)$. Let $\gamma_1, \gamma_2 \in (-\frac{\pi}{2}, \frac{\pi}{2})$ and consider the two-point boundary value problem of the ODE

$$\begin{cases} \omega = \dfrac{\varphi''}{1 + r^2\varphi'^2} + \dfrac{2\varphi' + r^2\varphi'^3}{r(1 + r^2\varphi'^2)} + A\dfrac{\sqrt{1 + r^2\varphi'^2}}{r}, & \widetilde{H} < r < \widetilde{G}, \\ \varphi'(\widetilde{H}) = \dfrac{\tan\gamma_1}{\widetilde{H}}, \quad \varphi'(\widetilde{G}) = \dfrac{-\tan\gamma_2}{\widetilde{G}}, \quad \varphi(\widetilde{H}) = 0. \end{cases} \tag{2.3}$$

If there exist ω and $\varphi(r) = \varphi(r; \omega, \widetilde{H}, \widetilde{G}, \gamma_1, \gamma_2)$ satisfying (2.3), then we call the pair $(\omega, \varphi(r))$ a solution of (2.3). This solution determines a rotating wave $\varphi(r) + \omega t$ of (1.2) in annulus $\{(r, \theta) | \widetilde{H} < r < \widetilde{G}\}$, and its graph contacts $r = \widetilde{H}$ (resp., $r = \widetilde{G}$) with angle $\frac{\pi}{2} + \gamma_1$ (resp., $\frac{\pi}{2} + \gamma_2$).

LEMMA 2.3. (i) *Assume* (1.9) *holds and* ε *is small. Let* $\gamma_i \in [0, \alpha_i + \zeta_i\varepsilon)$ $(i = 1, 2)$ *for* $\zeta_i = O(1)$. *Then* (2.3) *has a unique solution* $(\omega_l, \varphi(r; \omega_l))$, *and* $\omega_l > 0$.

(ii) *Let* $\gamma_i = -\beta_i + O(\varepsilon)$ $(i = 1, 2)$. *Then* (2.3) *has a unique solution* $(\omega_u, \varphi(r; \omega_u))$, *and* $\omega_u > 0$.

(iii) $\omega_l < \omega_u$.

Proof. Set $\psi(r) = \varphi'(r)$, and consider the following initial value problem:

$$\begin{cases} \psi' = \omega(1 + r^2\psi^2) - \dfrac{2\psi + r^2\psi^3}{r} - A\dfrac{(1 + r^2\psi^2)^{3/2}}{r}, & r \ge \widetilde{H}, \\ \psi(\widetilde{H}) = \dfrac{\tan\gamma_1}{\widetilde{H}}. \end{cases} \tag{2.4}$$

For each ω, denote the solution of (2.4) by $\psi(r;\omega)$. It is clear that $\psi(r;\omega)$ is strictly increasing in ω.

(i) First, when $\omega \geq A/\widetilde{H}$, we have $\psi(r) \geq 0$ $(r > \tilde{H})$.

Next, when $\omega = 0$, the solution of (2.4) is

$$\psi(r;0) = \frac{\frac{L}{r} - \frac{Ar}{2}}{r\sqrt{1-(\frac{L}{r} - \frac{Ar}{2})^2}} \qquad \text{for } \ r \in [\widetilde{H}, \tilde{r}),$$

where $L = \frac{A}{2}\widetilde{H}^2 + \widetilde{H}\sin\gamma_1$ and $\tilde{r} = \frac{1+\sqrt{1+2AL}}{A}$. If $\tilde{r} \leq \widetilde{G}$, then for some $\omega_1 > 0$, $\psi(r;\omega_1)$ is defined on $[\widetilde{H}, \widetilde{G}]$ and $\psi(\widetilde{G};\omega_1) < \frac{-\tan\gamma_2}{\widetilde{G}}$. If $\tilde{r} > \widetilde{G}$, then $\psi(\widetilde{G};0) < \frac{-\tan\gamma_2}{\widetilde{G}}$ if and only if the following holds:

$$A > \frac{2\widetilde{H}\sin\gamma_1 + 2\widetilde{G}\sin\gamma_2}{\widetilde{G}^2 - \widetilde{H}^2}. \tag{2.5}$$

This is true since (1.9) holds and ε is small. Then we always have $\omega_2 \geq 0$ such that $\psi(\widetilde{G};\omega_2) < \frac{-\tan\gamma_2}{\widetilde{G}}$.

Therefore, there is a unique $\omega_l > 0$ such that the solution $\psi(r;\omega_l)$ of (2.4) is defined on $[\widetilde{H}, \widetilde{G}]$ and $\psi(\widetilde{G};\omega_l) = \frac{-\tan\gamma_2}{\widetilde{G}}$, which determines a solution of (2.3): $\varphi(r) = \int_{\widetilde{H}}^{r} \psi(\zeta;\omega_l)d\zeta$.

(ii) can be proved in a way similar to (i) above, and (iii) is verified by an easy analysis of (2.4). □

Now we use this lemma to construct lower and upper solutions. We show that for appropriate choice of $\widetilde{H}$, $\widetilde{G}$, γ_1, and γ_2, the rotating wave $\varphi(r;\omega,\widetilde{H},\widetilde{G},\gamma_1,\gamma_2) + \omega t$, given by the unique solution of (2.3), is a lower/upper solution of (1.2)–(1.3).

Remark 2.1. We remark that it is complicated to state the *optimal* lower solution (cf. section 2.5). Here by *optimal* lower solution we mean a rotating wave whose graph contacts $\partial\Omega_m$ with angles not smaller than $\frac{\pi}{2}$ and, at some points, exactly $\frac{\pi}{2}$. For example, even if we choose a rotating wave whose graph contacts $\partial_1\Omega_m$ perpendicularly at point $(H - h(s_1), s_1)$ where $h'(s_1) = H\tan\alpha_1$, and contacts $\partial_2\Omega_m$ perpendicularly at point $(G + g(s_2), s_2)$ where $g'(s_2) = G\tan\alpha_2$, we are not yet clear whether the angles between the graph of the rotating wave and $\partial\Omega_m$ are larger than $\frac{\pi}{2}$ at other places, because this depends on the geometry of $\partial\Omega_m$ and the shape of the rotating wave. So for simplicity, instead of constructing the *optimal* lower solution, we construct a *good* lower solution, which is a rotating wave whose graph contacts $\partial\Omega_m$ with angles not smaller than $\frac{\pi}{2}$, and equals $\frac{\pi}{2} + O(\varepsilon)$ at some points.

In what follows, we shall use many positive constants μ, C, ζ_1, ζ_2, etc., which may be different from line to line and may depend on some of h, g, H, G, A, m and sometimes on t. When such a constant depends only on h', g', H, G, A, or even if it depends on h, g, and m but can be replaced by another constant independent of h, g, and m (as $m \to \infty$), then we will omit the dependence on h, g, h', g', H, G, A, m and just write it simply as μ or C; only when μ or C really depends on m and t do we write it out clearly, i.e., $\mu(t)$, $C(t,m)$, etc.

LEMMA 2.4. *Assume* (1.9) *holds and* ε *is small. Then* (1.2)–(1.3) *has a lower solution* $\widehat{\varphi}(r) + \widehat{\omega}t$ *and upper solution* $\widetilde{\varphi}(r) + \widetilde{\omega}t$.

Let $u(r,t)$ $(t \in [0,T))$ *be the solution of* (1.2)–(1.3) *with initial data* $u_0(r)$ *as in subsection* 2.1.2. *Then for* $t_0 \in [0,T)$ *and* $t \in [0, T - t_0)$ *we have*

$$\widehat{\omega}t - C \leq u(r, t+t_0) - u(r,t_0) \leq \widetilde{\omega}t + C \tag{2.6}$$

for some $C > 0$.

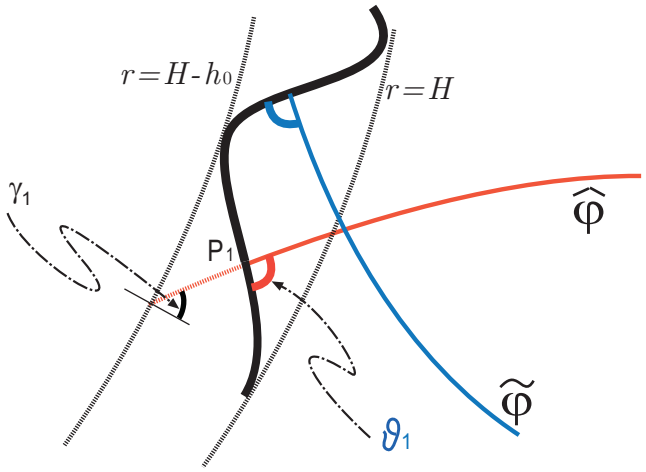


FIG. 2.1. Good *lower solution*.

Proof. Denote $h_0 = \max_s h(s)$, $g_0 = \max_s g(s)$. Consider

$$(2.7)\quad \begin{cases} \omega = \dfrac{\varphi''}{1+r^2\varphi'^2} + \dfrac{2\varphi' + r^2\varphi'^3}{r(1+r^2\varphi'^2)} + A\dfrac{\sqrt{1+r^2\varphi'^2}}{r}, \quad H-h_0 < r < G+g_0, \\ \varphi'(H-h_0) = \dfrac{\tan\gamma_1}{H-h_0}, \qquad \varphi'(G+g_0) = \dfrac{-\tan\gamma_2}{G+g_0}, \qquad \varphi(H-h_0) = 0, \end{cases}$$

with $\gamma_i = \alpha_i + \zeta_i\varepsilon$ $(i = 1, 2)$. Denote the unique solution of this problem by $(\widehat{\omega}, \widehat{\varphi}(r))$. Suppose that the graph of $\widehat{\varphi}(r)$ contacts $\partial_1\Omega_m$ (resp., $\partial_2\Omega_m$) at P_1 (resp., P_2) with angle ϑ_1 (resp., ϑ_2) (see Figure 2.1).

Then a careful analysis shows that, for large ζ_i $(i = 1, 2)$ with order $O(1)$, we have $\vartheta_i \geq \frac{\pi}{2}$ $(i = 1, 2)$, and there exist s_1 and s_2 such that

$$(2.8)\quad \vartheta_1 = \frac{\pi}{2} + O(\varepsilon) \ \text{ at } (H - h(s_1), s_1), \qquad \vartheta_2 = \frac{\pi}{2} + O(\varepsilon) \ \text{ at } (G + g(s_2), s_2).$$

Hence $\widehat{\varphi}(r) + \widehat{\omega}t$ is a *good* lower solution of (1.2)–(1.3).

In a similar way, using (ii) in Lemma 2.3 one can find a solution $(\widetilde{\omega}, \widetilde{\varphi}(r))$ of (2.3) with $\gamma_i = -\beta_i - \zeta_i\varepsilon$. It is easy to see that when $\zeta_i > 0$ is large, the graph of $\widetilde{\varphi}(r) + \widetilde{\omega}t$ contacts $\partial\Omega_\varepsilon$ with angles smaller than $\frac{\pi}{2}$, and equal to $\frac{\pi}{2} + O(\varepsilon)$ at some points. This means that $\widetilde{\varphi}(r) + \widetilde{\omega}t$ is an upper solution of (1.2)–(1.3).

Now if $u(r, t)$ is the solution of (1.2)–(1.3) on $[0, T)$, then we denote $\operatorname{osc} u(r, t) := \max_r u(r, t) - \min_r u(r, t)$ to be the oscillation of $u(r, t)$. By (i) of Lemma 2.5 below, which has nothing to do with this lemma, there exists C such that

$$\operatorname{osc} \widehat{\varphi}(r), \quad \operatorname{osc} \widetilde{\varphi}(r), \quad \operatorname{osc} u(r, t) \leq C.$$

For any $t_0 \in [0, T)$ and $t \in [0, T - t_0)$, we have

$$\widehat{\varphi}(r) + u(0, t_0) - 2C \leq u(r, t_0) \leq \widetilde{\varphi}(r) + u(0, t_0) + 2C,$$

$$\widehat{\varphi}(r) + \widehat{\omega}t + u(0, t_0) - 2C \leq u(r, t + t_0) \leq \widetilde{\varphi}(r) + \widetilde{\omega}t + u(0, t_0) + 2C.$$

Hence

$$\widehat{\omega}t - 6C \leq u(r, t + t_0) - u(r, t_0) \leq \widetilde{\omega}t + 6C. \qquad \square$$

2.1.4. A priori estimates and global existence. In this subsection, we give some a priori estimates and then prove the global existence of solutions of (1.2)–(1.3). Since the proofs are very long and are not our main purpose in this paper, we move them to the appendices.

For any $T > 0$, denote $Q_T := \{(r,t) \mid \eta_1(t) < r < \eta_2(t) \text{ and } 0 < t \le T\}$.

LEMMA 2.5. *Let $u(r,t) \in C^{2,1}(\overline{Q_T})$ be a solution of (1.2)–(1.3) with initial data $u_0(r)$. Then for $(r,t) \in \overline{Q_T}$ we have*

(i) $|u_r(r,t)| < \overline{M}$;

(ii) *there exist $\mu(T) > 0$ and $C > 0$ (independent of T) such that $|u_r|_\mu \le C$;*

(iii) *there exist $\mu(T) > 0$ and $C > 0$ (independent of T) such that $|u|_\mu \le C$.*

Based on Lemmas 2.4 and 2.5, we have the following result.

LEMMA 2.6. *Assume (1.9) holds. Then (1.2)–(1.3) with initial data u_0 has a unique, global solution $u(r,t)$ which satisfies $u_t(\cdot,t) \ge 0$.*

Moreover, for any $T > 0$, there exist positive constants $\mu(T)$, C_1, and C_2 (C_1, C_2 are independent of T) such that

$$u \in C^{2+\mu,1+\frac{\mu}{2}}(\overline{Q_T}) \qquad \text{and} \qquad \|u(r,t)\|_{C^{2+\mu,1+\frac{\mu}{2}}(\overline{Q_T})} \le C_1 T + C_2. \tag{2.9}$$

2.2. Existence of entire solutions. A solution defined on $t \in (-\infty,\infty)$ is called an *entire solution*. We use the *renormalization method* to show the existence of entire solutions.

LEMMA 2.7. *Equations (1.2)–(1.3) have an entire solution $U(r,t)$ such that $U_t(r,t) \ge 0$ and*

$$\widehat{\omega}t - \widehat{C} \le U(r,t+t_0) - U(r,t_0) \le \widetilde{\omega}t + \widetilde{C} \qquad \text{for } t_0 \in \mathbf{R} \text{ and } t \ge 0, \tag{2.10}$$

for some $\widehat{C}, \widetilde{C} > 0$.

Proof. Let u be the global solution of (1.2)–(1.3) obtained in Lemma 2.6. Take $t_n \to \infty$ in the following way:

$$\max_r u(r,t_n) = n \cdot 2\pi\varepsilon \qquad (n = n_0, n_0+1, \ldots),$$

where n_0 is a large integer. Set

$$u_n(r,t) := u(r,t+t_n) - n \cdot 2\pi\varepsilon.$$

Then u_n also satisfies (1.2)–(1.3) for $-t_n \le t < \infty$, and

$$\max_r u_n(r,0) = 0, \qquad \frac{\partial u_n}{\partial t} \ge 0 \qquad (n = n_0, n_0+1, \ldots).$$

By (2.6), there exist $\widehat{C}, \widetilde{C} > 0$ such that

$$\widehat{\omega}t - \widehat{C} \le u_n(r,t) \le \widetilde{\omega}t + \widetilde{C} \qquad \text{for} \quad -t_n \le t < \infty. \tag{2.11}$$

For any given $T > 0$, consider the problem about u_n (for large n) on $[-T,T]$. One can see that (i)–(iii) of Lemma 2.5 and (2.9) remain valid for u_n, the constant μ depends on T, and neither μ nor C depend on n. Therefore there exist $\mu = \mu(T) > 0$, $U(r,t) \in C^{2+\mu,1+\frac{\mu}{2}}(\overline{Q_T^U})$, and a sequence $n_j \to \infty$ ($j \to \infty$) such that

$$u_{n_j}(r,t) \to U(r,t) \qquad \text{in} \quad C^{2+\mu,1+\frac{\mu}{2}}(\overline{Q_T^U}),$$

where $Q_T^U := \{(r,t) \mid t \in [-T,T], r \text{ with } (r,U(r,t)) \in \Omega_m\}$.

Taking $T \to \infty$ and using Cantor's diagonal argument, one finds that there exists a sequence, still writing it as n_j ($n_j \to \infty$ as $j \to \infty$) and $U(r,t) \in C^{2,1}(\overline{Q_\infty^U})$ with $Q_\infty^U = \lim_{T\to\infty} Q_T^z$, such that for any $T > 0$, $U(r,t) \in C^{2+\mu(T),1+\frac{\mu(T)}{2}}(\overline{Q_T^U})$ for some $\mu(T) > 0$ and

$$u_{n_j}(r,t) \to U(r,t) \qquad \text{in} \quad C^{2+\mu(T),1+\frac{\mu(T)}{2}}(\overline{Q_T^U}).$$

Hence U is an entire solution of (1.2)–(1.3). It is also easy to see that U satisfies (2.10), $\max_r U(r,0) = 0$, and satisfies all the conclusions for u in Lemmas 2.5 and 2.6. □

2.3. Uniqueness of entire solution. Assume $U(r,t)$ and $W(r,t)$ are two entire solutions of (1.2)–(1.3) satisfying (2.10) and the conclusions for u in Lemmas 2.5 and 2.6. We shall prove that U is a time-shift of W. Define

$$\Lambda_{U,W}(t) := \inf\{\Lambda > 0 \mid \exists\, a \in \mathbf{R} \text{ such that } U(r,t+a) \le W(r,t) \le U(r,t+a+\Lambda)\}.$$

LEMMA 2.8. (i) *$\Lambda_{U,W}(t)$ is monotone decreasing, and there exists $M > 0$ such that $0 \le \Lambda_{U,W}(t) \le M$ for $t \in \mathbf{R}$.*

(ii) *If $\Lambda_{U,W}(t_0) = 0$ for some t_0, then there exists $a \in \mathbf{R}$ such that $U(r,t+a) \equiv W(r,t)$ for $t \ge t_0$. If $\Lambda_{U,W}(t_0) > 0$ for some t_0, then $\Lambda_{U,W}(t) > 0$ and is strictly decreasing for $t < t_0$.*

Proof. (i) For any $t \in \mathbf{R}$, by the definition of $\Lambda_{U,W}(t)$ there exist r_1 and r_2 such that

$$U(r_1, t+a) = W(r_1,t), \quad W(r_2,t) = U(r_2, t+a+\Lambda_{U,W}(t)),$$

so

$$\begin{aligned}
&\max_r U(r,t+a+\Lambda_{U,W}(t)) - \min_r U(r,t+a)\\
&\le U(r_2,t+a+\Lambda_{U,W}(t)) + \operatorname{osc} U - (U(r_1,t+a) - \operatorname{osc} U)\\
&\le W(r_2,t) - W(r_1,t) + 2\overline{M}(G-H) \le 3\overline{M}(G-H).
\end{aligned}$$

On the other hand, (2.10) implies that

$$\begin{aligned}
&\max_r U(r,t+a+\Lambda_{U,W}(t)) - \min_r U(r,t+a)\\
&\ge U(r,t+a+\Lambda_{U,W}(t)) - U(r,t+a) \ge \widehat{\omega}\cdot\Lambda_{U,W}(t) - \widehat{C}.
\end{aligned}$$

Hence

$$0 \le \Lambda_{U,W}(t) \le \frac{3\overline{M}(G-H) + \widehat{C}}{\widehat{\omega}}.$$

(ii) The first statement in (ii) is clear by the uniqueness. If $\Lambda_{U,W}(t_0) > 0$ for some t_0, then for any $t < t_0$ we have

$$U(r,t+a) \le W(r,t) \le U(r,t+a+\Lambda_{U,W}(t)) \qquad \text{for some } a = a(t).$$

By the strong comparison principle, after time $\tau = t_0 - t > 0$ we have

$$U(r,t+\tau+a(t)) < W(r,t+\tau) < U(r,t+\tau+a(t)+\Lambda_{U,W}(t)),$$

i.e.,

$$U(r,t_0+a(t)) < W(r,t_0) < U(r,t_0+a(t)+\Lambda_{U,W}(t)).$$

By the definition of $\Lambda_{U,W}(t_0)$ we have

$$U(r,t_0+a(t_0)) \le W(r,t) \le U(r,t_0+a(t_0)+\Lambda_{U,W}(t_0))$$

for some $a(t_0)$; each of the two equalities holds at some r. Since $U_t \ge 0$ we have

$$a(t) < a(t_0) \qquad \text{and} \qquad a(t)+\Lambda_{U,W}(t) > a(t_0)+\Lambda_{U,W}(t_0),$$

and so $\Lambda_{U,W}(t) > \Lambda_{U,W}(t_0)$. □

It is also easy to know the following.

LEMMA 2.9. *Let U_n and W_n be two sequences of entire solutions of* (1.2)–(1.3). *If $U_n \to U_\infty$ for $t \in \mathbf{R}$ and r with $(r, U_\infty(r,t)) \in \Omega_m$, $W_n \to W_\infty$ for all $t \in \mathbf{R}$ and r with $(r, W_\infty(r,t)) \in \Omega_m$, then $\Lambda_{U_n,W_n}(t) \to \Lambda_{U_\infty,W_\infty}(t)$ for every t.*

Our aim in this section is to prove the following.

LEMMA 2.10. *$W(r,t)$ is a time-shift of $U(r,t)$.*

Proof. We need to show only that $\Lambda_{U,W}(t) = 0$ for all $t \in \mathbf{R}$. If this is not true, then $\Lambda_{U,W}(t_0) > 0$ for some t_0.

By the monotonicity and boundedness of $\Lambda_{U,W}(t)$, we have $\lim_{t\to-\infty} \Lambda_{U,W}(t) =: \bar\Lambda$ for some $\bar\Lambda$ satisfying $0 < \Lambda_{U,W}(t_0) < \bar\Lambda \le M$.

Set $l_n = \left[\max_r U(r,-n) \cdot \frac{m}{2\pi}\right]$ and define

$$U_n(r,t) := U(r,t-n) - \frac{2\pi}{m} l_n, \qquad W_n(r,t) := W(r,t-n) - \frac{2\pi}{m} l_n.$$

Then both of U_n and W_n satisfy the inequalities for u_n in (2.11), and so a discussion similar to that in the proof of Lemma 2.7 shows that there exists a sequence $n_j \to \infty$ ($j \to \infty$) and $U_\infty, W_\infty \in C^{2,1}$, which are entire solutions of (1.2)–(1.3), such that as $j \to \infty$,

$$U_{n_j}(r,t) \to U_\infty(r,t) \qquad \text{for } t \in \mathbf{R} \quad \text{and} \quad r \text{ with } (r, U_\infty(r,t)) \in \Omega_m,$$

$$W_{n_j}(r,t) \to W_\infty(r,t) \qquad \text{for } t \in \mathbf{R} \quad \text{and} \quad r \text{ with } (r, W_\infty(r,t)) \in \Omega_m.$$

It follows from Lemma 2.9 that $\Lambda_{U_\infty,W_\infty}(t) = \lim_{j\to\infty} \Lambda_{U_{n_j},W_{n_j}}(t)$.

On the other hand, $\Lambda_{U_{n_j},W_{n_j}}(t) = \Lambda_{U,W}(t-n_j)$, so $\Lambda_{U_\infty,W_\infty}(t) = \lim_{j\to\infty} \Lambda_{U,W}(t-n_j) = \bar\Lambda$, that is, $\Lambda_{U_\infty,W_\infty}(t) \equiv \bar\Lambda$ ($t \in \mathbf{R}$). Applying (ii) of Lemma 2.8 to functions U_∞ and W_∞ we see that this is true only if $\bar\Lambda = 0$, a contradiction to $\bar\Lambda > \Lambda_{U,W}(t_0) > 0$.

Therefore, $\Lambda_{U,W}(t) = 0$ (for all $t \in \mathbf{R}$), and so there exists a_0 such that $U(r,t+a_0) \equiv W(r,t)$ for $t \in \mathbf{R}$. □

2.4. Existence and uniqueness of periodic rotating wave.

Proof for the sufficiency part of Theorem 1.1. In previous sections we obtained an entire solution $U(r,t)$ of (1.2)–(1.3). Clearly, $U(r,t)+\frac{2\pi}{m}$ is also an entire solution; Lemma 2.10 implies that $U(r,t)+\frac{2\pi}{m}$ is a time-shift of $U(r,t)$, i.e., there exists $T_m > 0$ such that

$$U(r,t) + \frac{2\pi}{m} = U(r,t+T_m) \qquad \text{for } t \in \mathbf{R} \quad \text{and} \quad r \text{ with } (r, U(r,t)) \in \Omega_m.$$

In other words, $U(r,t)$ is a periodic rotating wave.

The uniqueness follows from the uniqueness of the entire solution.

2.5. Necessity of (1.9). We show that (1.9) is a necessary condition for the existence of counterclockwise periodic rotating waves with average rotating speeds $O(1)$.

(i) Assume $A = \frac{2H\sin\alpha_1+2G\sin\alpha_2}{G^2-H^2}$.

From the proof of Lemma 2.3, one sees that when $\widetilde{H} = H, \widetilde{G} = G, \gamma_1 = \alpha_1, \gamma_2 = \alpha_2$, (2.3) has the solution $(0, \varphi(r; 0, H, G, \alpha_1, \alpha_2))$.

Extend this solution $\varphi(r)$ a little beyond $[H, G]$; suppose its graph contacts $\partial_1\Omega_m$ (resp., $\partial_2\Omega_m$) at a point $P_1' = (H - h(s_1'), s_1')$ for some s_1' (resp., $P_2' = (G + g(s_2'), s_2')$ for some s_2'). Choose $s_i \in (s_i', s_i' + 2\pi\varepsilon)$ $(i = 1, 2)$ such that $h'(s_1) = H\tan\alpha_1$, $g'(s_2) = G\tan\alpha_2$. Denote $P_1 = (H - h(s_1), s_1)$, $P_2 = (G + g(s_2), s_2)$. Then when we move the extended graph of $\varphi(r)$ such that it contacts $\partial_1\Omega_m$ at P_1, the angle ϑ_1 between the graph of $\varphi(r)$ and $\partial_1\Omega_m$ will be $\frac{\pi}{2} + O(\varepsilon)$. A similar conclusion is true at $P_2 \in \partial_2\Omega_m$.

Let us see the shape of $\partial_i\Omega_m$ near P_i $(i = 1, 2)$. First, at P_1, $h'(s_1) = \max_s h'(s) = H\tan\alpha_1$, $h''(s_1) = 0$ and $h'''(s_1) = O(\frac{1}{\varepsilon^2})$. Then for $\Delta s \in (-\varepsilon^{3/2}, \varepsilon^{3/2})$, we have

$$\begin{aligned} h'(s_1 + \Delta s) &= H\tan\alpha_1 + \frac{h'''(s_1)}{2}(\Delta s)^2 + h^{(4)}(s^*)(\Delta s)^3 \\ &= H\tan\alpha_1 + O\left(\left(\frac{\Delta s}{\varepsilon}\right)^2\right) = h'(s_1) + O(\varepsilon). \end{aligned}$$

A similar discussion is valid on $\partial_2\Omega_m$ near P_2.

Since the solution of (2.7) depends on γ_i continuously, we know that when $\gamma_i = \alpha_i - \zeta_i\varepsilon$ ($\zeta_i > 0$ are large), there is a unique $\omega^s > 0, \omega^s = O(\varepsilon)$ and $\varphi^s(r; \omega^s)$ solves (2.7). Moreover, the graph of $\varphi^s(r) + \omega^s t$ contacts $\partial_1\Omega_m$ at points $(H - h(s), s)$ for $s \in (s_1 - \varepsilon^{3/2}, s_1 + \varepsilon^{3/2})$ with angles less than $\frac{\pi}{2}$, and contacts $\partial_2\Omega_m$ at points $(G + g(s), s)$ for $s \in (s_2 - \varepsilon^{3/2}, s_2 + \varepsilon^{3/2})$ with angles less than $\frac{\pi}{2}$.

Suppose the end points of $\varphi^s(r) + \omega^s t$ reach $s_i + \varepsilon^{3/2}$ at time t_i $(i = 1, 2)$ and denote $t_3 := \min\{t_1, t_2\}$; then the discussion in the previous paragraph shows that $\varphi^s(r) + \omega^s t$ is an upper solution on $t \in [0, t_3]$. In this period the speed of $\varphi^s(r) + \omega^s t$ is $\omega^s = O(\varepsilon)$, and the θ-distance of this period is $O(\varepsilon^{3/2})$. Hence the upper solution $\varphi^s(r) + \omega^s t$ uses more time than $C\varepsilon^{1/2}$ (for some $C > 0$) to pass this period. Any rotating wave will be blocked by this kind of upper solution in each period of Ω_m such that the rotating wave passes one period in time greater than $C\varepsilon^{1/2}$. In other words, its average speed is at most $O(\varepsilon^{1/2})$. This is not the case we are interested in here.

(ii) In case $A < \frac{2H\sin\alpha_1+2G\sin\alpha_2}{G^2-H^2}$, one can even find a curve (like $\varphi^s(r)$ above) such that it rotates *clockwise temporarily* in a small period, and it blocks *counterclockwise* rotation.

Consequently, (1.9) is a necessary condition.

2.6. Stability. One has the following stability result by general theory in [19].

THEOREM 2.11. *Let $U(r, t)$ be a counterclockwise periodic rotating wave of* (1.2)–(1.3). *Then $U(r, t)$ is asymptotically stable in the C^1 sense. More precisely, any solution $u(r, t)$ of* (1.2)–(1.3) *whose initial data $u(r, 0)$ is sufficiently close to $U(r, t_0)$ for some $t_0 \in \mathbf{R}$ satisfies*

$$\lim_{t\to\infty} \|U(\cdot, t + t_1) - u(\cdot, t)\|_{C^1} = 0$$

for some $t_1 \in \mathbf{R}$, where $\|\cdot\|_{C^1}$ is understood in the following sense: $U(r, t)$ and $u(r, t)$ correspond to $V(z, t)$ $(z \in (0, 1))$ and $v(z, t)$ $(z \in (0, 1))$, respectively, by using (A.2)

in the appendices. The above limit is understood as

$$\lim_{t\to\infty} \|V(z,\, t+t_1) - v(z,\, t)\|_{C^1} = 0.$$

3. Estimate of average speed: Proof of Theorem 1.2.

3.1. Proof of (i) in Theorem 1.2. Recall that in section 2.1 we construct a *good* lower solution $\widehat{\varphi}(r) + \widehat{\omega}t$, so the average rotating speed ω_m satisfies $\omega_m \geq \widehat{\omega}$. Denote by $(\omega^*, \varphi^*(r))$ the solution of (1.11) (that is, (2.3) with $\widetilde{H} = H,\ \widetilde{G} = G,\ \gamma_1 = \alpha_1,\ \gamma_2 = \alpha_2$). Then it is easy to see by the proofs of Lemmas 2.3 and 2.4 that

$$\omega^* = \widehat{\omega} + O(\varepsilon), \quad \varphi^* = \widehat{\varphi} + C + O(\varepsilon), \quad \varphi^{*\prime} = \widehat{\varphi}\,' + O(\varepsilon). \tag{3.1}$$

Therefore $\omega_m > \omega^* - \frac{C}{m}$ for some $C > 0$.

Also from the proofs of Lemmas 2.3 and 2.4 we have $\widehat{\omega} < \omega_0$ and so $\omega^* < \omega_0$.

This proves the first and the third inequalities of (1.10).

3.1.1. Upper solution. Now we use $\widehat{\varphi}(r) + \widehat{\omega}t$ to construct an upper solution. Let $U(r,t)$ be the periodic rotating wave of (1.2)–(1.3). We note that $U(r,t)\big|_{[H,G]}$ is nothing but the solution of

$$\begin{cases} \tilde{u}_t = \dfrac{\tilde{u}_{rr}}{1+r^2\tilde{u}_r^2} + \dfrac{2\tilde{u}_r + r^2\tilde{u}_r^3}{r(1+r^2\tilde{u}_r^2)} + A\dfrac{\sqrt{1+r^2\tilde{u}_r^2}}{r}, & t>0,\ H<r<G, \\ \tilde{u}(H,t) = U(H,t),\quad \tilde{u}(G,t) = U(G,t), & t>0, \\ \tilde{u}(r,0) = U(r,0), & H<r<G. \end{cases} \tag{3.2}$$

Without loss of generality, we may assume $U(r,0) \leq \widehat{\varphi}(r)$ for $r \in [H,G]$ and $U(r_0,0) = \widehat{\varphi}(r_0)$ for some $r_0 \in [H,G]$. Recall $\varepsilon = 1/m$ and define

$$w(r,t) = E\sqrt{\varepsilon}\,\left(1 - e^{-a^2t}\,\sin(ar+b)\right) + aEF\sqrt{\varepsilon}\,t \quad \text{for } r\in[H,G],\ t\geq 0, \tag{3.3}$$

where $E = O(1)$ is determined later, $a = \frac{\pi}{G-H},\ b = \frac{-\pi H}{G-H}$, and

$$F = \max_{H-h_0\leq r\leq G+g_0} |\,F(r)\,| + 2, \tag{3.4}$$

where

$$F = \frac{2(r\widehat{\varphi}'' + 2\widehat{\varphi}' + r^2(\widehat{\varphi}')^3)r^2\widehat{\varphi}' - 2 - 2r^2(\widehat{\varphi}')^2 - 3r^2(\widehat{\varphi}')^2(1+r^2(\widehat{\varphi}')^2)}{r(1+r^2(\widehat{\varphi}')^2)^2} - \frac{Ar\widehat{\varphi}'}{\sqrt{1+r^2(\widehat{\varphi}')^2}}.$$

It is clear that

$$w - aEF\sqrt{\varepsilon}t \geq 0, \qquad w_t = w_{rr} + aEF\sqrt{\varepsilon} > 0, \qquad \min w(r,0) = 0.$$

LEMMA 3.1. *$\bar{u}(r,t) := w(r,t) + \widehat{\varphi}(r) + \widehat{\omega}t$ is an upper solution of (3.2) on $t\in[0,1]$, and hence*

$$\bar{u}(r,t) \geq U(r,t) \qquad \textit{for } r\in[H,G],\ t\in[0,1]. \tag{3.5}$$

Proof. To prove the lemma, it suffices to show that

$$\bar{u}_t \geq \frac{\bar{u}_{rr}}{1+r^2\bar{u}_r^2} + \frac{2\bar{u}_r + r^2\bar{u}_r^3}{r(1+r^2\bar{u}_r^2)} + A\frac{\sqrt{1+r^2\bar{u}_r^2}}{r} \qquad \text{for } H<r<G,\ t>0, \tag{3.6}$$

and

$$U(H,t) \le \bar{u}(H,t), \qquad t \in [0,1], \tag{3.7}$$

$$U(G,t) \le \bar{u}(G,t), \qquad t \in [0,1]. \tag{3.8}$$

We first prove (3.6). Direct calculation shows that

$$\begin{aligned} &\bar{u}_t - \frac{\bar{u}_{rr}}{1+r^2\bar{u}_r^2} - \frac{2\bar{u}_r + r^2\bar{u}_r^3}{r(1+r^2\bar{u}_r^2)} - A\frac{\sqrt{1+r^2\bar{u}_r^2}}{r} \\ &= w_t - \frac{w_{rr}}{1+r^2(w_r+\widehat{\varphi}')^2} + F_1(r)w_r + O(\varepsilon) \ge 0, \end{aligned}$$

where

$$\begin{aligned} F_1(r) = &\frac{2(r\widehat{\varphi}'' + 2\widehat{\varphi}' + r^2(\widehat{\varphi}')^3)r^2\widehat{\varphi}' - 2 - 2r^2(\widehat{\varphi}')^2 - 3r^2(\widehat{\varphi}')^2(1+r^2(\widehat{\varphi}')^2)}{r(1+r^2(\widehat{\varphi}')^2)(1+r^2(w_r+\widehat{\varphi}')^2)} \\ &- \frac{2Ar\widehat{\varphi}'}{\sqrt{1+r^2(\widehat{\varphi}')^2}+\sqrt{1+r^2(w_r+\widehat{\varphi}')^2}} = F(r) + O(\sqrt{\varepsilon}) \end{aligned}$$

satisfies

$$|F_1(r)| < F - 1 \quad \text{(note that } |w_r| \le aE\sqrt{\varepsilon}\,\text{)}.$$

Next we prove (3.7) and (3.8). Suppose that they hold on $t \in [0,\tau]$ for some $\tau < 1$; we show that they hold in fact on $t \in [0,1]$.

Construct an arc $\theta = \zeta(r)$ as follows (see Figure 3.1). Assume that $h'(s_1) = H\tan\alpha_1$. Denote $P = (H - h(s_1), s_1) \in \partial_1\Omega_m$. Choose $\zeta(r)$ to be the arc with curvature $-A$ that contacts $\partial_1\Omega_m$ perpendicularly at P. Without loss of generality, assume that $\zeta(H) = \theta_0 \in (-2\pi\varepsilon, 2\pi\varepsilon)$. Then P and (H,θ_0) are on $\zeta(r)$.

By (2.8), we know that both $\widehat{\varphi}$ and $\zeta(r)$ have almost the same slope at P:

$$\widehat{\varphi}\,'(H - h(s_1)) - \zeta'(H - h(s_1)) = O(\varepsilon).$$

Hence, there exists $M > 0$ such that

$$|\widehat{\varphi}\,'(H + l\sqrt{\varepsilon}) - \zeta'(H + l\sqrt{\varepsilon})| \le (M-1)\sqrt{\varepsilon} \qquad \text{for any } \ l \in [0,1]. \tag{3.9}$$

Choose $D(\tau) > 0$, such that $\zeta(r) + D(\tau)$ intersects $\bar{u}(r,\tau)$ at $r = H + \sqrt{\varepsilon}$, that is, $\zeta(H+\sqrt{\varepsilon}) + D(\tau) = \bar{u}(H+\sqrt{\varepsilon},\tau)$ (see Figure 3.1). Then by (3.9) we have

$$\begin{aligned} D(\tau) &= w(H+\sqrt{\varepsilon},\tau) + \widehat{\varphi}(H+\sqrt{\varepsilon}) + \widehat{\omega}\tau - \zeta(H+\sqrt{\varepsilon}) \\ &= \bar{u}(H,\tau) + w_r(H,\tau)\sqrt{\varepsilon} + \widehat{\varphi}(H+\sqrt{\varepsilon}) - \widehat{\varphi}(H) - \zeta(H+\sqrt{\varepsilon}) + o(\varepsilon) \\ &< \bar{u}(H,\tau) - aEe^{-a^2}\varepsilon + (M+2\pi)\varepsilon. \end{aligned}$$

Therefore when we choose E satisfying $aEe^{-a^2} > 8\pi + M$ we have $D(\tau) < \bar{u}(H,\tau) - 6\pi\varepsilon$.

Since $\zeta(r)$ contacts $\partial_1\Omega_m$ at P perpendicularly, there exists $\delta \in [0, 2\pi\varepsilon)$ such that the graph of $\zeta(r) + D(\tau) + \delta$ contacts $\partial_1\Omega_m$ perpendicularly and hence $\zeta(r) + D(\tau) + \delta$ is stationary. So

$$U(H+\sqrt{\varepsilon},\tau) \le \bar{u}(H+\sqrt{\varepsilon},\tau) \le \zeta(H+\sqrt{\varepsilon}) + D(\tau) + \delta$$

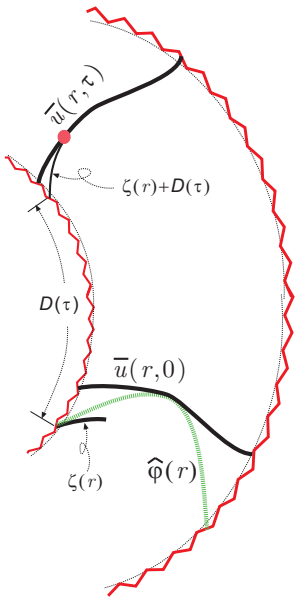


FIG. 3.1. *Upper solution.*

implies that $U(r,\tau) \le \zeta(r) + D(\tau) + \delta$ for $H < r < H + \sqrt{\varepsilon}$. Especially,

$$U(H,\tau) \le \zeta(H) + D(\tau) + \delta \le \bar{u}(H,\tau) - 2\pi\varepsilon.$$

So

$$\bar{u}(H,\tau+t) \ge \bar{u}(H,\tau) \ge U(H,\tau) + 2\pi\varepsilon \ge U(H,\tau+t) \qquad \text{for} \quad t \in [0,T_m].$$

This means that (3.7) holds on $t \in [0,\tau+T_m]$ provided $\tau < 1$. Repeating the above discussion finite times, we obtain (3.7).

Similarly, we can show (3.8) provided that E in the definition of w is chosen large enough. ∎

3.1.2. Proof of the second inequality in (1.10).

$$U(r,1) - \widehat{\varphi}(r) \le \bar{u}(r,1) - \widehat{\varphi}(r) \le (E+aEF)\sqrt{\varepsilon} + \widehat{\omega} \le \left[\frac{(E+aEF)\sqrt{\varepsilon}+\widehat{\omega}}{2\pi\varepsilon} + 1\right] \cdot 2\pi\varepsilon,$$

where $[\cdot]$ is a Gauss function. On the other hand, by the periodicity

$$U\left(r, \left[\frac{(E+aEF)\sqrt{\varepsilon}+\widehat{\omega}}{2\pi\varepsilon} + 1\right] \cdot T_m\right) \le \widehat{\varphi}(r) + \left[\frac{(E+aEF)\sqrt{\varepsilon}+\widehat{\omega}}{2\pi\varepsilon} + 1\right] \cdot 2\pi\varepsilon,$$

where " $=$ " holds at $r = r_0$. Hence time 1 is smaller than time $\left[\frac{(E+aEF)\sqrt{\varepsilon}+\widehat{\omega}}{2\pi\varepsilon}+1\right] \cdot T_m$, and so

$$\omega_m = \frac{2\pi}{mT_m} \le \widehat{\omega} + \frac{E+aEF+1}{\sqrt{m}}. \tag{3.10}$$

Therefore (3.1) implies that

$$\omega_m \le \omega^* + \frac{E+aEF+2}{\sqrt{m}}.$$

This completes the proof of (i) in Theorem 1.2.

3.2. Proof of (ii) in Theorem 1.2. In this subsection, we write the unique periodic rotating wave constructed in Lemma 2.7 by $U_m(r,t)$. Clearly, the period T_m of $U_m(r,t)$ is smaller than 1 when $m > m_0$ for some m_0.

For any given $T > 0$, by Lemma 2.7, U_m $(m = m_0, m_0+1, \ldots)$ are bounded in $C^{2+\mu,1+\frac{\mu}{2}}(\overline{Q_T^m})$, where $\mu = \mu(T)$ is independent of m and $Q_T^m := \{(r,t) | t \in [-T, T+1],\ r \text{ with } (r, U_m(r,t)) \in \Omega_m\}$. So there exists a sequence $\{m_i\}_{i=0}^{\infty}$ and $\mathcal{U}(r,t) \in C^{2,1}([H,G]\times[-T,T+1])$ such that

$$\|U_{m_i} - \mathcal{U}\|_{C^{2,1}([H,G]\times[-T,T+1])} \to 0 \qquad \text{as } \ i \to \infty.$$

Thus, for any $(r,t) \in [H,G]\times[-T,T]$, when $i \to \infty$ we have

$$\omega^* \leftarrow \omega_{m_i} = \frac{2\pi}{m_i T_{m_i}} = \frac{1}{T_{m_i}} \int_t^{T_{m_i}+t} \frac{\partial U_{m_i}(r,t)}{\partial t} dt = \frac{\partial U_{m_i}(r,s)}{\partial t} \to \mathcal{U}_t(r,t)$$

with $s \in (t, T_{m_i}+t)$. This means that $\mathcal{U}(r,t) = \mathcal{U}(r) + \omega^* t$ in $[H,G]\times[-T,T]$ for some $\mathcal{U}(r) \in C^2([H,G])$.

U_{m_i} is the solution of (1.2)–(1.3); taking limit $i \to \infty$ in these equations we have

$$\text{(3.11)} \quad \begin{cases} \omega^* = \dfrac{\mathcal{U}_{rr}}{1+r^2\mathcal{U}_r^2} + \dfrac{2\mathcal{U}_r + r^2\mathcal{U}_r^3}{r(1+r^2\mathcal{U}_r^2)} + A\dfrac{\sqrt{1+r^2\mathcal{U}_r^2}}{r}, & H \le r \le G, \\ \mathcal{U}_r(H) \in \left[\dfrac{-\tan\beta_1}{H},\ \dfrac{\tan\alpha_1}{H}\right], \qquad \mathcal{U}_r(G) \in \left[\dfrac{-\tan\alpha_2}{G},\ \dfrac{\tan\beta_2}{G}\right]. \end{cases}$$

Comparing it with (1.11) we find that when $\mathcal{U}_r(H) < \frac{\tan\alpha_1}{H}$, $\mathcal{U}_r(r) < \varphi_r^*(r;\omega^*)$ and so $\mathcal{U}_r(G) < \frac{-\tan\alpha_2}{G}$. Therefore, (3.11) has a solution if and only if

$$\mathcal{U}_r(H) = \frac{\tan\alpha_1}{H}, \quad \mathcal{U}_r(G) = \frac{-\tan\alpha_2}{G},$$

and the solution $\mathcal{U}(r)$ is nothing but $\varphi^*(r;\omega^*) + C$ for some C. Recall that we require $\max_r U_m(r,0) = 0$ in the proof of Lemma 2.7; hence $\max_r \mathcal{U}(r) = 0$, which implies that $C = -\max_r \varphi^*(r;\omega^*)$.

For any sequence $\{U_{m_j}\}$, there is a subsequence $\{U_{m_{j_k}}\}$ that converges to the same homogenized limit $\varphi^*(r;\omega^*)+\omega^* t - \max_r \varphi^*(r;\omega^*)$ in $C^{2,1}([H,G]\times[-T,T])$ as $k \to \infty$. Consequently, $U_m(r,t) \to \varphi^*(r;\omega^*) + \omega^* t - \max_r \varphi^*(r;\omega^*)$ in $C^{2,1}([H,G]\times[-T,T])$ as $m \to \infty$. This proves (ii) of Theorem 1.2.

Appendix A. Proof of Lemma 2.5.

(i) *of Lemma* 2.5. Set $w = u_r$; then for $t \in (0,T]$ and $\eta_1(t) < r < \eta_2(t)$ we have

$$\text{(A.1)} \quad \begin{cases} w_t = \dfrac{w_{rr}}{1+r^2w^2} + aw_r - \dfrac{2+5r^2w^2+r^4w^4}{r^2(1+r^2w^2)^2}w - \dfrac{A}{r^2\sqrt{1+r^2w^2}}, \\ w(\eta_1(t),t) = h'(u(\eta_1(t),t))/\eta_1^2(t), \quad w(\eta_2,t) = -g'(u(\eta_2(t),t))/\eta_2^2(t), \\ w(r,0) = u_0'(r) \qquad\qquad \text{for } \ r \ \text{ with } \ (r,u_0(r)) \in \Omega_m, \end{cases}$$

where $a = a(r,u,w)$ is a smooth function. By the boundary condition we have

$$-\frac{\tan\beta_1}{H} - \frac{C}{m} < w(\eta_1(t),t) < \frac{\tan\alpha_1}{H} + \frac{C}{m}$$

for some $C > 0$. Hence $|w(\eta_1(t), t)| \le M_1 + \frac{C}{m} < \overline{M} - \frac{1}{2}$. Similarly, $|w(\eta_2(t), t)| \le M_1 + \frac{C}{m} < \overline{M} - \frac{1}{2}$.

On the other hand, if w takes the maximum at (r_1, t) with $r_1 \in (\eta_1(t), \eta_2(t))$ and $w(r_1, t) > \overline{M} - \frac{1}{2}$, then we have

$$w_t(r_1, t) < \frac{w_{rr}(r_1, t)}{1 + r^2 w^2(r_1, t)} + a w_r(r_1, t) \le 0,$$

which implies that w will never be larger than $\overline{M} - \frac{1}{4}$. Similarly, if w takes the minimum at (r_2, t) with $r_2 \in (\eta_1(t), \eta_2(t))$ and $w(r_2, t) < -\overline{M} + \frac{1}{2}$, then, by the definition of M_2 in (1.8), we have

$$w_t(r_2, t) \ge -\frac{2 + 5r^2w^2 + r^4w^4}{r^2(1 + r^2w^2)^2} w - \frac{A}{r^2\sqrt{1 + r^2w^2}}\bigg|_{r=r_2} > 0.$$

Therefore, w will never be smaller than $-\overline{M} + \frac{1}{4}$. Thus we obtain (i) of Lemma 2.5.

We prove (ii) and (iii) of Lemma 2.5 by converting the problem of u into a problem about $v(z, t)$ $(z \in (0, 1))$.

Introduce a new variable

$$z := \frac{r - H + h(\theta)}{J(\theta)}$$

with $J(\theta) = G - H + g(\theta) + h(\theta)$. Then in the new coordinates (z, θ), domain Ω_m is expressed as a domain $\{(z, \theta) \mid z \in (0, 1), \theta \in \mathbf{R}\}$. Now, given a solution $u(r, t)$ of (1.2), we define a new unknown $v(z, t)$ by

$$v(z, t) = u(r(z, t), t), \tag{A.2}$$

where $r(z, t)$ is the inverse function of

$$z(r, t) = \frac{r - H + h(u(r, t))}{J(u(r, t))}.$$

Such an inverse function exists if

$$\frac{\partial z}{\partial r} := \frac{J(u) + [(G + g)h' + (H - h)g' - r(g' + h')]u_r}{J^2(u)} \ne 0,$$

that is,

$$J(u) + [(G + g)h' + (H - h)g' - r(g' + h')]u_r \ne 0. \tag{A.3}$$

We will see later that (A.3) always holds for any solution of (1.2)–(1.3) with appropriate initial data. From $\frac{\partial z}{\partial r} \ne 0$, we have inverse function $r(z, t)$ and

$$\frac{\partial r}{\partial z} = J(v) + (g' + h')z v_z - h' v_z \ne 0.$$

As is easily seen, we have

$$u_r = v_z \frac{\partial z}{\partial r} = \frac{v_z}{J(v)} + \frac{[(G + g)h' + (H - h)g' - r(g' + h')]u_r v_z}{J^2(v)}$$

and hence

$$u_r = \frac{J(v)v_z}{J^2(v) - [(G+g)h' + (H-h)g' - r(g'+h')]v_z} = \frac{v_z}{J(v) + I(z,v)v_z},$$

where $I(z,v) = z(g'(v) + h'(v)) - h'(v)$. Similarly,

$$u_{rr} = \frac{Jv_{zz} - 2(g'+h')v_z^2 - (z(g''+h'') - h'')v_z^3}{(J+Iv_z)^3}, \quad u_t = \frac{Jv_t}{J+Iv_z}$$

since

$$\frac{\partial z}{\partial t} = \frac{h' - z(g'+h')}{J} u_t.$$

Hence the problem about u reduces to a problem about v,

$$v_t = \frac{v_{zz}}{K} + f(z,v,v_z) \qquad \text{for } 0 < z < 1,\ t > 0, \tag{A.4}$$

with

$$f(z,v,p) = \frac{(Jz+H-h)p^3 - 2(g'+h')p^2 - [z(g''+h'') - h'']p^3}{JK} + \frac{2p(J+Ip)^2}{JK(Jz+H-h)}$$

and

$$K(z,v,p) = (J+Ip)^2 + p^2(Jz+H-h)^2.$$

Boundary conditions (1.3) reduce to

$$v_z(0,t) = \frac{Jh'}{(H-h)^2 + (h')^2}, \qquad v_z(1,t) = \frac{-Jg'}{(G+g)^2 + (g')^2}. \tag{A.5}$$

Any solution $u(r,t)$ of (1.2)–(1.3) satisfying (A.3) defines a solution $v(z,t)$ of (A.4)–(A.5) by the relation (A.2). Conversely, if $v(z,t)$ is a solution of (A.4)–(A.5), then the function $u(r,t)$ defined by

$$u(r,t) = v\left(\frac{r - H + h(u)}{J(u)}, t\right) \tag{A.6}$$

is a solution of (1.2)–(1.3). For u to be well defined by (A.6), we need to assume that

$$\frac{\partial}{\partial u}\left(u - v\left(\frac{r - H + h(u)}{J(u)}, t\right)\right) \neq 0$$

or, equivalently,

$$J(v) + [z(g'+h') - h']v_z \neq 0. \tag{A.7}$$

We will see later that any solution of (A.4)–(A.5) with appropriate initial data satisfies (A.7) everywhere.

Using initial data $u_0(r)$ constructed in subsection 2.1.2, one can define a smooth function $v_0(z)$ by (A.2). Since

$$\frac{v_{0z}(z)}{J(v_0(z)) + I(z, v_0(z))v_{0z}(z)} \equiv u_{0r}(r) = O(\varepsilon) \qquad \text{for } z \in [0,1],$$

we have $v_{0z}(z) = O(\varepsilon)$. In what follows we consider problem (A.4)–(A.5) with initial data $v_0(z)$. First, we give an a priori estimate of v_z by (i) of Lemma 2.5.

LEMMA A.1. *Let $v(z,t)$ be a solution of* (A.4)–(A.5) *with initial data v_0 on some time-interval $0 \le t < t_1$. Then*

(i) *there exists $\sigma = \sigma(v_0) > 0$ such that*

$$J(v) + I(z,v)v_z \ge \sigma \qquad \text{for } t \in [0,t_1),\ z \in [0,1];$$

(ii) *there exists $\varrho \in (0,1)$ such that*

$$|v_z(z,t)| \le \frac{G - H + \max(g(s) + h(s))}{\varrho} \cdot \overline{M} \qquad \text{for } t \in [0,t_1),\ z \in [0,1].$$

Proof. (i) Suppose that

$$J(v(z,t)) + I(z,v(z,t))v_z(z,t) \ge \sigma_1 > 0 \qquad \text{for } \ 0 \le t < \tilde{t} < t_1,\ z \in [0,1];$$

then $u(r,t)$ is well defined by (A.6) on $[0,\tilde{t})$ with

$$u_r(r,t) = \frac{v_z(z,t)}{J(v(z,t)) + I(z,v(z,t))v_z(z,t)}.$$

$u(r,t)$ is a solution of (1.2)–(1.3) and satisfies (i) of Lemma 2.5 on $t \in [0,\tilde{t})$. This means that

$$\left| \frac{v_z(z,t)}{J(v(z,t)) + I(z,v(z,t))v_z(z,t)} \right| < \overline{M} \qquad \text{for } t \in [0,\tilde{t}),\ z \in [0,1]. \tag{A.8}$$

If there exists $z_0 \in [0,1]$ such that $J(v(z_0,t)) + I(z_0,v(z_0,t))v_z(z_0,t) \to 0$ as $t \nearrow \tilde{t}$, then (A.8) indicates that $v_z(z_0,t) \to 0$ as $t \nearrow \tilde{t}$. However, this implies that

$$J(v(z_0,t)) + I(z_0,v(z_0,t))v_z(z_0,t) \to J(v(z_0,t)) \ge G - H > 0 \qquad \text{as } t \nearrow \tilde{t},$$

a contradiction. This proves (i).

(ii) From (i), we can define a function $u(r,t)$ by (A.6) on $0 \le t < t_1$, which is a solution of (1.2)–(1.3) and satisfies (i) of Lemma 2.5 on $0 \le t < t_1$.

By (1.6), there exists a $\varrho > 0$ such that $|h'|\overline{M} < 1 - 2\varrho$, $|g'|\overline{M} < 1 - 2\varrho$. Therefore, for large m, (i) of Lemma 2.5 implies that

$$\left| \frac{(G+g)h' + (H-h)g' - r(g'+h')}{G - H + g + h} \cdot u_r \right| < \left| \frac{(G-r)h'}{G-H} - \frac{(r-H)g'}{G-H} \right| \overline{M} + \varrho < 1 - \varrho. \tag{A.9}$$

Hence $|u_r| = |v_z| \cdot \left|\frac{\partial z}{\partial r}\right| > \varrho |v_z| / J(v)$ and then

$$|v_z(z,t)| < \frac{G - H + \max(g(s) + h(s))}{\varrho} \overline{M} \qquad \text{for } z \in [0,1],\ t \in [0,t_1).$$

Remark A.1. In fact, from the proof we know that the left-hand side of (A.3) is positive, that is, $\frac{\partial z}{\partial r} > 0$, so $\frac{\partial r}{\partial z} > 0$.

Next we give the Hölder estimate for v_z by a result in [13].

THEOREM A.2 (see [13, Theorem 13.16]). *Let $\Omega = \omega \times (0,T)$ for some domain $\omega \subset \mathbf{R}^n$ with $\partial\omega \in H_2$. Let $w \in C^{2,1}(\overline{\Omega})$ be a solution of*

$$\text{(A.10)}\qquad \begin{cases} w_t = \operatorname{div} A(x,t,w,Dw) + B(x,t,w,Dw) & \text{in } \ \Omega, \\ A(x,t,w,Dw)\cdot\gamma + \psi(x,t,w) = 0 & \text{on } \ \partial\omega\times(0,T), \\ w = \varphi & \text{on } \ \omega\times\{0\}, \end{cases}$$

with A a C^1 function of (x,t,w,p), ψ a C^1 function of (x,t,w), and A and B uniformly continuous with respect to (x,t,w,p), γ the inner normal to $\partial\omega$. Suppose that there are positive constants L, λ_L, Λ_L, and μ_L such that $|w|+|Dw| \le L$ and

$$\text{(A.11)}\qquad \frac{\partial A^i}{\partial p_j}\xi_i\xi_j \ge \lambda_L|\xi|^2, \qquad |A_p| \le \Lambda_L,$$

$$\text{(A.12)}\qquad |A_w| + |A_x| + |A_t| + |B| \le \mu_L, \qquad |\psi_w| + |\psi_x| + |\psi_t| + |\psi| \le \mu_L$$

for (x,t,w,p) with $|w|+|p| \le L$. Suppose also that

$$\text{(A.13)}\qquad |A(x,t,w,p) - A(x,s,w,p)| + |\psi(x,t,w) - \psi(x,s,w)| \le \mu_L|t-s|^{\frac{1}{2}}$$

for $(x,w,p) \in \partial\omega \times \mathbf{R} \times \mathbf{R}^n$ with $|w|+|p| \le L$ and all $s,t \in (0,T)$. If also $\varphi \in H_2$ and

$$A(x,t,\varphi,D\varphi)\cdot\gamma + \psi(x,t,\varphi) = 0 \quad \text{on } \ \partial\omega\times\{0\},$$

then there are positive constants $\alpha = \alpha(L,n,\lambda_L,\Lambda_L)$ and $C = C(n,\lambda_L,\Lambda_L,\mu_L,\Omega,|\varphi|_2)$ such that $|Du|_\alpha \le C$.

Define

$$A(z,v,p) = \frac{p}{K(z,v,p)} + \int_0^p \frac{pK_p(z,v,p)}{K^2(z,v,p)}dp,$$

$$\mathcal{H}(v) = \frac{J(v)h'(v)}{(H-h(v))^2 + h'(v)^2}, \quad \mathcal{G}(v) = \frac{-J(v)g'(v)}{(G+g(v))^2 + g'(v)^2}.$$

Then (A.4)–(A.5) with initial data $v_0(z)$ can be written in divergence form:

$$\text{(A.14)}\qquad \begin{cases} v_t = \dfrac{\partial}{\partial z}A(z,v,v_z) + B(z,v,v_z), & z \in (0,1), t > 0, \\ v_z(0,t) = \mathcal{H}(v(0,t)), \quad v_z(1,t) = \mathcal{G}(v(1,t)), & t > 0, \\ v(z,0) = v_0(z), & z \in (0,1), \end{cases}$$

where $\frac{\partial A}{\partial z}$ denotes the partial derivative of A on z when we regard A as a function of z and t, and a careful calculation shows that $B(z,v,p)$ is a smooth and bounded function provided v_z is bounded.

LEMMA A.3. *Let $v(z,t)$ be a solution of* (A.14) *on some time-interval $0 \le t \le T$; then there exist $\mu(T) > 0$ and $C > 0$ (independent of T) such that*

$$\text{(A.15)}\qquad |v_z|_\mu \le C \qquad \text{in } \ [0,1]\times[0,T].$$

Proof. First, (2.6) gives the a priori estimate of $u \equiv v$:

$$\widehat{\omega}t - C \le v(z,t) < \widetilde{\omega}t + C \qquad \text{for } z \in [0,1],\ t \in [0,T].$$

As T goes to $+\infty$, v also goes to $+\infty$. However, one important thing should be noticed: in our problem (A.14), v appears always in a form such as $g(v), h(v), g'(v), h'(v)$, etc. Therefore, the unboundedness of v does not cause the unboundedness of the coefficients in (A.14).

Second, the a priori estimate for v_z is given by (ii) in Lemma A.1: $|v_z(z,t)| \le C$ for $z \in [0,1]$ and $t \in [0,T]$.

A trivial and careful calculation shows that Theorem A.2 stated above is applicable to (A.14) and none of the constants λ, Λ, and μ depend on T (since the bound of z and v_z are independent of T, and v appears in the form $g(v), h(v)$, etc.). So there exist $\mu = \mu(T)$ and $C > 0$ (independent of T) such that (A.15) holds. ∎

Proof of (ii) *of Lemma* 2.5. By $u_r = \frac{v_z}{J(v)+I(z,v)v_z}$ and the a priori estimates of v, v_z, and $|v_z|_\mu$ we have $\mu(T) > 0$ and $C > 0$ (independent of T) such that $|u_r|_{\mu(T)} \le C$ for $(r,t) \in \overline{Q_T}$.

Proof of (iii) *of Lemma* 2.5. Choose a constant $\sigma > 0$ such that

$$\sigma > \max\{\max_s |\mathcal{H}'(s)|,\ \max_s |\mathcal{G}'(s)|\} + 1$$

and define $w(z,t) := v(z,t)e^{\sigma z}$. Then w is a solution of

$$\text{(A.16)} \qquad \begin{cases} w_t = \dfrac{\partial}{\partial z}\widetilde{A}(z,w,w_z) + \widetilde{B}(z,w,w_z), & z \in (0,1), t > 0, \\ w_z(0,t) = \mathcal{H}(v(0,t)) + \sigma w(0,t), & t > 0, \\ w_z(1,t) = \mathcal{G}(v(1,t))e^{\sigma} + \sigma w(1,t), & t > 0, \\ w(z,0) = v_0(z)e^{\sigma z}, & z \in (0,1), \end{cases}$$

where $\widetilde{A}$ and $\widetilde{B}$ are smooth functions like A and B. Theorem A.2 stated above can also be used for this problem and so we have $|w_z|_{\mu(T)} \le C$ for $(z,t) \in \overline{Q_T}$, where $\mu(T) > 0$ and $C > 0$ (independent of T).

Especially, at $z = 0$ we have $|w_z(0,t) - w_z(0,s)| \le C|t-s|^{\mu(T)}$. On the other hand,

$$\begin{aligned} |w_z(0,t) - w_z(0,s)| &= |\mathcal{H}(v(0,t)) - \mathcal{H}(v(0,s)) + \sigma w(0,t) - \sigma w(0,s)| \\ &= |\sigma + \mathcal{H}'(\varsigma)| \cdot |w(0,t) - w(0,s)| \ge |w(0,t) - w(0,s)|. \end{aligned}$$

Hence $|w(0,t) - w(0,s)| \le C|t-s|^{\mu(T)}$. Similarly we have $|w(1,t) - w(1,s)| \le C|t-s|^{\mu(T)}$. Finally by Theorem 1.1 of Chapter V in [14] we have $|w|_{\mu(T)} \le C$. This implies that $|v|_{\mu(T)} \le C$ for another $C > 0$. ∎

Appendix B. Proof of Lemma 2.6. It is clear that to prove Lemma 2.6 we only need to prove similar conclusions for v.

LEMMA B.1. *Assume* (1.9) *holds; then* (A.14) *has a unique, global solution* $v(z,t)$ *satisfying* $v_t(\cdot,t) \ge 0$. *Moreover, for any* $T > 0$, *let* $Q_T^z := (0,1) \times (0,T]$; *then there exist positive constants* $\mu(T)$, C_1, *and* C_2 (C_1, C_2 *are independent of* T) *such that*

(i) $v \in C^{2+\mu, 1+\frac{\mu}{2}}(\overline{Q_T^z})$;

(ii) $\|v(z,t)\|_{C^{2+\mu,1+\frac{\mu}{2}}(\overline{Q_T^z})} \le C_1 T + C_2$.

Proof. First, by Theorem 8.1 in [21] and the above a priori estimates for v, we have a unique, global solution v of (A.14). Moreover for any $T > 0$, $v \in C^\infty([0,T], H^\infty(0,1))$, where $H^\infty(0,1) := \bigcap_{k=0}^\infty H^k(0,1)$ is a Fréchet space with norms $(\|\cdot\|_k)_{k=0}^\infty$, and $H^k(0,1)$ is a Sobolev space with norm $\|\cdot\|_k$ (cf. [21]). Embedding theorem implies that we indeed obtain a global solution of (A.14): $v \in C^\infty(\overline{Q_\infty^z})$, where $Q_\infty^z = (0,1)\times(0,\infty)$. (i) is proved.

The result $u_t(\cdot,t) \geq 0$ for $t > 0$ in subsection 2.1.2 implies that $v_t(z,t) \geq 0$.

Now we use the a priori estimates in Lemma 2.5 and the interior estimate (see, for example, Theorem 5 of Chapter 3 in [7]) for problem (A.4)–(A.5) with initial data $v_0(z)$; then for any $T > 0$, there exist $\mu = \mu(T) > 0$ and $C_1 > 0$, $C_2 > 0$ (C_1, C_2 are independent of T) such that

$$\|v\|_{C^{2+\mu,1+\frac{\mu}{2}}(Q_T^z)} \leq C_1 T + C_2.$$

By the smoothness of v on $\overline{Q_T^z}$, we indeed obtain $\|v\|_{C^{2+\mu,1+\frac{\mu}{2}}(\overline{Q_T^z})} \leq C_1 T + C_2$. ∎

Acknowledgments. This work was completed while the author was in RIES of Hokkaido University, Japan. The author would like to thank Professor Yasumasa Nishiura for his hospitality, and Professors Hiroshi Matano and Yuki Kuramoto for helpful discussions. He also thanks the referees for several valuable suggestions.

REFERENCES

[1] S. Angenent, *Parabolic equations for curves on surfaces.* I. *Curves with p-integrable curvature*, Ann. of Math. (2), 132 (1990), pp. 451–483.

[2] S. Angenent, *Parabolic equations for curves on surfaces.* II. *Intersections, blow-up and generalized solutions*, Ann. of Math. (2), 133 (1991), pp. 171–215.

[3] W. K. Burton, N. Cabrera, and F. C. Frank, *The growth of crystals and equilibrium structure of their surfaces*, Philos. Trans. Roy. Soc. London. Ser. A., 243 (1951), pp. 299–358.

[4] D. Cioranescu and J. Saint Jean Paulin, *Homogenization of Reticulated Structures*, Appl. Math. Sci., 136, Springer-Verlag, New York, 1999.

[5] K.-S. Chou and X.-P. Zhu, *The Curve Shorting Problem*, Chapman & Hall/CRC, Boca Raton, FL, 2001.

[6] E. Dulos, J. Boissonade, and P. De Kepper, *Dynamics and morphology of sustained Two-dimensional wavetrains*, Phys. A, 188 (1992), pp. 120–131.

[7] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[8] M. Gage and R. S. Hamilton, *The heat equation shrinking convex plane curves*, J. Differential Geom., 23 (1986), pp. 69–96.

[9] Y. Giga, N. Ishimura, and Y. Kohsaka, *Spiral solutions for a weakly anisotropic curvature flow equation*, Adv. Math. Sci. Appl., 12 (2002), pp. 393–408.

[10] M. A. Grayson, *The heat equation shrinks embedded plane curves to round points*, J. Differential Geom., 26 (1987), pp. 285–314.

[11] J. P. Keener, *A geometrical theory for spiral waves in excitable media*, SIAM J. Appl. Math., 46 (1986), pp. 1039–1056.

[12] J. P. Keener and J. J. Tyson, *Spiral waves in the Belousov-Zhabotinskii reaction*, Phys. D, 21 (1986), pp. 307–324.

[13] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.

[14] O. A. Ladyzhenskia, V. A. Solonnikov, and N. N. Uraltseva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[15] B. Lou, H. Matano, and K. Nakamura, *Periodic Traveling Waves of a Mean Curvature Flow in a Band Domain with Periodically Undulating Boundaries*, preprint.

[16] H. Matano, *Traveling waves in spatially inhomogeneous diffusive media with bistable nonlinearity* I, Discrete Contin. Dyn. Syst., submitted.

[17] W. W. Mullins, *Two-dimensional motion of idealized grain boundaries*, J. Appl. Phys., 27 (1956), pp. 900–904.

[18] Z. Noszticzius, W. Horsthemke, W. D. McCormick, et al., *Sustained spiral waves in an annulus gel reactor: A chemical pinwheel*, Nature, 329 (1987), pp. 619–620.
[19] T. Ogiwara and H. Matano, *Monotonicity and convergence results in order-preserving systems in the presence of symmetry*, Discrete Contin. Dynam. Systems, 5 (1999), pp. 1–34.
[20] T. Ogiwara and K.-I. Nakamura, *Spiral traveling wave solutions of nonlinear diffusion equations related to a model of spiral crystal growth*, Publ. Res. Inst. Math. Sci., 39 (2003), pp. 767–783.
[21] M. Poppenberg, *Nash-Moser techniques for nonlinear boundary-value problems*, Electronic J. Differential Equations, 2003 (54) (2003), pp. 1–34.
[22] J. J. Tyson and J. P. Keener, *Singular perturbation theory of traveling waves in excitable media*, Phys. D, 32 (1988), pp. 327–361.
[23] J. Xin, *Front propagation in heterogeneous media*, SIAM Rev., 42 (2000), pp. 161–230.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 717–740

PACKET FLOW ON TELECOMMUNICATION NETWORKS*

CIRO D'APICE†, ROSANNA MANZO†, AND BENEDETTO PICCOLI‡

Abstract. The aim of this paper is to introduce a macroscopic fluid dynamic model dealing with the flow of information on a telecommunication network encoded in packets. Taking an intermediate time and space scale, we propose a model similar to that introduced recently for car traffic; see [G. M. Coclite, M. Garavello, and B. Piccoli, *SIAM J. Math. Anal.*, 36 (2005), pp. 1862–1886]. For dynamics at nodes we consider two "routing algorithms" and prove existence of solutions to Cauchy problems. The main difference among the two algorithms is the possibility of redirecting packets of the second algorithm, which in turn implies stability, i.e., Lipschitz continuous dependence on initial data, not granted for solutions using the first algorithm.

Key words. traffic flow on networks, conservation laws, fluid dynamic models

AMS subject classifications. 35L65, 35L67, 90B20

DOI. 10.1137/050631628

1. Introduction.

The aim of this paper is to introduce a macroscopic fluid dynamic model dealing with the flow of information on a telecommunication network encoded in packets. There are some recent works on traffic flow on road networks (see [10, 11, 14, 15, 16, 18]) that are based on macroscopic description via car densities and other conserved quantities. Our idea is to look at the network at an intermediate time scale so that packet transmission happens at a faster level but the equilibria of the whole network are reached only as asymptotic. This permits us to construct a model again relying on macroscopic description.

There exist various approaches to traffic flow on telecommunication networks, in particular for the Internet and with special focus on properties of control congestion algorithms such as TCP/IP; see, for example, [4, 17, 25]. Our idea is rather to take a large number of nodes, which use some simple routing algorithm, and via some limiting procedure obtain a partial differential equation for the packet density on the network. First we focus on a straight transmission line and justify the limiting procedure. Then we consider a network and introduce two routing algorithms for nodes with many entering and exiting lines. Let us start with the basic assumptions.

A network is formed by a finite collection of transmission lines and nodes (or routers). We assume that each node receives and sends information encoded in packets. Each packet can thus be seen as a particle on the network, but we have to take into account specific issues of telecommunications. Having in mind the Internet as our key model, we assume the following:

(1) Each packet travels on the network with a fixed speed and with assigned final destination.
(2) Nodes receive, process, and then forward packets. Packets may be lost with a probability increasing with the number of packets to be processed. Each lost packet is sent again.

*Received by the editors May 16, 2005; accepted for publication (in revised form) February 13, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/sima/38-3/63162.html
†Department of Information Engineering and Applied Mathematics, University of Salerno, Fisciano (SA), Italy (dapice@diima.unisa.it, manzo@diima.unisa.it).
‡Istituto per le Applicazioni del Calcolo "Mauro Picone," Consiglio Nazionale delle Ricerche, Roma, Italy (b.piccoli@iac.cnr.it).

We first model the behavior of a single straight transmission line on which there are some consecutive nodes. Each node sends packets to the following one a first time, then packets which are lost in this process are sent a second time, and so on. The important point is that each packet is sent until it reaches the next node; thus, looking from a macroscopic level, it is assumed that packets are conserved. This leads from the microscopic dynamics to the simple model consisting of a single conservation law:

$$\rho_t + f(\rho)_x = 0, \tag{1}$$

where ρ is the packet density, v is the velocity, and $f(\rho) = v\rho$ is the flux. Since the packet transmission velocity on the line is assumed constant, we can derive an average transmission velocity among nodes considering the amount of packets that may be lost. More precisely, assigning a loss probability as function of the density, it is possible to compute a velocity function and thus a flux function.

The conclusion is rigorously justified only for constant density, but is assumed to hold in general. This corresponds to the hypotheses that macroscopic density waves move at a velocity much smaller than the packet transmission velocity. In section 2 we derive some models and then focus the rest of the paper on a particular model that implies equivalence between the total variation of density and of flux. Even if our limiting procedure is not completely rigorous, there are other approaches, as that of [3] for supply chains, which lead to conservation laws. Moreover, since our method to solve problems at nodes is based only on flux values, every limiting procedure, which leads to a conservation law formulation, may be used to treat the problem on a network.

The aim is then to consider complex networks; thus we need to introduce a way of solving dynamics at nodes in which many lines intersect. For this, respecting rule 2, we propose two different routing algorithms:

(RA1) Packets from incoming lines are sent to outgoing ones according to their final destination (without taking into account possible high loads of outgoing lines).

(RA2) Packets are sent to outgoing lines in order to maximize the flux through the node.

The main differences of the two algorithms are the following. The first one simply sends each packet to the outgoing line which is naturally chosen according to the final destination of the packet itself. The algorithm is blind to possible overloads of some outgoing lines and, by some abuse of notation, is similar to the behavior of a "switch." The second algorithm, on the contrary, sends packets to outgoing lines taking into account the loads, and thus possibly redirecting packets. Again by some abuse of notation, this is similar to a "router" behavior.

Routing algorithm (RA1) can be described by two rules and was already used in [11] for car traffic. In particular a traffic distribution matrix A is given, which describes the percentage of packets from an incoming line that are addressed to an outgoing one. For existence of solutions to the Cauchy problem on the network, we have to restrict ourselves to the case of simple nodes with two incoming and two outgoing lines, but, differently from [11], we can obtain a precise bound on the total variation of density, thanks to the assumption on the flux function, and then derive existence of solutions to the Cauchy problem more directly by wave-front tracking. However, Lipschitz continuous dependence of solutions is not granted.

Then we analyze routing algorithm (RA2). Notice that this second algorithm

was not considered for car traffic, because redirection of cars is not expected from a modeling point of view (except special situations, as closure of a road).

In order to determine unique solutions to Riemann problems, some additional parameters are introduced, called, respectively, priority parameters and traffic distribution parameters. The former describe priorities among incoming lines, while the latter have the same meaning of the traffic distribution matrix.

The advantage of this second algorithm is that the flux variation at a node is conserved for interaction of waves from transmission lines. This permits us both to obtain estimates on the total variation of density, thus to construct solutions again by wave-front tracking, and also to obtain uniqueness and Lipschitz continuous dependence of solutions. The latter result is achieved by the method introduced in [6, 8], which considers a Riemannian-type metric on L^1. More precisely, the distance among solutions is measured by paths in L^1 which admit some generalized tangent vectors. The key point is that the norms of tangent vectors are known to decrease inside each line (i.e., for scalar conservation laws), while for interactions with nodes its evolution is determined by flux variation. As explained in section 5.2.1, other known methods to treat uniqueness for scalar conservation laws seem not to work for the network case.

The obtained results show the strong effect of the routing algorithm. More precisely, the choice of a "router"-type algorithm, i.e., (RA2), implies stability of solutions, with respect of perturbation of the data, as opposed to the instability obtained with the "switch" types.

The paper is organized as follows. Section 2 describes the dynamics of packet density on a single transmission line. Section 3 gives general definitions of a network and of the Riemann solver. Then we describe the two routing algorithms in section 4, giving explicit unique solutions to Riemann problems. Finally, section 5 provides the needed estimates for constructing solutions to Cauchy problems and for obtaining continuous dependence for the second algorithm.

2. Packet loss and velocity functions on transmission lines. We model a transmission line by a sequence of nodes N_k, representing routers, and edges which connect consecutive nodes. Thus the transmission line is represented by a real interval I union of many edges and nodes.

Each node (router) sends and receives packets. Following rule 1, we assume that packets flow at constant velocity from each node N_k to N_{k+1}. Taking a discrete time scale for the evolution, the state at time t_i is described by the packet quantities $R_k(t_i)$ on nodes N_k, and transmission happens among consecutive nodes between two discrete times. Therefore, to determine the dynamics on I we need to describe the effect of packet loss on the velocity of transmission function.

As for the Internet, we assume that each node N_k sends again packets that are lost by the following node N_{k+1}. Therefore the number of packets is conserved, i.e., at macroscopic level we expect (1) to hold. More precisely, we assume that there exists a function $p : [0, R_{max}] \to [0, 1]$ which assigns the packet loss probability as a function of the number of packets.

Let us focus now on two consecutive nodes and introduce some notation. Suppose that δ is the distance between nodes N_k and N_{k+1}. Let Δt_0 be the transmission time of packets from node N_k to node N_{k+1} if they are sent with success at the first attempt, and let Δt_{av} be the average transmission time when some packets are lost by N_{k+1}. Finally, we denote by $\bar{v} = \frac{\delta}{\Delta t_0}$ and $v = \frac{\delta}{\Delta t_{av}}$ the packet velocity in the two cases.

At the first attempt, the packets sent by node N_k reach with success node N_{k+1}

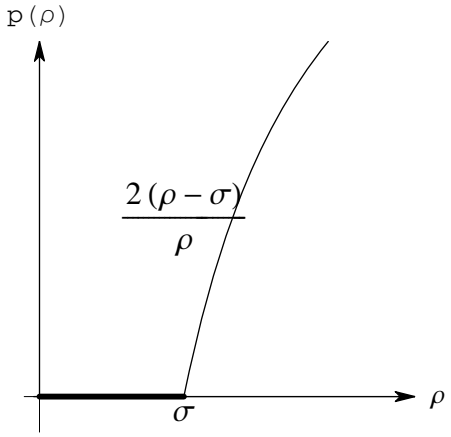


FIG. 1. *Packet loss function.*

with probability $(1-p)$ and are lost by node N_{k+1} with probability p. At the second attempt there are p of the total number of packets left to be sent again and $(1-p)p$ are sent with success while p^2 are lost. Going on at the nth attempt, $(1-p)p^{n-1}$ packets are sent successfully and p^n are lost. The average transmission time is equal to

$$\Delta t_{av} = \sum_{n=1}^{+\infty} n\Delta t_0 (1-p)p^{n-1} = \frac{\Delta t_0}{1-p}, \tag{2}$$

from which we get that the transmission velocity is given by

$$v = \frac{\delta}{\Delta t_{av}} = \frac{\delta}{\Delta t_0}(1-p) = \bar{v}(1-p). \tag{3}$$

The above reasoning works for the entire line if $R_k(t_0) = R$ for all k. In fact, one gets immediately that $R_k(t_i) = R$ for all i and k. Thus the following holds.

LEMMA 1. *Assume that $R_k(t_0) = R$ for all k. Then the average transmission time and velocity are given by* (2) *and* (3).

Clearly Lemma 1 gives an average velocity only if the density is constant. However, we assume the conclusion holds in general for the macroscopic velocity and use this together with (1). This assumption is not completely justified, but it is reasonable if the transmission velocity of packets is expected to be much bigger than the macroscopic velocity.

We may also assign the loss probability directly as a function of the packet density; then the corresponding flux is easily determined. Such loss probability should vanish for low load levels of nodes and reach the value 1 for $R = R_{max}$. We show some choice of packet loss functions and the corresponding macroscopic fluxes.

Example 2. Let us suppose that the packet loss probability is given by

$$p(\rho) = \begin{cases} 0, & 0 \le \rho \le \sigma, \\ \frac{2(\rho-\sigma)}{\rho}, & \sigma \le \rho \le \rho_{\max}, \end{cases}$$

for some $\sigma \in]0, \rho_{\max}[$; see Figure 1. Then the average transmission velocity is equal to

$$v(\rho) = \bar{v}(1-p(\rho)) = \begin{cases} \bar{v}, & 0 \le \rho \le \sigma, \\ \bar{v}\frac{(2\sigma-\rho)}{\rho}, & \sigma \le \rho \le \rho_{\max}. \end{cases}$$

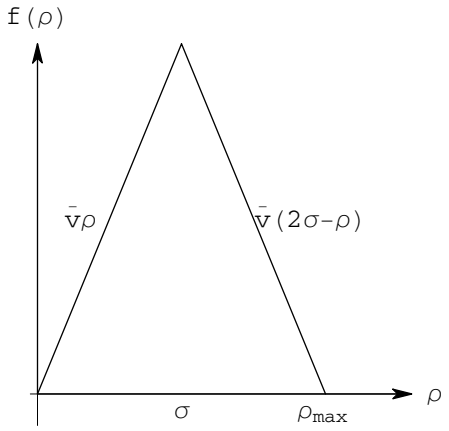


FIG. 2. *Flux function.*

Imposing that

$$v(\rho_{\max}) = \bar{v}\frac{(2\sigma - \rho_{\max})}{\rho_{\max}} = 0,$$

we get that $\sigma = \frac{\rho_{\max}}{2}$. Since $f(\rho) = v(\rho)\rho$ it follows that (see Figure 2)

$$f(\rho) = \begin{cases} \bar{v}\rho, & 0 \le \rho \le \sigma, \\ \bar{v}(2\sigma - \rho), & \sigma \le \rho \le \rho_{\max}. \end{cases}$$

The fundamental diagram (i.e., the expression of the flux as a function of the density) of Example 2 was extensively used in traffic flow literature (see [13, 20]) and is sometimes called the Daganzo–Newell flux.

Example 3. Suppose that

$$p(\rho) = \begin{cases} 0, & 0 \le \rho \le \sigma, \\ \frac{\rho - \sigma}{\sigma}, & \sigma \le \rho \le \rho_{\max}. \end{cases}$$

It follows that

$$v(\rho) = \begin{cases} \bar{v}, & 0 \le \rho \le \sigma, \\ \frac{\bar{v}(2\sigma - \rho)}{\sigma}, & \sigma \le \rho \le \rho_{\max}, \end{cases}$$

and

$$f(\rho) = \begin{cases} \bar{v}\rho, & 0 \le \rho \le \sigma, \\ \frac{\bar{v}\rho(2\sigma - \rho)}{\sigma}, & \sigma \le \rho \le \rho_{\max}. \end{cases}$$

Example 4. Suppose that

$$p(\rho) = \begin{cases} 0, & 0 \le \rho \le \sigma, \\ \frac{(\rho - \sigma)^2}{\sigma^2}, & \sigma \le \rho \le \rho_{\max}. \end{cases}$$

It follows that

$$v(\rho) = \begin{cases} \bar{v}, & 0 \le \rho \le \sigma, \\ \frac{\bar{v}\rho(2\sigma - \rho)}{\sigma^2}, & \sigma \le \rho \le \rho_{\max}, \end{cases}$$

and

$$f(\rho) = \begin{cases} \bar{v}\rho, & 0 \le \rho \le \sigma, \\ \frac{\bar{v}\rho^2(2\sigma-\rho)}{\sigma^2}, & \sigma \le \rho \le \rho_{\max}. \end{cases}$$

Remark 5. Examples 2 and 3 lead to fluxes which are not $\mathcal{C}^1$; the opposite happens for Example 4. Notice that only for Example 2 does the corresponding flux have the property that $f'(\rho\pm) \neq 0$ for every ρ. Thus the density variation along discontinuities not crossing σ is equivalent to the flux ones.

In what follows we suppose that measures on packet loss probability lead to the formulation of Example 2. This allows us to control the variation of the density function in terms of the variation of the flux function, as shown later.

We can suppose for simplicity that $\rho_{\max} = 1$, so we have the following assumptions on the flux:

(F) $f : [0,1] \to R$, $\quad f(\rho) = \begin{cases} \bar{v}\rho, & 0 \le \rho \le \sigma, \\ \bar{v}(2\sigma - \rho), & \sigma \le \rho \le 1, \end{cases}$
$f(0) = f(1) = 0$. Thus $\sigma = \frac{1}{2}$ is the unique maximum point.

3. Telecommunication networks. We consider a telecommunication network that is modeled by a finite set of intervals $I_i = [a_i, b_i] \subset \mathbb{R}$, $i = 1, \dots, N$, $a_i < b_i$, possibly with either $a_i = -\infty$ or $b_i = +\infty$, on which we consider the model of the previous section, i.e., equation (1) with assumption (F). The network evolution is described by a finite set of functions ρ_i defined on $[0, +\infty[\times I_i$.

On each transmission line I_i we want ρ_i to be a weak entropic solution of (1), that is, for every function $\varphi : [0, +\infty[\times I_i \to \mathbb{R}$ smooth, positive with compact support on $]0, +\infty[\times]a_i, b_i[$,

$$\int_0^{+\infty} \int_{a_i}^{b_i} \left(\rho_i \frac{\partial \varphi}{\partial t} + f(\rho_i) \frac{\partial \varphi}{\partial x} \right) dxdt = 0, \tag{4}$$

and for every $k \in \mathbb{R}$ and every $\tilde{\varphi} : [0, +\infty[\times I_i \to \mathbb{R}$ smooth, positive with compact support on $]0, +\infty[\times]a_i, b_i[$,

$$\int_0^{+\infty} \int_{a_i}^{b_i} \left(|\rho_i - k| \frac{\partial \tilde{\varphi}}{\partial t} + sgn(\rho_i - k)\left(f(\rho_i) - f(k)\right) \frac{\partial \tilde{\varphi}}{\partial x} \right) dxdt \ge 0. \tag{5}$$

It is well known that, for (1) on $\mathbb{R}$ and for every initial data in L^∞, there exists a unique weak entropic solution depending in a continuous way on the initial data in L^1_{loc}. Moreover, for initial data in $L^\infty \cap L^1$ we have Lipschitz continuous dependence in L^1.

We assume that the transmission lines are connected by some junctions. Each junction J is given by a finite number of incoming transmission lines and a finite number of outgoing transmission lines; thus we identify J with $((i_1, \dots, i_n), (j_1, \dots, j_m))$, where the first n-tuple indicates the set of incoming transmission lines and the second m-tuple indicates the set of outgoing transmission lines. Each transmission line can be incoming for at most one junction and outgoing for at most one junction. Hence the complete model is given by a couple $(\mathcal{I}, \mathcal{J})$, where $\mathcal{I} = \{I_i : i = 1, \dots, N\}$ is the collection of transmission lines and $\mathcal{J}$ is the collection of junctions.

Now we discuss how to define solutions at junctions. For this, fix a junction J with n incoming transmission lines, say $I_1, \dots, I_n$, and m outgoing transmission lines, say

$I_{n+1}, \dots, I_{n+m}$. A weak solution at J is a collection of functions $\rho_l : [0, +\infty[\times I_l \to \mathbb{R}$, $l = 1, \dots, n+m$, such that

$$\sum_{l=1}^{n+m} \left(\int_0^{+\infty} \int_{a_l}^{b_l} \left(\rho_l \frac{\partial \varphi_l}{\partial t} + f(\rho_l) \frac{\partial \varphi_l}{\partial x} \right) dx dt \right) = 0 \tag{6}$$

for every φ_l, $l = 1, \dots, n+m$, smooth having compact support in $]0, +\infty[\times]a_l, b_l]$ for $l = 1, \dots, n$ (incoming transmission lines) and in $]0, +\infty[\times [a_l, b_l[$ for $l = n+1, \dots, n+m$ (outgoing transmission lines), which are also smooth across the junction, i.e.,

$$\varphi_i(\cdot, b_i) = \varphi_j(\cdot, a_j), \quad \frac{\partial \varphi_i}{\partial x}(\cdot, b_i) = \frac{\partial \varphi_j}{\partial x}(\cdot, a_j), \quad i = 1, \dots, n, j = n+1, \dots, n+m.$$

Remark 6. Let $\rho = (\rho_1, \dots, \rho_{n+m})$ be a weak solution at the junction such that each $x \to \rho_i(t, x)$ has bounded variation. We can deduce that ρ satisfies the Rankine–Hugoniot condition at the junction J, namely,

$$\sum_{i=1}^{n} f(\rho_i(t, b_i-)) = \sum_{j=n+1}^{n+m} f(\rho_j(t, a_j+)) \tag{7}$$

for almost every $t > 0$.

For a scalar conservation law a Riemann problem is a Cauchy problem for an initial data of Heaviside type that is piecewise constant with only one discontinuity. One looks for centered solutions, i.e., $\rho(t, x) = \phi(\frac{x}{t})$ formed by simple waves, which are the building blocks for constructing solutions to the Cauchy problem via wave-front tracking algorithm. These solutions are formed by continuous waves called rarefactions and by traveling discontinuities called shocks. The speed of waves are related to the values of f'; see [7].

Analogously, we call the Riemann problem for a junction the Cauchy problem corresponding to an initial data which is constant on each transmission line.

DEFINITION 7. *A Riemann solver for the junction J is a map $RS : [0,1]^n \times [0,1]^m \to [0,1]^n \times [0,1]^m$ that associates to Riemann data $\rho_0 = (\rho_{1,0}, \dots, \rho_{n+m,0})$ at J a vector $\hat{\rho} = (\hat{\rho}_1, \dots, \hat{\rho}_{n+m})$ so that the solution on an incoming transmission line I_i, $i = 1, \dots, n$, is given by the wave $(\rho_{i,0}, \hat{\rho}_i)$ and on an outgoing one I_j, $j = n+1, \dots, n+m$, is given by the wave $(\hat{\rho}_j, \rho_{j,0})$. We require the following consistency condition:*

(CC) $RS(RS(\rho_0)) = RS(\rho_0)$.

Remark 8. The condition (CC) is necessary for providing a good definition of Riemann solver and thus also for uniqueness.

Assume, for example, that $RS(\rho) = \rho'$ and $RS(\rho') = \rho$ for some Riemann data $\rho \neq \rho'$. To solve the Riemann problem with datum ρ, one should use the boundary datum ρ' at the junction. In turn, when ρ' starts propagating into lines, one should go back to ρ, and so on and so forth. A solution would thus not exist.

The same kind of problem happens for uniqueness.

Once a Riemann solver is assigned we can define admissible solutions at J.

DEFINITION 9. *Assume a Riemann solver RS is assigned. Let $\rho = (\rho_1, \dots, \rho_{n+m})$ be such that $\rho_i(t, \cdot)$ is of bounded variation for every $t \geq 0$. Then ρ is an admissible weak solution of* (1) *related to RS at the junction J if and only if the following properties hold:*

(i) *ρ is a weak solution at junction J;*

(ii) *for almost every* t *setting*

$$\rho_J(t) = (\rho_1(\cdot, b_1-), \dots, \rho_n(\cdot, b_n-), \rho_{n+1}(\cdot, a_{n+1}+), \dots, \rho_{n+m}(\cdot, a_{n+m}+))$$

we have

$$RS(\rho_J(t)) = \rho_J(t).$$

For every transmission line $I_i = [a_i, b_i]$, if $a_i > -\infty$ and I_i is not the outgoing transmission line of any junction, or $b_i < +\infty$ and I_i is not the incoming transmission line of any junction, then a boundary datum $\psi_i : [0, +\infty[\to \mathbb{R}$ is given. We ask ρ_i to satisfy $\rho_i(t, a_i) = \psi_i(t)$ (or $\rho_i(t, b_i) = \psi_i(t)$) in the sense of [5]. The treatment of boundary data in the sense of [5] can be done as in [1, 2], and thus only the case without boundary data is considered. All the stated results hold also for the case with boundary data with obvious modifications.

Our aim is to solve the Cauchy problem on $[0, +\infty[$ for given initial and boundary data as in the next definition.

DEFINITION 10. *Given* $\bar\rho_i : I_i \to [0, 1]$, $i = 1, \dots, N$, *measurable functions, a collection of functions* $\rho = (\rho_1, \dots, \rho_N)$, *with* $\rho_i : [0, +\infty[\times I_i \to [0, 1]$ *continuous as functions from* $[0, +\infty[$ *into* L^1_{loc}, *is an admissible solution to the Cauchy problem on the network if* ρ_i *is a weak entropic solution to* (1) *on* I_i, $\rho_i(0, x) = \bar\rho_i(x)$ *a.e., and at each junction* ρ *is a weak solution and is an admissible weak solution in case of bounded variation.*

Remark 11. It is possible to generalize all definitions and results of upcoming sections to the case of different fluxes f_i for each line I_i. In fact, all statements are in terms of values of fluxes at junctions; thus it is sufficient that the ranges of fluxes intersect.

4. Riemann solvers at junctions. In this section we describe two different Riemann solvers at a junction that represent two different routing algorithms:

(RA1) We assume that
  - (A) the traffic from incoming transmission lines is distributed on outgoing transmission lines according to fixed coefficients;
  - (B) respecting (A), the router chooses to send packets in order to maximize fluxes (i.e., the number of packets which are processed).

(RA2) We assume that the number of packets through the junction is maximized over both incoming and outgoing lines.

Once solutions to Riemann problems are given, one can use a wave-front tracking algorithm to construct a sequence of approximate solutions. To pass to the limit one has to bound the number of waves and the BV norm of approximate solutions; see [7, 11]. In the next section we prove a BV bound on the density for the case of junctions with two incoming and two outgoing transmission lines, for both the routing algorithms.

4.1. Riemann solver for algorithm (RA1). The Riemann solver for algorithm (RA1) has already been described in [10, 11], where traffic problems for road networks have been analyzed, using different assumptions on the flux function.

Consider a junction J in which there are n transmission lines with incoming traffic and m transmission lines with outgoing traffic. To deal with (A) we fix a traffic distribution matrix $A \doteq \{\alpha_{ji}\}_{j=n+1,\dots,n+m, i=1,\dots,n} \in \mathbb{R}^{m\times n}$ such that

$$0 < \alpha_{ji} < 1, \quad \sum_{j=n+1}^{n+m} \alpha_{ji} = 1$$

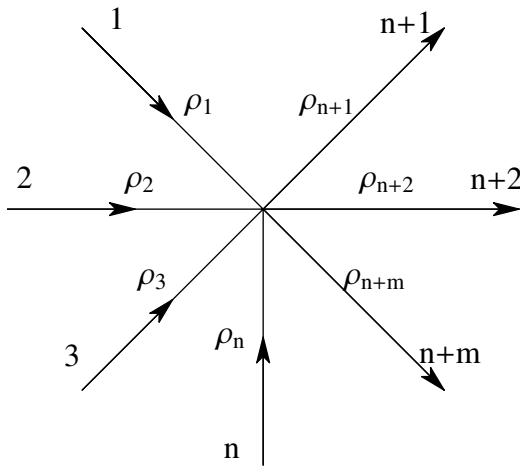


FIG. 3. *A junction.*

for each $i = 1, \dots, n$ and $j = n+1, \dots, n+m$, where α_{ji} is the percentage of packets arriving from the ith incoming transmission line that take the jth outgoing transmission line.

For simplicity we indicate by

$$(t,x) \in \mathbb{R}_+ \times I_i \to \rho_i(t,x) \in [0,1], \quad i = 1, \dots, n,$$

the densities of the packets on the transmission lines with incoming traffic and by

$$(t,x) \in \mathbb{R}_+ \times I_j \to \rho_j(t,x) \in [0,1], \quad j = n+1, \dots, n+m,$$

those on transmission lines with outgoing traffic; see Figure 3.

We need some more notation.

DEFINITION 12. *Let $\tau : [0,1] \to [0,1]$ be the map such that*

1. *$f(\tau(\rho)) = f(\rho)$ for every $\rho \in [0,1]$;*
2. *$\tau(\rho) \neq \rho$ for every $\rho \in [0,1] \backslash \{\sigma\}$.*

Clearly, τ is well defined and satisfies

$$0 \le \rho \le \sigma \Leftrightarrow \sigma \le \tau(\rho) \le 1,$$
$$\sigma \le \rho \le 1 \Leftrightarrow 0 \le \tau(\rho) \le \sigma.$$

To state the main result of this section we need some assumption on the matrix A (satisfied under generic conditions for $m = n$). Let $\{e_1, \dots, e_n\}$ be the canonical basis of $\mathbb{R}^n$ and for every subset $V \subset \mathbb{R}^n$ indicate by $V^\perp$ its orthogonal. Define for every $i = 1, \dots, n$, $H_i = \{e_i\}^\perp$, i.e., the coordinate hyperplane orthogonal to e_i, and for every $j = n+1, \dots, n+m$, let $\alpha_j = \{\alpha_{j1}, \dots, \alpha_{jn}\} \in \mathbb{R}^n$ and define $H_j = \{\alpha_j\}^\perp$. Let $\mathcal{K}$ be the set of indices $k = (k_1, \dots, k_l)$, $1 \le l \le n-1$, such that $0 \le k_1 < k_2 < \cdots < k_l \le n+m$ and for every $k \in \mathcal{K}$ set $H_k = \bigcap_{h=1}^{l} H_h$. Letting $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$, we assume

(C) for every $k \in \mathcal{K}$, $\mathbf{1} \notin H_k^\perp$.

THEOREM 13 (Theorem 3.1 of [11]). *Consider a junction J; assume that the flux $f : [0,1] \to \mathbb{R}$ satisfies* (F) *and the matrix A satisfies condition* (C). *For every*

$\rho_{1,0},\dots,\rho_{n+m,0} \in [0,1]$, there exists a unique admissible centered weak solution, in the sense of Definition 9; $\rho = (\rho_1,\dots,\rho_{n+m})$ of (1) at the junction J such that

$$\rho_1(0,\cdot) \equiv \rho_{1,0},\dots,\rho_{n+m}(0,\cdot) \equiv \rho_{n+m,0}.$$

Moreover, there exists a unique $(n+m)$-tuple $(\hat\rho_1,\dots,\hat\rho_{n+m}) \in [0,1]^{n+m}$ such that

$$\hat\rho_i \in \begin{cases} \{\rho_{i,0}\} \cup]\tau(\rho_{i,0}),1] & \text{if } 0 \le \rho_{i,0} \le \sigma, \\ [\sigma,1] & \text{if } \sigma \le \rho_{i,0} \le 1, \end{cases} \quad i = 1,\dots,n, \tag{8}$$

and

$$\hat\rho_j \in \begin{cases} [0,\sigma] & \text{if } 0 \le \rho_{j,0} \le \sigma, \\ \{\rho_{j,0}\} \cup [0,\tau(\rho_{j,0}))[& \text{if } \sigma \le \rho_{j,0} \le 1, \end{cases} \quad j = n+1,\dots,n+m, \tag{9}$$

and on each incoming line I_i, $i = 1,\dots,n$, the solution consists of the single wave $(\rho_{i,0},\hat\rho_i)$, while on each outgoing line I_j, $j = n+1,\dots,n+m$, the solution consists of the single wave $(\hat\rho_j,\rho_{j,0})$.

Condition (C) on A cannot hold for crossings with two incoming and one outgoing transmission lines. Following [10], it is possible to introduce a further parameter whose meaning is the following. When the number of packets is too big to let all of them go through the crossing, there is a priority rule that describes the percentage of packets, going through the crossings, that comes from the first line. Since the construction happens to be a special case of that in the next section, we omit the details and refer the reader to [10] or to the next section.

4.2. Riemann solver for algorithm (RA2). To solve Riemann problems according to (RA2) we need some additional parameters called priority and traffic distribution parameters. For simplicity of exposition, consider first a junction J in which there are two transmission lines with incoming traffic and two transmission lines with outgoing traffic. In this case we have only one priority parameter $q \in]0,1[$ and one traffic distribution parameter $\alpha \in]0,1[$. We denote by $\rho_i(t,x)$, $i = 1,2$, and $\rho_j(t,x)$, $j = 3,4$, the traffic densities, respectively, on the incoming transmission lines and on the outgoing ones and by $(\rho_{1,0},\rho_{2,0},\rho_{3,0},\rho_{4,0})$ the initial datum.

Define $\gamma_i^{\max}$ and $\gamma_j^{\max}$ as follows:

$$\gamma_i^{\max} = \begin{cases} f(\rho_{i,0}) & \text{if } \rho_{i,0} \in [0,\sigma], \\ f(\sigma) & \text{if } \rho_{i,0} \in]\sigma,1], \end{cases} \quad i = 1,2, \tag{10}$$

and

$$\gamma_j^{\max} = \begin{cases} f(\sigma) & \text{if } \rho_{j,0} \in [0,\sigma], \\ f(\rho_{j,0}) & \text{if } \rho_{j,0} \in]\sigma,1], \end{cases} \quad j = 3,4. \tag{11}$$

The quantities $\gamma_i^{\max}$ and $\gamma_j^{\max}$ represent the maximum flux that can be obtained by a single wave solution on each transmission line. In order to maximize the number of packets through the junction over incoming and outgoing lines we define

$$\Gamma = \min\{\Gamma_{in}^{\max},\Gamma_{out}^{\max}\},$$

where $\Gamma_{in}^{\max} = \gamma_1^{\max} + \gamma_2^{\max}$ and $\Gamma_{out}^{\max} = \gamma_3^{\max} + \gamma_4^{\max}$. Thus we want to have Γ as a flux through the junction.

Reasoning as in Theorem 13, one can easily see that to solve the Riemann problem, it is enough to determine the fluxes $\hat\gamma_i = f(\hat\rho_i)$, $i = 1,2$. In fact, to have simple

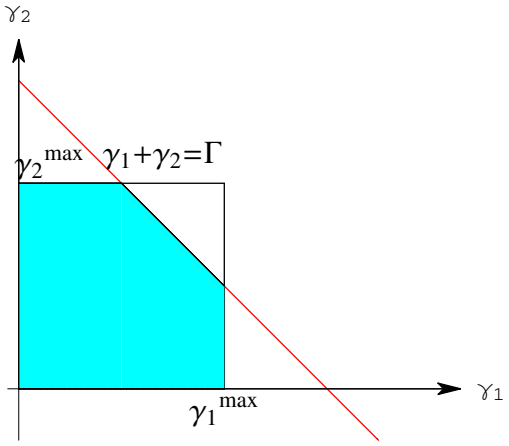


FIG. 4. *Case* $\Gamma_{in}^{\max} > \Gamma$.

waves with the appropriate velocities, i.e., negative on incoming lines and positive on outgoing ones, we get the constraints (8), (9). We have to distinguish two cases.

Case I. $\Gamma_{in}^{\max} = \Gamma$.

Case II. $\Gamma_{in}^{\max} > \Gamma$.

In the first case we set $\hat{\gamma}_i = \gamma_i^{\max}$, $i = 1, 2$.

Let us analyze the second case in which we use the priority parameter q.

Not all packets can enter the junction, so let C be the amount of packets that can go through. Then qC packets come from the first incoming line and $(1-q)C$ packets from the second. Consider the space (γ_1, γ_2) and define the following lines:

$$r_q : \gamma_2 = \frac{1-q}{q}\gamma_1,$$

$$r_\Gamma : \gamma_1 + \gamma_2 = \Gamma.$$

Define P to be the point of intersection of the lines r_q and r_Γ. Recall that the final fluxes should belong to the region (see Figure 4):

$$\Omega = \{(\gamma_1, \gamma_2) : 0 \le \gamma_i \le \gamma_i^{\max}, i = 1, 2\}.$$

We distinguish two cases.

Case (a). P belongs to Ω.

Case (b). P is outside Ω.

In the first case we set $(\hat{\gamma}_1, \hat{\gamma}_2) = P$, while in the second case we set $(\hat{\gamma}_1, \hat{\gamma}_2) = Q$, with $Q = proj_{\Omega \cap r_\Gamma}(P)$, where *proj* is the usual projection on a convex set; see Figure 5.

The reasoning can be repeated also in the case of n incoming lines. In $\mathbb{R}^n$ the line r_q is given by $r_q = tv_q$, $t \in \mathbb{R}$, with $v_q \in \Delta_{n-1}$, where

$$\Delta_{n-1} = \left\{ (\gamma_1, \ldots, \gamma_n) : \gamma_i \ge 0, i = 1, \ldots, n, \sum_{i=1}^{n} \gamma_i = 1 \right\}$$

is the $(n-1)$-dimensional simplex and

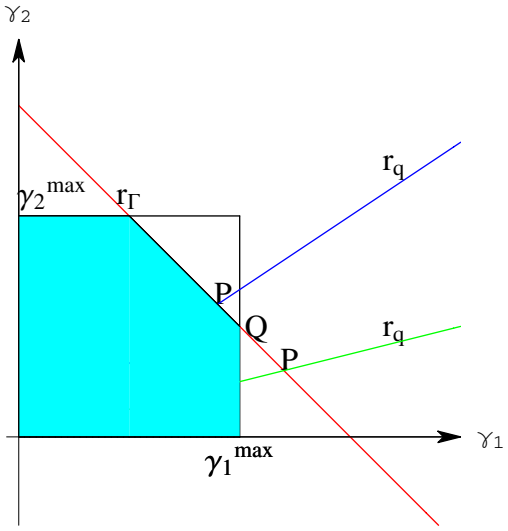


FIG. 5. *P belongs to Ω and P is outside Ω.*

$$H_\Gamma = \left\{ (\gamma_1, \dots, \gamma_n) : \sum_{i=1}^{n} \gamma_i = \Gamma \right\}$$

is a hyperplane where $\Gamma = \min\{\sum_{in} \gamma_i^{\max}, \sum_{out} \gamma_j^{\max}\}$. Since $v_q \in \Delta_{n-1}$, there exists a unique point $P = r_q \cap H_\Gamma$. If $P \in \Omega$, then we set $(\hat{\gamma}_1, \dots, \hat{\gamma}_n) = P$. If $P \notin \Omega$, then we set $(\hat{\gamma}_1, \dots, \hat{\gamma}_n) = Q = proj_{\Omega \cap H_\Gamma}(P)$, the projection over the subset $\Omega \cap H_\Gamma$. Observe that the projection is unique since $\Omega \cap H_\Gamma$ is a closed convex subset of H_Γ.

Remark 14. A possible alternative definition in the case $P \notin \Omega$ is to set $(\hat{\gamma}_1, \dots, \hat{\gamma}_n)$ as one of the vertices of $\Omega \cap H_\Gamma$.

Let us now determine $\hat{\gamma}_j$, $j = 3, 4$. As for the incoming transmission lines we have to distinguish two cases.

Case I. $\Gamma_{out}^{\max} = \Gamma$.

Case II. $\Gamma_{out}^{\max} > \Gamma$.

In the first case $\hat{\gamma}_j = \gamma_j^{\max}$, $j = 3, 4$. Let us determine $\hat{\gamma}_j$ in the second case. Recall α the traffic distribution parameter. Since not all packets can go on the outgoing transmission lines, we let C be the amount that goes through. Then αC packets go on the outgoing line I_3 and $(1-\alpha)C$ on the outgoing line I_4.

Now we can proceed exactly as in the previous case with q replaced by α. More precisely, we define r_α by the equation $\gamma_4 = \frac{1-\alpha}{\alpha}\gamma_3$, r_Γ by $\gamma_3 + \gamma_4 = \Gamma$, and P to be the point of intersection of the lines r_α and r_Γ. Setting $\Omega = \{(\gamma_3, \gamma_4) : 0 \le \gamma_j \le \gamma_j^{\max}, j = 3, 4\}$, we distinguish two cases.

Case (a). P belongs to Ω.

Case (b). P is outside Ω.

In the first case we set $(\hat{\gamma}_3, \hat{\gamma}_4) = P$, while in the second case we set $(\hat{\gamma}_3, \hat{\gamma}_4) = Q$, where $Q = proj_{\Omega \cap r_\Gamma}(P)$. Again, we can extend to the case of m outgoing lines as for the incoming lines defining the hyperplane $H_\Gamma = \{(\gamma_{n+1}, \dots, \gamma_{n+m}) : \sum_{j=n+1}^{n+m} \gamma_j = \Gamma\}$ and choosing a vector $v_\alpha \in \Delta_{m-1}$.

Remark 15. An alternative way of choosing the vector v_α is the following. We assume that a traffic distribution matrix A is assigned; then we compute $\hat{\gamma}_1, \dots, \hat{\gamma}_n$

as before and choose $v_\alpha \in \Delta_{m-1}$ by

$$v_\alpha = \Delta_{m-1} \cap \{tA(\hat\gamma_1, \dots, \hat\gamma_n) : t \in \mathbb{R}\}.$$

The solution to Riemann problems in this section is consistent, as shown by the next lemma.

LEMMA 16. (CC) *holds for the Riemann solver for* (RA2) *defined in this section.*

Proof. Let $\rho_0 = (\rho_{1,0}, \dots, \rho_{4,0})$ be the initial datum and $\hat\rho = RS(\rho_0)$. Assume, first, that $\Gamma < \Gamma_{in}^{\max}$. Define $\hat\gamma_i^{\max}$ to be the maximum flux on I_i given by a wave with left datum $\hat\rho_i$ and then set $\hat\Gamma_{in}^{\max} = \hat\gamma_1^{\max} + \hat\gamma_2^{\max}$. Then $\hat\Gamma_{in}^{\max} \geq \Gamma_{in}^{\max}$. Indeed if $\rho_{i,0} \in [0, \sigma[$, then $\hat\rho_i \in \{\rho_{i,0}\} \cup]\tau(\rho_{i,0}), \rho_{\max}]$ and $\hat\gamma_i^{\max} \geq \gamma_i^{\max} = f(\rho_{i,0})$, while if $\rho_{i,0} \in [\sigma, \rho_{\max}]$, then $\hat\rho_i \in [\sigma, \rho_{\max}]$ and so $\hat\gamma_i^{\max} = f(\sigma) = \gamma_i^{\max}$. The case $\Gamma < \Gamma_{out}^{\max}$ is treated similarly. ∎

5. Estimates on density variation. In this section we derive estimates on the total variation of the densities along a wave-front tracking approximate solution (constructed as in [11]) for both routing algorithms. This allows us to construct the solutions to the Cauchy problem in the standard way; see [7]. From now on, we assume that every junction has exactly two incoming transmission lines and two outgoing ones. This hypothesis is crucial, because the presence of more complicated junctions may provoke additional increases of the total variation of the flux and so of the density. The case where junctions have at most two incoming transmission lines and at most two outgoing ones can be treated in the same way.

From now on we fix a telecommunication network $(\mathcal{I}, \mathcal{J})$, with each node having at most two incoming and at most two outgoing lines, and a wave-front tracking approximate solution ρ, defined on the telecommunication network.

5.1. Algorithm (RA1). We first introduce the following.

DEFINITION 17. *For every transmission line I_i, $i = 1, \dots, N$, we indicate by*

$$\left(\rho_-^\beta, \rho_+^\beta\right), \quad \beta \in A = A(\rho, t, i), \quad A \text{ finite set},$$

the discontinuities on line I_i at time t, and by $x^\beta(t)$, $\lambda^\beta(t)$, $\beta \in A$, respectively, their positions and velocities at time t. We also refer to the wave β to indicate the discontinuity $(\rho_-^\beta, \rho_+^\beta)$.

We have the following.

LEMMA 18. *For some $K > 0$, we have*

$$\begin{aligned} TV(f(\rho(t,\cdot))) &\leq e^{Kt} TV(f(\rho(0+,\cdot))) \\ &\leq e^{Kt}(TV(f(\rho(0,\cdot))) + 2Nf(\sigma)) \end{aligned}$$

for each $t \geq 0$, where N is the total number of transmission lines of the network.

For the proof see Lemma 18 in [11]. To estimate the total variation of densities and to pass to the limit we need some additional notation.

DEFINITION 19. *For every line I_i, we define two curves $Y_-^{i,\rho}(t), Y_+^{i,\rho}(t)$, called the boundary of external flux (BEF), in the following way. We set the initial condition $Y_-^{i,\rho}(0) = a_i$, $Y_+^{i,\rho}(0) = b_i$ (if $a_i = -\infty$, then $Y_-^{i,\rho} \equiv -\infty$ and if $b_i = +\infty$, then $Y_+^{i,\rho} \equiv +\infty$). We let $Y_\pm^{i,\rho}(t)$ follow the generalized characteristic as defined in* [12], *letting $Y_-^{i,\rho}(t) = a_i$ (resp., $Y_+^{i,\rho}(t) = b_i$) if the generalized characteristic reaches the boundary and $f'(\rho(t, a_i)) < 0$ (resp., $f'(\rho(t, b_i)) > 0$). (In this way $Y_\pm^{i,\rho}(t)$ may coincide with a_i or b_i for some time intervals.) Let $\bar t$ be the first time such that*

$Y^{i,\rho}_-(\bar t) = Y^{i,\rho}_+(\bar t)$ *(possibly* $\bar t = +\infty$*). Then we let* $Y^{i,\rho}_\pm$ *be defined on* $[0, \bar t]$*. Finally, we define the sets*

$$D^i_1(\rho) = \left\{(t,x) : t \in [0, \bar t[: Y^{i,\rho}_-(t) < x < Y^{i,\rho}_+(t)\right\}$$

and

$$D^i_2(\rho) = [0, +\infty[\times [a_i, b_i] \setminus D^i_1(\rho).$$

Clearly $Y^{i,\rho}_\pm(t)$ bound the set on which the datum is not influenced by other transmission lines through the junctions.

DEFINITION 20. *Fix a transmission line* I_i*,* $i = 1, \dots, N$*, and a junction* J*. A wave* β *in* I_i *is said to be a big wave if*

$$sgn(\rho^\beta_- - \sigma) \cdot sgn(\rho^\beta_+ - \sigma) \le 0,$$

where $sgn(0) = 0$*. We say that an incoming transmission line* I_i *has a bad datum at* J *at time* $t > 0$ *if*

$$\rho_i(t, b_i-) \in [0, \sigma[\,,$$

while we say that an outgoing transmission line I_j *has a bad datum at* J *at time* $t > 0$ *if*

$$\rho_j(t, a_j+) \in]\sigma, 1]\,.$$

Our aim is now to bound, for each line I_i, the number of big waves inside the region $D^i_2(\rho)$, i.e., those generated by the influence of external lines.

LEMMA 21. *Let* $\bar t$ *be the time at which the two BEFs* $Y^{i,\rho}_\pm$ *interact. Assume* $\bar t < +\infty$*,* $Y^{i,\rho}_\pm(\bar t) \in]a_i, b_i[$*, and define*

$$\hat\rho_{out} = \rho\left(Y^{i,\rho}_\pm(\bar t)-\right), \quad \hat\rho_{in} = \rho\left(Y^{i,\rho}_\pm(\bar t)+\right), \quad \rho^* = \lim_{t\uparrow\bar t} \rho\left(Y^{i,\rho}_-(t)+\right) = \lim_{t\uparrow\bar t} \rho\left(Y^{i,\rho}_+(t)-\right).$$

If $\hat\rho_{in}$*, respectively,* $\hat\rho_{out}$*, is a bad datum for* I_i *as incoming line, respectively, for* I_i *as outgoing line, then there exists no value* ρ^* *of the density such that*

$$\lambda(\hat\rho_{out}, \rho^*) > \lambda(\rho^*, \hat\rho_{in}).$$

Proof. Since $\hat\rho_{out}$ and $\hat\rho_{in}$ are bad data for, respectively, an outgoing transmission line and an incoming transmission line, it follows that

$$\hat\rho_{out} \in]\sigma, 1]\,, \quad \hat\rho_{in} \in [0, \sigma[\,.$$

Observe that $\hat\rho_{out}$ and ρ^* must be connected by a single wave, and thus $\rho^* \ge \sigma$; otherwise the wave would be split in a fan of rarefaction shocks.

Similarly, ρ^* and $\hat\rho_{in}$ must be connected by a single wave, and thus $\rho^* \le \sigma$; otherwise the wave would be split in a fan of rarefaction shocks.

Finally, $\rho^* = \sigma$, but then

$$\lambda(\hat\rho_{out}, \rho^*) \le 0 \le \lambda(\rho^*, \hat\rho_{in})$$

and the conclusion holds. ∎

LEMMA 22. *For every $t \geq 0$, there are at most two big waves on*

$$\left\{x : (t,x) \in D_2^i(\rho)\right\} \subseteq [a_i, b_i]\,.$$

Proof. A big wave can originate at time t on transmission line I_i from J only if the line I_i has a bad datum at J at time t. If this happens, then, from Theorem 13, line I_i has not a bad datum at J up to the time in which a big wave is absorbed from I_i. This concludes the proof if $D_2^i(\rho)$ is formed by two connected components.

It remains to consider the time at which the two BEFs interact. By Lemma 21 we have that not both connected components can contain a big wave. Thus again there are at most two big waves. □

Up to now, we did not make use of assumption (F), which is necessary for the next lemma.

LEMMA 23. *Assume* (F)*; then for some $K > 0$, we have*

$$TV(\rho(t,\cdot)) \leq TV(\rho(0,\cdot)) + 2N\left(\frac{e^{Kt}f(\sigma)}{\bar{v}} + 1\right)$$

for each $t \geq 0$, where N is the total number of transmission lines of the network.

Proof. Let $TV(h;[a,b])$ denote the total variation of the function h over the interval $[a,b]$ and define

$$TV^j(\rho(t)) = \sum_i TV(\rho(t); D_j^i(\rho(t))), \qquad j = 1,2,$$

which are, respectively, the total variation of $\rho(t)$ due to the evolution only inside each line I_i and by interaction with junctions. Clearly,

$$TV(\rho(t)) = TV^1(\rho(t)) + TV^2(\rho(t)).$$

Since $D_1^i(\rho(t))$ is not influenced by external lines, we are in the situation of a conservation law on $\mathbb{R}$, hence

$$TV^1(\rho(t)) \leq TV(\rho(0)).$$

Let $B(t)$ denote the number of big waves generated from junctions, i.e., the number of big waves in $\bigcup_i D_2^i(\rho(t))$. Then by the chain rule for BV functions and Lemma 18,

$$TV^2(\rho(t)) \leq \frac{1}{\bar{v}} TV^2(f(\rho(t)) + B(t) \leq \frac{1}{\bar{v}} e^{Kt}(TV^2(f(\rho(0+))) + B(t). \tag{12}$$

Now that $TV^2(\rho(0)) = 0$, and thus, using again Lemmas 18 and 22, the following relation holds:

$$TV^2(\rho(t)) \leq \frac{1}{\bar{v}} e^{Kt} 2N f(\sigma) + 2N. \tag{13}$$

Finally we get

$$TV(\rho(t)) = TV^1(\rho(t)) + TV^2(\rho(t)) \leq TV(\rho(0)) + 2N\left(\frac{e^{Kt}f(\sigma)}{\bar{v}} + 1\right). \qquad \square$$

Thanks to Lemma 23 and the Lipschitz continuous dependence in L^1_{loc} of wave-front tracking approximations, we can apply the Helly theorem, as in [7], to get existence of solutions.

THEOREM 24. *Fix a telecommunication network* (I, J) *and assume* (F). *Given* $T > 0$, *for every initial data there exists an admissible solution to the Cauchy problem on the network defined on* $[0, T]$.

Let us observe that there is no Lipschitz continuous dependence by initial data with respect to the L^1 norm. In fact it is possible to choose two piecewise constant initial data, which are exactly the same except for a shift of a discontinuity, such that the L^1-distance of the two corresponding solutions increases by an arbitrary multiplicative factor (see [11]).

5.2. Algorithm (RA2). Let us now estimate the flux total variation and the density total variation for routing algorithm (RA2). We can define BEFs, bad data, and big waves as in the previous section.

Fix a junction J with two incoming transmission lines I_1 and I_2 and two outgoing ones I_3 and I_4.

Suppose that at some time $\bar{t}$ a wave interacts with the junction J and let $(\rho_1^-, \rho_2^-, \rho_3^-, \rho_4^-)$ and $(\rho_1^+, \rho_2^+, \rho_3^+, \rho_4^+)$ indicate the equilibrium configurations at the junction J before and after the interaction, respectively. Introduce the following notation:

$$\gamma_i^\pm = f(\rho_i^\pm), \quad \Gamma_{in}^\pm = \gamma_{1,\max}^\pm + \gamma_{2,\max}^\pm, \quad \Gamma_{out}^\pm = \gamma_{3,\max}^\pm + \gamma_{4,\max}^\pm,$$

$$\Gamma^\pm = \min\{\Gamma_{in}^\pm, \Gamma_{out}^\pm\},$$

where $\gamma_{i,\max}^\pm$, $i = 1, 2$, and $\gamma_{j,\max}^\pm$, $j = 3, 4$, are defined as in (10) and (11). In general $-$ and $+$ denote the values before and after the interaction, while by Δ we indicate the variation, i.e., the value after the interaction minus the value before. For example, $\Delta\Gamma = \Gamma^+ - \Gamma^-$. Let us denote by $TV(f)^\pm = TV(f(\rho(\bar{t}\pm, \cdot)))$ the flux variation of waves before and after the interaction, and by

$$TV(f)_{in}^\pm = TV(f(\rho_1(\bar{t}\pm, \cdot))) + TV(f(\rho_2(\bar{t}\pm, \cdot))),$$

$$TV(f)_{out}^\pm = TV(f(\rho_3(\bar{t}\pm, \cdot))) + TV(f(\rho_4(\bar{t}\pm, \cdot)))$$

the flux variation of waves before and after the interaction, respectively, on incoming and outgoing lines.

Let us prove some estimates which are used later to control the total variation of the density function. For simplicity, from now on we assume that

(A) the wave interacting at time $\bar{t}$ with J comes from line 1 and we let ρ_1 be the value on the left of the wave.

The case of a wave from an outgoing line can be treated similarly.

LEMMA 25. *We have*

$$sgn\,(\Delta\gamma_3) \cdot sgn\,(\Delta\gamma_4) \geq 0.$$

Proof. To prove the lemma it is enough to observe that a variation of γ_3 is due to a movement along the line r_q or along $\gamma_3 = c_1$ or $\gamma_4 = c_2$ with c_1 and c_2 constant. In each case $\Delta\gamma_3$ and $\Delta\gamma_4$ have the same sign. □

In the same way we can prove the following lemma.

LEMMA 26. *We have*

$$sgn(\gamma_1^+ - \gamma_1) \cdot sgn(\Delta\gamma_2) \geq 0,$$

where $\gamma_1 = f(\rho_1)$.

LEMMA 27. *It holds that*

$$TV(f)^+_{out} = |\Delta\Gamma|.$$

Proof. To prove the lemma it is enough to observe that

$$\Gamma^- = \gamma_3^- + \gamma_4^-, \quad \Gamma^+ = \gamma_3^+ + \gamma_4^+,$$

$$|\Delta\Gamma| = |\Gamma^+ - \Gamma^-| = |(\gamma_3^+ - \gamma_3^-) + (\gamma_4^+ - \gamma_4^-)|,$$

from which, by Lemma 25, we have

$$|\Delta\Gamma| = |\Delta\gamma_3| + |\Delta\gamma_4| = TV(f)^+_{out}. \qquad \square$$

LEMMA 28. *We have*

$$TV(f)^-_{in} = TV(f)^+_{in} + |\Delta\Gamma|. \tag{14}$$

Proof. Clearly since the wave on the first line has positive velocity, we have $0 \le \rho_1 \le \sigma$. Since $\rho_1 \le \sigma$, observe that the maximum flux for ρ_1^+, which is the solution with initial data ρ_1, is given by $\gamma_{1,\max} = f(\rho_1)$. Also

$$TV(f)^- = TV(f)^-_{in} = |\gamma_1 - \gamma_1^-|.$$

We have two possibilities.

Case 1. $\rho_1^- \le \sigma$.

Case 2. $\rho_1^- > \sigma$.

Let us first analyze Case 1. Then we further split it into two subcases.

Case 1(a). $\rho_1 < \rho_1^-$.

Case 1(b). $\rho_1 > \rho_1^-$.

If 1(a) holds true, since $\rho_1 < \rho_1^-$, we get $\gamma_{1,\max} = f(\rho_1) < f(\rho_1^-) = \gamma_{1,\max}^-$ and one of the following holds.

Case 1(a.1). $\Gamma^- = \Gamma^-_{in}$.

Case 1(a.2). $\Gamma^- = \Gamma^-_{out}$.

In Case 1(a.1) from $\gamma_{1,\max} < \gamma_{1,\max}^-$ and $\Gamma^- = \Gamma^-_{in}$, it follows that $\Gamma^+ = \Gamma^+_{in}$, from which $\gamma_2^+ = \gamma_2^-$, $\gamma_1^+ = \gamma_1$ and then $TV(f)^+_{in} = 0$.

In Case 1(a.2) we have $\gamma_{1,\max} < \gamma_{1,\max}^-$, and hence $\Gamma^-_{in} \ge \Gamma^-$ and $\gamma_{1,\max} + \gamma_{2,\max}^- < \Gamma^-_{in}$. The following distinction must be considered.

Case 1(a.2.1). $\gamma_{1,\max} + \gamma_{2,\max}^- \ge \Gamma^-$.

Case 1(a.2.2). $\gamma_{1,\max} + \gamma_{2,\max}^- < \Gamma^-$.

If Case 1(a.2.1) holds, from $\gamma_{1,\max} + \gamma_{2,\max}^- \ge \Gamma^-$, we have that $\Gamma^+ = \Gamma^-$, from which $|\Delta\Gamma| = 0$. By Lemma 26 the conclusion holds.

In the opposite Case 1(a.2.2) from $\gamma_{1,\max} + \gamma_{2,\max}^- < \Gamma^-$, one gets $\Gamma^+ = \gamma_{1,\max} + \gamma_{2,\max}^-$, from which it follows that $TV(f)^+_{in} = 0$. Then $|\Delta\Gamma| = \left|\gamma_1^- - \gamma_1\right| = TV(f)^-_{in}$. Case 1(a) is thus finished.

Let us now focus on Case 1(b). We have to distinguish two possibilities.

Case 1(b.1). $\Gamma^- = \Gamma^-_{out}$.

Case 1(b.2). $\Gamma^- = \Gamma^-_{in}$.

If Case 1(b.1) holds, from $\Gamma^- = \Gamma^-_{out}$ it follows that $\gamma_{1,\max} + \gamma_{2,\max}^- > \Gamma^-_{in}$. Then $\Gamma^+ = \Gamma^-$; hence $|\Delta\Gamma| = 0$ and by Lemma 26 the conclusion holds.

In Case 1(b.2), we have $\gamma_{1,\max} + \gamma_{2,\max}^- > \Gamma^-_{in}$ and $\Gamma^-_{out} \ge \Gamma^-_{in}$, and the following cases may happen.

Case 1(b.2.1). $\gamma_{1,\max} + \gamma^-_{2,\max} \le \Gamma^-_{out}$.

Case 1(b.2.2). $\gamma_{1,\max} + \gamma^-_{2,\max} > \Gamma^-_{out}$.

Consider Case 1(b.2.1) first. From $\gamma_{1,\max} + \gamma^-_{2,\max} \le \Gamma^-_{out}$, one has $TV(f)^+_{in} = 0$, and hence $|\Delta\Gamma| = \left|\gamma_1 - \gamma_1^-\right| = TV(f)^-_{in}$.

In Case 1(b.2.2), from $\gamma_{1,\max}+\gamma^-_{2,\max} > \Gamma^-_{out}$ we obtain $\Gamma^+ = \Gamma^+_{out}$. By Lemma 25,

$$TV(f)^+_{in} = \gamma_{1,\max} + \gamma^-_{2,\max} - \Gamma^-_{out},$$

$$TV(f)^-_{in} = \gamma_{1,\max} - \gamma^-_{1,\max},$$

and hence

$$\begin{aligned} TV(f)^-_{in} - TV(f)^+_{in} &= -\gamma^-_{1,\max} - \gamma^-_{2,\max} + \Gamma^-_{out} \\ &= \Gamma^+ - \Gamma^-_{in} = \Gamma^+ - \Gamma^- = |\Delta\Gamma|. \end{aligned}$$

Let us analyze Case (2). Since $\rho_1^- > \sigma$ it follows that $\rho_1 < \tau(\rho_1^-) < \sigma$. Observe that $\gamma_1 = f(\rho_1) < f(\rho_1^-) = \gamma_1^-$ and $\gamma^-_{1,\max} = f(\sigma)$, $\gamma_{1,\max} = f(\rho_1)$.

We have to distinguish two cases.

Case 2(a). $\Gamma^- = \Gamma^-_{in}$.

Case 2(b). $\Gamma^- = \Gamma^-_{out}$.

If Case 2(a) holds, then one gets $\gamma_{1,\max} + \gamma^-_{2,\max} < \Gamma^-$, from which it follows that $\Gamma^+ = \gamma_{1,\max} + \gamma^-_{2,\max}$. Hence $TV(f)^+_{in} = 0$ and the conclusion holds.

For the opposite Case (2b), we have $\gamma_{1,\max} + \gamma^-_{2,\max} < \Gamma^-_{in}$ and $\Gamma^-_{in} \ge \Gamma^-_{out}$. Hence the following two cases are possible.

Case 2(b.1). $\gamma_{1,\max} + \gamma^-_{2,\max} \ge \Gamma^-_{out}$.

Case 2(b.2). $\gamma_{1,\max} + \gamma^-_{2,\max} < \Gamma^-_{out}$.

In Case 2(b.1), from $\gamma_{1,\max} + \gamma^-_{2,\max} \ge \Gamma^-_{out}$, it follows that $\Gamma^+ = \Gamma^-$. The latter implies $|\Delta\Gamma| = 0$ and the conclusion follows from Lemma 26.

In Case 2(b.2) from $\gamma_{1,\max}+\gamma^-_{2,\max} < \Gamma^-_{out}$, we obtain $\Gamma^+ = \gamma_{1,\max}+\gamma^-_{2,\max}$. Thus, by Lemma 26, we get

$$TV(f)^+_{out} = \Gamma^+ - (\gamma_1 + \gamma_2^-) = (\gamma_{1,\max} + \gamma^-_{2,\max}) - (\gamma_1 + \gamma_2^-) = \gamma^-_{2,\max} - \gamma_2^-.$$

It follows that

$$\begin{aligned} |\Delta\Gamma| = \Gamma^- - \Gamma^+ &= \gamma_1^- + \gamma_2^- - (\gamma_{1,\max} + \gamma^-_{2,\max}) \\ = (\gamma_1^- - \gamma_{1,\max}) + (\gamma_2^- - \gamma^-_{2,\max}) &= TV(f)^-_{in} - TV(f)^+_{out}, \end{aligned}$$

and the conclusion holds. The proof is thus complete. ∎

From the above results, we are ready to state the following.

LEMMA 29. *The flux variation $TV(f)$ is conserved along wave-front tracking approximations.*

Notice that this result is much stronger than that obtained for routing algorithm (RA1), for which only an exponential in time bound for the flux variation is achieved.

Proof. From Lemmas 27 and 28 we get

$$TV(f)^- = TV(f)^-_{in} = TV(f)^+_{in} + |\Delta\Gamma| = TV(f)^+. \qquad ∎$$

The estimate on the number of big waves is valid also for algorithm (RA2); thus we bound the total variation of the densities as follows.

THEOREM 30. *Consider a telecommunication network* $(\mathcal{I}, \mathcal{J})$ *and assume* (F). *Let* ρ *be a wave-front tracking approximate solution. Then*

$$TV(\rho(t,\cdot)) \le TV(\rho(0,\cdot)) + 2N\left(\frac{f(\sigma)}{\bar{v}} + 1\right)$$

for each $t \ge 0$*, where* N *is the total number of transmission lines of the network. Moreover, given* $T > 0$*, there exists an admissible solution to the Cauchy problem on the network defined on* $[0,T]$ *for every initial data.*

5.2.1. Uniqueness and Lipschitz continuous dependence. The aim of this section is to prove Lipschitz continuous dependence by initial data for solutions to the Cauchy problem on the network, controlling for any two approximate solutions ρ, ρ' how their distance varies in time. We use the method introduced in [8], which is based on a Riemannian type distance on L^1. There are various alternative methods to treat uniqueness and continuous dependence for the case of scalar conservation laws on the real line, including Kruzkov entropies (cf. [7]), viscous approximations (cf. [22, 23]), and Bressan–Liu–Yang functionals (see [9]). No one of these methods seems to work for the network case. In fact, the Kruzkov method requires one to estimate integrals on a region in $\mathbb{R}^2$, which now is replaced by an integral on the topological space obtained by the product of the network and $\mathbb{R}$. On the other hand, it is not clear how to define a viscous solution on the network, in particular how to treat boundary data at nodes, and how to pass to the limit. Finally, a Bressan–Liu–Yang-type functional requires one to introduce a definition of approaching waves, but, on a general network, with complicated topology, every wave is potentially approaching each other.

The basic idea is to estimate the L^1-distance viewing L^1 as a Riemannian manifold. We consider the subspace of piecewise constant functions and "generalized tangent vectors" consisting of two components (v, ξ), where $v \in L^1$ describes the L^1 infinitesimal displacement, while $\xi \in \mathbb{R}^n$ describes the infinitesimal displacement of discontinuities. For example, take a family of piecewise constant functions $\theta \to \rho^\theta$, $\theta \in [0,1]$, each of which has the same number of jumps, say at the points $x_1^\theta < \cdots < x_N^\theta$. Assume that the following functions are well defined (Figure 6):

$$L^1 \ni v^\theta(x) \doteq \lim_{h\to 0} \frac{\rho^{\theta+h}(x) - \rho^\theta(x)}{h},$$

as well as the numbers

$$\xi_\beta^\theta \doteq \lim_{h\to 0} \frac{x_\beta^{\theta+h} - x_\beta^\theta}{h}, \qquad \beta = 1, \dots, N.$$

Then we say that γ admits tangent vectors $(v^\theta, \xi^\theta) \in T_{\rho^\theta} \doteq L^1(\mathbb{R}; \mathbb{R}^n) \times \mathbb{R}^n$. In general a path such as $\theta \to \rho^\theta$ is not differentiable with respect to the usual differential structure of L^1; in fact if $\xi_\beta^\theta \ne 0$, as $h \to 0$ the ratio $\left[\rho^{\theta+h}(x) - \rho^\theta\right]/h$ does not converge to any limit in L^1.

Moreover, we can compute the L^1-length of the path $\gamma : \theta \to \rho^\theta$ in the following way:

$$\|\gamma\|_{L^1} = \int_0^1 \left\|v^\theta\right\|_{L^1} d\theta + \sum_{\beta=1}^{N} \int_0^1 \left|\rho^\theta(x_\beta+) - \rho^\theta(x_\beta-)\right| \left|\xi_\beta^\theta\right| d\theta. \tag{15}$$

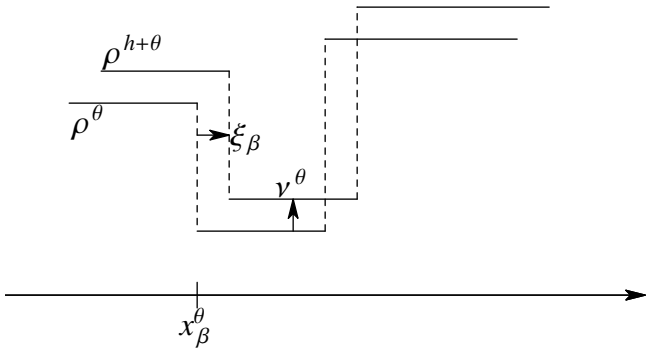


FIG. 6. *Construction of "generalized tangent vectors."*

According to (15), in order to compute the L^1-length of a path γ, we integrate the norm of its tangent vector, which is defined as follows:

$$\|(v,\xi)\| \doteq \|v\|_{L^1} + \sum_{\beta=1}^{N} |\Delta\rho_\beta|\,|\xi_\beta|\,,$$

where $\Delta\rho_\beta = \rho(x_\beta+) - \rho(x_\beta-)$ is the jump across the discontinuity x_β.

Let us introduce the following definition.

DEFINITION 31. *We say that a continuous map $\gamma : \theta \to \rho^\theta \doteq \gamma(\theta)$ from $[0,1]$ into L^1_{loc} is a regular path if the following holds. All functions ρ^θ are piecewise constant, with the same number of jumps, say at $x^\theta_1 < \cdots < x^\theta_N$, and coincide outside some fixed interval $]-M, M[$. Moreover, the function $\theta \to \rho^\theta_x$ is continuous from $[0,1]$ into L^1, and the map $\theta \to \rho^\theta$ admits a generalized tangent vector $D\gamma(\theta) = (v^\theta, \xi^\theta) \in T_{\gamma(\theta)} = L^1(\mathbb{R};\mathbb{R}^n) \times \mathbb{R}^N$, continuously depending on θ.*

Given two piecewise constant functions ρ and ρ', call $\Omega(\rho,\rho')$ the family of all regular paths $\gamma : [0,1] \to \gamma(t)$ with $\gamma(0) = \rho$, $\gamma(1) = \rho'$. The Riemannian distance between ρ and ρ' is given by

$$d(\rho,\rho') \doteq \inf\left\{\|\gamma\|_{L^1}\,, \gamma \in \Omega(\rho,\rho')\right\}.$$

To define d on all L^1, for given $\rho, \rho' \in L^1$ we set

$$d(\rho,\rho') \doteq \inf\{\|\gamma\|_{L^1} + \|\rho - \tilde\rho\|_{L^1} + \|\rho' - \tilde\rho'\|_{L^1} : \\ \tilde\rho, \tilde\rho' \text{ piecewise constant functions}, \gamma \in \Omega(\tilde\rho, \tilde\rho')\}\,.$$

It is easy to check that this distance coincides with the distance of L^1. For the systems case, one has to introduce weights; see [8].

Now we are ready to estimate the L^1-distance among solutions, studying the evolution of norms of tangent vectors along wave-front tracking approximations. Take ρ, ρ' piecewise constant functions and let $\gamma_0(\vartheta) = \rho^\vartheta$ be a regular path joining $\rho = \rho^0$ with $\rho' = \rho^1$. Define $\rho^\vartheta(t,x)$ to be a wave-front tracking approximate solution with initial data ρ^ϑ and let $\gamma_t(\vartheta) = \rho^\vartheta(t,\cdot)$.

If we can prove that, for every γ_0 (regular path) and every $t \geq 0$, γ_t is a regular path and

$$\|\gamma_t\|_{L^1} \leq \|\gamma_0\|_{L^1}\,, \tag{16}$$

then for every $t \geq 0$

$$\|\rho(t,\cdot) - \rho'(t,\cdot)\|_{L^1} \leq \inf_{\gamma_t} \|\gamma_t\|_{L^1} \leq \inf_{\gamma_0} \|\gamma_0\|_{L^1} = \|\rho(0,\cdot) - \rho'(0,\cdot)\|_{L^1}\,. \tag{17}$$

To obtain (16), hence (17), it is enough to prove that, for every tangent vector $(v,\xi)(t)$ to any regular path γ_t, one has

$$\|(v,\xi)(t)\| \leq \|(v,\xi)(0)\|\,, \tag{18}$$

i.e., the norm of a tangent vector does not increase in time. Moreover, if (17) is established, then uniqueness and Lipschitz continuous dependence of solutions to Cauchy problems is straightforwardly achieved passing to the limit on the wave-front tracking approximate solutions.

The same reasoning can be used on the network. If $\rho = (\rho_1, \ldots, \rho_N)$ is a solution on the network, then we set

$$\|\rho\|_{L^1} = \sum_i \|\rho_i\|_{L^1(I_i)}\,.$$

To estimate the distance among wave-front tracking solutions it is thus enough to prove (18). We prove the latter estimating the evolution of the tangent vector norm at each time. For this, we fix a time $\bar{t} \geq 0$ and, without loss of generality, treat the following cases:

(a) no interaction of waves takes place in any transmission line at $\bar{t}$, and no wave interacts with a junction;
(b) two waves interact at $\bar{t}$ on a transmission line, and no other interaction takes place;
(c) a wave interacts with a junction at $\bar{t}$, and no other interaction takes place.

In case (a) we can prove

$$\left[\frac{d}{dt}\,\|(v,\xi)(t)\|\right]_{t=\bar{t}} \leq 0,$$

while in cases (b) and (c), letting $(v,\xi)^{\pm}$ be the tangent vector before $(-)$ and after $(+)$ the interaction, we prove

$$\left\|(v,\xi)^+\right\| \leq \left\|(v,\xi)^-\right\|.$$

Let us first analyze case (a). Denote by x_β, σ_β, and ξ_β, respectively, the positions, sizes, and shifts of the discontinuities of the wave-front tracking approximate solution. Following [8] we get

$$\begin{aligned}
&\frac{d}{dt}\left\{\int |v(t,x)|\,dx + \sum_{\beta=1}^{N} |\xi_\beta|\,|\sigma_\beta|\right\} \\
&= -\left\{\sum_\beta \left(\lambda(\rho^-) - \dot{x}_\beta\right)\left|v^-\right| + \sum_\beta \left(\dot{x}_\beta - \lambda(\rho^+)\right)\left|v^+\right|\right\} \\
&\quad + \sum_\beta D\lambda(\rho^-,\rho^+)\cdot(v^-,v^+)\,(\operatorname{sign}\xi_\beta)\,|\sigma_\beta|\,,
\end{aligned}$$

with $\sigma_\beta = \rho^+ - \rho^-$, $\rho^\pm \doteq \rho(x_\beta \pm)$ and similarly for $v^\pm$. If the waves respect the Rankine–Hugoniot conditions, then

$$D\lambda(\rho^-, \rho^+)(v^-, v^+) = \left(\lambda(\rho^-) - \dot{x}_\beta\right) \frac{v^-}{|\sigma_\beta|} + \left(\dot{x}_\beta - \lambda(\rho^+)\right) \frac{v^+}{|\sigma_\beta|}$$

and

$$\frac{d}{dt} \left\{ \int |v(t,x)|\, dx + \sum_{\beta=1}^{N} |\xi_\beta|\, |\sigma_\beta| \right\} \leq 0. \tag{19}$$

Remark 32. To be precise, to obtain control over $TV(f)$ the wave-front tracking is slightly modified in the following way (see [11]). For every initial data ρ a sequence of piecewise constant approximations ρ_ν are constructed, converging to ρ in L^1. Then one chooses a sequence $\delta_\nu > 0$ converging to zero and construct wave-front tracking approximate solutions splitting rarefaction waves into a fan of rarefaction shocks, each of size at most δ_ν. If a rarefaction wave is originated at a junction with ρ^+ or ρ^- equal to σ, then we let $\dot{x}_\beta = 0$. However, since $\dot{x}_\beta = \frac{f(\sigma+\delta_\nu)-f(\sigma)}{\delta_\nu}$, $|\dot{x}_\beta - \bar{x}_\beta| = \delta_\nu$ we get

$$\frac{d}{dt} \left\{ \int |v(t,x)|\, dx + \sum_{\beta=1}^{N} |\xi_\beta|\, |\sigma_\beta| \right\} \leq 2\delta_\nu N,$$

where N is the number of transmission lines. In fact, by Lemma 22, there are at most two such waves on each transmission line. Hence the estimate (18) is obtained in the limit as ν tends to $+\infty$.

In case (b), we use the following lemma (see [11], for example).

LEMMA 33. *Let us consider in a transmission line two waves, with speeds λ_1 and λ_2, respectively, that interact, producing a wave with speed λ_3. If the first wave is shifted by ξ_1 and the second wave by ξ_2, then the shift of the resulting wave is given by*

$$\xi_3 = \frac{\lambda_3 - \lambda_2}{\lambda_1 - \lambda_2}\xi_1 + \frac{\lambda_1 - \lambda_3}{\lambda_1 - \lambda_2}\xi_2.$$

Moreover we have that

$$\Delta\rho_3\xi_3 = \Delta\rho_1\xi_1 + \Delta\rho_2\xi_2, \tag{20}$$

where $\Delta\rho_i$ are the signed strengths of the corresponding waves.

From (20) it follows that

$$|\Delta\rho_3\xi_3| \leq |\Delta\rho_1|\, |\xi_1| + |\Delta\rho_2|\, |\xi_2|\,,$$

from which

$$\left\|(v,\xi)^+\right\| \leq \left\|(v,\xi)^-\right\|. \tag{21}$$

For case (c) we report the lemma in [11].

LEMMA 34. *Let us consider a junction J with incoming lines I_1 and I_2 and outgoing lines I_3 and I_4. If a wave on a transmission line I_i ($i \in \{1, \dots, 4\}$) interacts*

with J and if ξ_i is the shift of the wave in I_i, then the shift ξ_j produced in a different line I_j ($j \in \{1,\dots,4\}\backslash\{i\}$) satisfies

$$\xi_j(\rho_j^+ - \rho_j^-) = \frac{\Delta\gamma_j}{\Delta\gamma_i}\xi_i(\rho_i^+ - \rho_i^-),$$

where $\Delta\gamma_l$ ($l \in \{i,j\}$) represents the variation of the flux in the line I_l and ρ_l^-, ρ_l^+ ($l \in \{i,j\}$) are the states at J in the line I_l, respectively, before and after the interaction.

Define $TV(f)^{\pm}$ to be the total variation of the flux of the solution before $(-)$ and after $(+)$ the interaction, and $TV(f)_i^{\pm}$ the same quantity on line I_i. Without loss of generality, we can assume that a wave from an incoming transmission line $\bar{\imath}$ interacts with a junction J and no other wave is present. Then $TV(f)^- = TV(f)_{\bar{\imath}}^-$ and $TV(f)^+ = \sum_j TV(f)_j^+$, where $TV(f)_j^+$ measures just the wave produced by the interaction. From Lemma 34 we have

$$|\xi_j|\,|\Delta\rho_j| = \frac{TV(f)_j^-}{TV(f)^-}\,|\xi_{\bar{\imath}}|\,|\Delta\rho_{\bar{\imath}}|\,.$$

Using Lemma 29 we conclude

$$\begin{aligned}\left\|(v,\xi)^+\right\| &= \|v\|_{L^1} + \sum_j |\xi_j|\,|\Delta\rho_j| = \|v\|_{L^1} + \sum_j \frac{TV(f)_j^-}{TV(f)_{\bar{\imath}}^-}\,|\xi_{\bar{\imath}}|\,|\Delta\rho_{\bar{\imath}}| \\ &= \|v\|_{L^1} + \frac{TV(f)^+}{TV(f)^-}\,|\xi_{\bar{\imath}}|\,|\Delta\rho_{\bar{\imath}}| = \left\|(v,\xi)^-\right\|.\end{aligned} \tag{22}$$

From (19), (21), and (22), we get the following.

THEOREM 35. *Consider a telecommunication network $(\mathcal{I},\mathcal{J})$ and assume* (F). *Then the solutions to Cauchy problems on the networks are unique and depend in a Lipschitz continuous way on initial data.*

REFERENCES

[1] D. AMADORI, *Initial-boundary value problems for systems of conservation laws*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 1–42.

[2] F. ANCONA AND A. MARSON, *Scalar nonlinear conservation laws with integrable boundary data*, Nonlinear Anal., 35 (1999), pp. 687–710.

[3] D. ARMBRUSTER, P. DEGOND, AND C. RINGHOFER, *A model for the dynamics of large queuing networks and supply chains*, SIAM J. Appl. Math., 66 (2006), pp. 896–920.

[4] F. BACCELLI, D. HONG, AND Z. LIU, *Fixed Points Methods for the Simulation of the Sharing of a Local Loop by Large Number of Interacting TCP Connections*, Technical Report RR-4154, INRIA, 2001.

[5] C. BARDOS, A. Y. LE ROUX, AND J. C. NÉDÉLEC, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[6] A. BRESSAN, *A contractive metric for systems of conservation laws with coinciding shock and rarefaction curves*, J. Differential Equations, 106 (1993), pp. 332–366.

[7] A. BRESSAN, *Hyperbolic Systems of Conservation Laws—The One-Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[8] A. BRESSAN, G. CRASTA, AND B. PICCOLI, *Well-Posedness of the Cauchy Problem for $n \times n$ Systems of Conservation Laws*, Mem. Amer. Math. Soc., 146 (694) (2000).

[9] A. BRESSAN, T. P. LIU, AND T. YANG, *L^1 stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[10] Y. CHITOUR AND B. PICCOLI, *Traffic circles and timing of traffic lights for cars flow*, Discrete Contin. Dyn. Syst. Ser. B, 5 (2005), pp. 599–630.

[11] G. M. COCLITE, M. GARAVELLO, AND B. PICCOLI, *Traffic flow on a road network*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.

[12] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, 1999.
[13] C. Daganzo, *Fundamentals of Transportation and Traffic Operations*, Pergamon-Elsevier, Oxford, UK, 1997.
[14] M. Garavello and B. Piccoli, *Traffic flow on a road network using the Aw–Rascle model*, Comm. Partial Differential Equations, 31 (2006), pp. 243–275.
[15] M. Herty and A. Klar, *Modeling, simulation, and optimization of traffic flow networks*, SIAM J. Sci. Comput., 25 (2003), pp. 1066–1087.
[16] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.
[17] F. Kelly, A. K. Maulloo, and D. K. H. Tan, *Rate control in communication networks: Shadow prices, proportional fairness, and stability*, J. Oper. Res. Soc., 49 (1998), pp. 237–252.
[18] J. P. Lebacque, *The Godunov scheme and what it means for first order flow models*, in Transportation and Traffic Theory, Proceedings of the 13th ISTTT, J. B. Lesort, ed., Pergamon, Oxford, UK, 1996, pp. 647–677.
[19] M. J. Lighthill and G. B. Whitham, *On kinetic waves.* II. *Theory of traffic flows on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.
[20] G. F. Newell, *Traffic Flow on Transportation Networks*, MIT Press, Cambridge, MA, 1980.
[21] P. I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.
[22] D. Serre, *Systems of Conservation Laws* I, Cambridge University Press, Cambridge, UK, 1999.
[23] D. Serre, *Systems of Conservation Laws* II, Cambridge University Press, Cambridge, UK, 2000.
[24] A. S. Tanenbaum, *Computer Networks*, Prentice Hall, Upper Saddle River, NJ, 2003.
[25] W. Willinger and V. Paxson, *Where mathematics meets the Internet*, Notices Amer. Math. Soc., 45 (1998), pp. 961–970.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 741–758

APPROXIMATE TRAVELING WAVES IN LINEAR REACTION-HYPERBOLIC EQUATIONS*

AVNER FRIEDMAN† AND GHEORGHE CRACIUN‡

Abstract. Linear reaction-hyperbolic equations arise in the transport of neurofilaments and membrane-bound organelles in axons. The profile of the solution was shown by simulations to be approximately that of a traveling wave; this was also suggested by formal calculations [M. C. Reed, S. Venakides, and J. J. Blum, *SIAM J. Appl. Math.*, 50 (1990), pp. 167–180]. In this paper we prove such a result rigorously.

Key words. axonal transport, hyperbolic equations, asymptotic approximations, traveling waves

AMS subject classifications. 35L45, 92C20, 92C40

DOI. 10.1137/050637947

1. Introduction.

This paper is concerned with the mathematical analysis of reaction-hyperbolic equations which describe transport of materials along a straight ray $l_0 = \{x : 0 < x < \infty\}$. The model is motivated from biology; it describes the transport of proteins and other molecules along the axon of a neuron. The proteins are formed near the nucleus of the cell, that is, at $x = 0$, and are transported to various locations along the axon, moving towards the synaptic end. Some material is also transported back, in retrograde motion. The transported materials include, for example, vesicles, membrane-bound organelles, and neurofilaments. Motor proteins attached to a vesicle (or a neurofilament) carry this cargo as they pace along a microtubule, step by step, energized by adenosine triphosphate (ATP) molecules. While some of the motors may be moving along a microtubule, others may be "resting" on-track, or even off-track, for a while. Thus the model has to deal with several populations of vesicles, depending in what state of motion they are. Earlier models of axonal transport were developed by Reed and Blum [10, 1, 2]. Using mass reaction laws and conservation of mass, they derived a system of hyperbolic equations and studied (mostly numerically) the particle concentration profile along the axon. The numerical results show that the transport of the particle concentrations has the profile of "approximate traveling waves"; experimentally, they arise from radiolabeling proteins in the soma and then observing the progress of the wave of label as it goes down the axon. The wave goes at constant velocity, but the front spreads so it is only approximately a traveling wave. Reed, Venakides, and Blum [11] considered a mathematical problem derived from such a transport model, in the biologically relevant case when the transition between the various populations is fast relative to the transport. Recent experimental results and computational models [4, 6, 8, 12, 13] also describe the dynamics of such transport.

*Received by the editors August 10, 2005; accepted for publication (in revised form) March 6, 2006; published electronically July 31, 2006. This work was supported by the National Science Foundation under agreement 0112050.

http://www.siam.org/journals/sima/38-3/63794.html

†Mathematical Biosciences Institute, Ohio State University, Columbus, OH 43210 (afriedman@mbi.ohio-state.edu).

‡Department of Mathematics and Department of Biomolecular Chemistry, University of Wisconsin, Madison, WI 53706 (craciun@math.wisc.edu).

Consider, for example, two populations p and q, with transition rates

$$p \underset{k_2}{\overset{k_1}{\rightleftharpoons}} q,$$

where q is moving and p is resting. Then the corresponding reaction-hyperbolic system describing their transport is given by

$$\begin{aligned} \varepsilon p_t &= -k_1 p + k_2 q, \\ \varepsilon\,(q_t + v_2 q_x) &= k_1 p - k_2 q, \end{aligned} \tag{1.1}$$

where ε is a small positive constant. The general transport problem for n species may be written in the form

$$\varepsilon\,(\partial_t + v_i \partial_x)\, p_i = \sum_{j=1}^{n} k_{ij} p_j \text{ for } 0 < x < \infty,\ t > 0,\ 1 \le i \le n, \tag{1.2}$$

where $k_{ij} \ge 0$ if $i \ne j$, the velocities v_i may be positive, negative, or zero, and

$$\sum_{i=1}^{n} k_{ij} = 0 \tag{1.3}$$

by conservation of mass; thus,

$$k_{jj} = -\sum_{\substack{i=1 \\ i \ne j}}^{n} k_{ij}.$$

We need to complement (1.2) with initial conditions, and with boundary conditions at $x = 0$ for each p_i for which $v_i > 0$.

The special case

$$\begin{aligned} &v_i = 0, \quad 1 \le i \le n-1, \\ &v_n > 0, \\ &p_i(x,0) = 0, \quad 1 \le i \le n,\ 0 < x < \infty \\ &p_n(0,t) = 1, \quad t > 0, \end{aligned} \tag{1.4}$$

was studied by Reed, Venakides, and Blum [11]. They derived formulas which suggest that if we write p_n in the form

$$p_n(x,t) = Q_\varepsilon\left(\frac{x - vt}{\sqrt{\varepsilon}}, t\right), \tag{1.5}$$

then

$$Q_\varepsilon(s,t) \to Q_0(s,t) \text{ as } \varepsilon \to \infty, \tag{1.6}$$

where Q_0 is a solution of the heat equation

$$\begin{aligned} &\partial_t Q_0 - \sigma^2 \partial_s^2 Q_0 = 0, \quad -\infty < s < \infty,\ t > 0, \\ &Q_0(s,0) = 1 \text{ if } s < 0, \\ &Q_0(s,0) = 0 \text{ if } s > 0; \end{aligned} \tag{1.7}$$

the parameters $v,\ \sigma^2$ are computed from the v_j and k_{ij}. Formula (1.5), together with (1.6), (1.7), shows the approximate traveling wave profile for the transport of the concentration $p_n(x,t)$. The assertion (1.5) in [11] was only formal, but Brooks [3] developed a probabilistic model which enabled her to prove (1.5), in some sense, in the special case of (1.1).

One of the aims of the present paper is to give a rigorous proof of (1.6) for the case (1.4), or more generally for the case when all $v_j \geq 0$, and

$$p_i(x,0) = \lambda_i q_0\left(\frac{x}{\sqrt{\varepsilon}}\right),\ 1 \leq i \leq n,$$
$$p_j(0,t) = \lambda_j \text{ if } v_j > 0,$$

where either $q_0 \equiv 0$ or $q_0(s)$ has compact support and $q_0(0) = 1$; here $(\lambda_1, \ldots, \lambda_n)$ is a vector with positive components and, at the same time, a generator of the null space of the matrix (k_{ij}).

The proof, given in section 3, is by PDE methods. Since, in the case $q_0 \equiv 0$, the function Q_0 is discontinuous only at the point $(0,0)$ while Q_ε is discontinuous along the half-line $x = vt,\ t \geq 0$, we don't expect the convergence in (1.6) to be in the uniform sense, at least not near the origin $(0,0)$; we shall prove the convergence in the weak L^r-sense for any $1 < r < \infty$.

In the case of (1.1) we shall prove, in section 2, in case q_0 is not identically zero, a "strong" convergence in (1.6), namely,

$$(1.8)\quad \int_{-\frac{vT}{\sqrt{\varepsilon}}}^{\infty} \left(Q_\varepsilon(s,t) - Q_0(s,t)\right)^2 ds + \int_0^T \int_{-\frac{vT}{\sqrt{\varepsilon}}}^{\infty} \left(\partial_s Q_\varepsilon(s,t) - \partial_s Q_0(s,t)\right)^2 ds dt \leq C\sqrt{\varepsilon}.$$

This estimate also holds for the case (1.2) if $n = 2$, i.e., for $v_2 > 0$ and $v_1 \geq 0$.

In [9] Pinsky considered the system (1.2) for $-\infty < x < \infty,\ t > 0$, with initial data

$$p_i(x,0) = f\left(\frac{x}{\sqrt{\varepsilon}}\right) \text{ for all } i,\ -\infty < x < \infty,$$

and proved that

$$|Q_\varepsilon(s,t) - Q_0(s,t)| \leq C\sqrt{\varepsilon},$$

where $Q_0(s,t)$ is a solution of a heat equation as in (1.7) for $-\infty < s < \infty,\ t > 0$, with $Q_0(s,0) = f(s),\ -\infty < s < \infty$. Pinsky's dynamical system is derived from a different (stochastic) model and, in particular, he assumes that $\sum_{j=1}^n k_{ij} = 0$ instead of (1.3). His proof is based on the construction of boundary layers and estimates via the Fourier transform in x; that proof does not extend to the system (1.2), (1.4), not even in the case $n = 2$.

We conclude the introduction by pointing out several open problems:

(i) Extend the strong convergence result to $n > 2$.
(ii) Extend the results of this paper to the case of forward *and* backward velocities (i.e., anterograde and retrograde transport).
(iii) Study the case where the kinetics is nonlinear.

2. The case $n = 2$. Throughout this paper we shall use the notation

$$D = \{(x,t) : \ 0 < x < \infty, \ t > 0\}.$$

Consider the process $p \underset{k_2}{\overset{k_1}{\rightleftharpoons}} q$, where k_1, k_2 are positive constants, with dynamics

$$\varepsilon(p_t + v_1 p_x) = -k_1 p + k_2 q, \tag{2.1}$$

$$\varepsilon(q_t + v_2 q_x) = k_1 p - k_2 q \tag{2.2}$$

in D, where v_1, v_2 are given constant velocities, and ε is a small positive number. We intend to prove that, as $\varepsilon \to 0$, $q(x,t)$ will behave like $Q(\frac{x-vt}{\sqrt{\varepsilon}}, t)$, where $Q(s,t)$ is a solution of a parabolic equation. Such a result requires, of course, that the initial values for q should be of the form $q_0(\frac{x}{\sqrt{\varepsilon}})$. We also choose initial data which are in equilibrium with respect to the process $p \underset{k_2}{\overset{k_1}{\rightleftharpoons}} q$. Thus, we assume that

$$p(x,0) = \frac{k_1}{k_2} q_0\left(\frac{x}{\sqrt{\varepsilon}}\right), \ q(x,0) = q_0\left(\frac{x}{\sqrt{\varepsilon}}\right) \text{ for } 0 \le x < \infty. \tag{2.3}$$

We further assume that $v_1 \ge 0, \ v_2 > 0, v_1 \ne v_2$ and prescribe the boundary condition

$$q(0,t) = 1 \text{ for } t > 0. \tag{2.4}$$

If $v_1 = 0$, then no boundary conditions are imposed on p; however, if $v_1 > 0$, then we prescribe the boundary condition $p(0,t) = \frac{k_2}{k_1}$ for $t > 0$.

We shall require that $q_0(s)$ is in C^4 $(0 \le s < \infty)$ and that

$$\begin{aligned} &0 \le q_0(s) \le 1 \text{ if } s > 0, \ q_0(s) = 0 \text{ if } s > A_0 \text{ for some } A_0 < \infty, \\ &q_0(0) = 1, \ \partial_s^j q_0(s)|_{s=0} = 0 \text{ if } 1 \le j \le 4. \end{aligned} \tag{2.5}$$

The last two conditions ensure that the initial and boundary data fit at $(0,0)$ up to order 4. Note that (2.5) implies that $0 \le q_0(s) \le 1$.

It will be convenient to first deal with the case

$$v_1 = 0. \tag{2.6}$$

However, in order not to repeat some of the calculations, we shall perform these calculations for general v_1.

By standard ODE arguments we deduce that the solution (p,q) is continuously differentiable in $\bar{D}$, that $(\partial_x p, \partial_x q)$ satisfy the same system as (p,q), and that

$$k_1 \partial_x p(x,0) = k_2 \partial_x q(x,0) = k_2 \partial_x q_0\left(\frac{x}{\sqrt{\varepsilon}}\right), \ 0 < x < \infty;$$

furthermore, if $v_1 = 0$, then

$$\varepsilon v_2 \partial_x q(0,t) = k_1 p(0,t) - k_2 q(0,t) \to 0 \text{ if } t \to 0.$$

Since $\partial_x q(x,0) \to 0$ as $x \to 0$, the initial-boundary data are continuous at $(0,0)$. It follows, as before, that the second order derivatives of p, q are continuous in $\bar{D}$, and similarly one can prove that also the third order derivatives of p, q are continuous in $\bar{D}$.

LEMMA 2.1 (maximum principle). *The following inequalities hold:*

$$0 \le p \le \frac{k_2}{k_1},\ 0 \le q \le 1.$$

Proof. Let us first consider the system (2.1)–(2.4) in $D_L = D \cap \{x < L\}$, for $0 < L < \infty$. We modify (2.1), (2.2) by subtracting a small positive number μ from the right-hand sides, and we denote the corresponding solution by (p_μ, q_μ). We claim that

$$p_\mu < \frac{k_2}{k_1}(1+\mu),\ q_\mu < 1+\mu. \tag{2.7}$$

Indeed, if this is not true, then consider a point (x_0, t_0) with the smallest t_0 for which equality occurs in (2.7) either for p_μ or for q_μ. Suppose, for definiteness, that $q_\mu(x_0, t_0) = 1+\mu$. Because of the boundary and initial conditions, $x_0 \neq 0$ and $t_0 \neq 0$. At (x_0, t_0) we have

$$\varepsilon(\partial_t + v_2\partial_x)q_\mu = k_1 p_\mu - k_2 q_\mu - \mu \le -\mu < 0.$$

Thus, there would be an earlier time at which $q_\mu \ge 1+\mu$, which is a contradiction. Taking $\mu \to 0$ and then $L \to \infty$ we obtain the inequalities $p \le \frac{k_2}{k_1},\ q \le 1$. Similarly one can prove, by taking $\mu < 0$, that $p \ge 0,\ q \ge 0$. ∎

We now proceed to derive the asymptotic behavior for $q(x,t)$ as $\varepsilon \to 0$. We begin by deriving a second order PDE for q. From (2.1), (2.2) we obtain

$$\varepsilon(\partial_t + v_1\partial_x)(\partial_t + v_2\partial_x)q + (k_1+k_2)\partial_t q - (k_1 v_2 + k_2 v_1)\partial_x q = 0. \tag{2.8}$$

Introduce a change of variables

$$s = \frac{x - vt}{\sqrt{\varepsilon}},\ Q_\varepsilon(s,t) = q(x,t),$$

where v will be determined later on. The domain D is transformed into the domain

$$\Omega = \left\{(s,t) : -\frac{vt}{\sqrt{\varepsilon}} < s < \infty,\ t > 0\right\}.$$

Later on we shall use the notation

$$\begin{aligned}
&\Omega_T = \Omega \cap \{t < T\},\\
&\Gamma_T \equiv \Gamma_{T,\varepsilon} = \left\{(s,t) : s = -\frac{vt}{\sqrt{\varepsilon}}, 0 < t < T\right\},\ \text{where } 0 < T \le \infty,\\
&L_0 = \{(s,0) : 0 < s < \infty\},\\
&s_t \equiv s_{t,\varepsilon} = -\frac{vt}{\sqrt{\varepsilon}}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\partial_t q = \partial_t Q_\varepsilon - \frac{v}{\sqrt{\varepsilon}}\partial_s Q_\varepsilon,\ \partial_x q = \frac{1}{\sqrt{\varepsilon}}\partial_s Q_\varepsilon,\ \text{and}\\
&\partial_t Q_\varepsilon = \partial_t q + v\partial_x q,\ \partial_s Q_\varepsilon = \sqrt{\varepsilon}\partial_x q.
\end{aligned} \tag{2.9}$$

It is easily seen that if we choose

$$v = \frac{k_1 v_2 + k_2 v_1}{k}, \text{ where } k = k_1 + k_2, \tag{2.10}$$

then we obtain the following equation for Q_ε:

$$\partial_t Q_\varepsilon - \sigma^2 \partial_s^2 Q_\varepsilon + \partial_{ts} K\sqrt{\varepsilon} Q_\varepsilon + \frac{\varepsilon}{k}\partial_t^2 Q_\varepsilon = 0 \text{ in } \Omega, \tag{2.11}$$

where

$$\sigma^2 = \frac{k_1 k_2}{k^3}(v_2 - v_1)^2, \; K = \frac{k_2 - k_1}{k^2}(v_2 - v_1). \tag{2.12}$$

Continuing with the assumption (2.6), we also have

$$Q_\varepsilon(s,0) = q_0(s), \; s > 0, \tag{2.13}$$

and, by (2.2), (2.3), (2.9), and (2.10),

$$\partial_t Q_\varepsilon(s,0) = -\frac{k_2(v_2 - v_1)}{k}\frac{1}{\sqrt{\varepsilon}}\partial_s q_0(s), \; s > 0. \tag{2.14}$$

Since $q(0,t) = 1$ the function $p(0,t) \equiv k_2/k_1$ is the solution of (2.1) at $x = 0$, so that also

$$\varepsilon \partial_x q(0,t) = \frac{\varepsilon}{v_2}(\partial_t q + v_2 \partial_x q) = \frac{1}{v_2}(k_1 p - k_2 q) = 0 \text{ at } (0,t).$$

Hence

$$Q_\varepsilon = 1, \partial_s Q_\varepsilon = \partial_t Q_\varepsilon = 0 \text{ on } \Gamma_T \text{ for any } T > 0. \tag{2.15}$$

If W, ϕ are bounded continuously differentiable functions in $\bar{\Omega}_T$, $T < \infty$, and $\phi(s,T) \equiv 0$, then, by integration by parts,

$$\iint_{\Omega_T} \partial_t W \cdot \phi ds dt = -\int_{\Gamma_T \cup L_0} W\phi ds - \iint_{\Omega_T} W \cdot \partial_t \phi ds dt,$$

where in the integrals along Γ_T and L_0 the variable s is increasing from left to right. Similarly, if $\phi(s,t)$ or $W(s,t)$ converges to zero as $s \to \infty$, $0 < t < T$, then

$$\iint_{\Omega_T} \partial_s W \cdot \phi ds dt = -\int_{\Gamma_T} W\phi dt - \iint_{\Omega_T} W \cdot \partial_s \phi ds dt,$$

where in the integral along Γ_T the variable t is increasing from $t = 0$ to $t = T$. With the above understanding of the integrals along Γ_T, we have

$$\int_{\Gamma_T} W\phi dt = \int_{\Gamma_T} \frac{\sqrt{\varepsilon}}{v} W\phi ds. \tag{2.16}$$

Let $\tilde{Q}_\varepsilon$ denote the unique bounded solution of the parabolic equation

$$\partial_t \tilde{Q}_\varepsilon - \sigma^2 \partial_s^2 \tilde{Q}_\varepsilon = 0 \text{ in } \Omega, \tag{2.17}$$

with the initial and boundary conditions

$$\tilde{Q}_\varepsilon(s,0) = q_0(s) \text{ on } L_0, \qquad \tilde{Q}_\varepsilon \equiv 1 \text{ on } \Gamma_\infty. \tag{2.18}$$

Recalling that $\partial^j q_0(s) = 0$ at $s = 0$ for $1 \le j \le 4$, one can easily verify that $\tilde{Q}_\varepsilon$ satisfies the consistency conditions of order 2 at $(s,t) = (0,0)$. Hence all the derivatives

$$\partial_t^k \partial_s^l \tilde{Q}_\varepsilon \quad (0 \le 2k + l \le 4) \quad \text{are continuous at } (0,0). \tag{2.19}$$

In the next two lemmas we prove that these functions are uniformly bounded in Ω, and that they converge uniformly to zero as $s \to \infty$, the uniformity being with respect to ε.

LEMMA 2.2. *The following inequalities hold:*

$$|\partial_t^k \partial_s^l \tilde{Q}_\varepsilon(s,t)| \le C \text{ in } \Omega \quad (0 \le 2k + l \le 4), \tag{2.20}$$

where C is a constant independent of ε.

Proof. Consider the function

$$U(x,t) = \tilde{Q}_\varepsilon(s,t), \text{ where } s = \frac{x - vt}{\sqrt{\varepsilon}}.$$

Since $\partial_t \tilde{Q}_\varepsilon = \partial_t U + v \partial_x U$, there holds

$$\partial_t U = \alpha \varepsilon \partial_x^2 U - \beta \partial_x U \text{ in } D, \tag{2.21}$$

where α, β are positive numbers depending only on k_i, v_i, and

$$U(x,0) = q_0\left(\frac{x}{\sqrt{\varepsilon}}\right), \quad x > 0, \tag{2.22}$$

$$U(0,t) = 1, \quad t > 0. \tag{2.23}$$

The function $V = \partial_x U$ satisfies the heat equation (2.21) and, since $\partial_t U(0,t) = 0$,

$$\partial_x V - \frac{\gamma}{\varepsilon} V = 0 \text{ at } x = 0,\ t > 0, \quad \text{where } \gamma = \frac{\beta}{\alpha}.$$

Hence V cannot take positive maximum or negative minimum at $(0,t),\ t > 0$. By the maximum principle we then deduce that

$$|\partial_x U| \le \frac{1}{\sqrt{\varepsilon}} \sup_{x>0} \left| q_0'\left(\frac{x}{\sqrt{\varepsilon}}\right)\right| \le \frac{C}{\sqrt{\varepsilon}}$$

on the set $\{(x,t) :\ x > 0,\ t > 0\}$.

We proceed to apply the above argument to the function $V_1 = \partial_t \partial_x U$, which is again a solution of the heat equation (2.21), with

$$\partial_x V_1 - \frac{\gamma}{\varepsilon} V_1 = \frac{1}{\alpha\varepsilon} \partial_t^2 U = 0 \text{ at } x = 0,\ t > 0.$$

By the maximum principle we then get

$$\begin{aligned} |\partial_x V_1|_{L^\infty(\mathbb{R}_2^+)} &\le \sup_{x>0} |V_1(x,0)| \\ &= \sup_{x>0} \left| \alpha\varepsilon \partial_x^3 q_0\left(\frac{x}{\sqrt{\varepsilon}}\right) - \beta \partial_x^2 q_0\left(\frac{x}{\sqrt{\varepsilon}}\right)\right| \\ &\le \frac{C}{\varepsilon}. \end{aligned}$$

Hence

$$\left|\partial_x(\partial_x^2 U) - \frac{\gamma}{\varepsilon}(\partial_x^2 U)\right| = \frac{|V_1|}{\alpha\varepsilon} \le \frac{C}{\varepsilon^2}.$$

Multiplying by $e^{-\gamma x/\varepsilon}$ and integrating in x over (x,∞) we get

$$|e^{-\gamma x/\varepsilon}\partial_x^2 U(x,t)| \le \frac{C}{\varepsilon}e^{-\gamma x/\varepsilon}.$$

Hence

$$|\partial_x^2 U(x,t)|_{L^\infty(\mathbb{R}_2^+)} \le \frac{C}{\varepsilon}.$$

Similarly, working with the function $V_2 = \partial_t^2 \partial_x U$, we deduce, by the maximum principle, that

$$|V_2|_{L^\infty(\mathbb{R}_2^+)} \le \frac{C}{\varepsilon^{3/2}}.$$

Hence

$$\left|\partial_x(\partial_t\partial_x^2 U) - \frac{\gamma}{\varepsilon}(\partial_t\partial_x^2 U)\right| \le \frac{C}{\varepsilon^{5/2}}.$$

As before, we deduce by integration that

$$|\partial_t\partial_x^2 U(x,t)|_{L^\infty(\mathbb{R}_2^+)} \le \frac{C}{\varepsilon^{3/2}}.$$

This implies, by (2.21), that

$$\left|\partial_x(\partial_x^3 U) - \frac{\gamma}{\varepsilon}(\partial_x^3 U)\right| \le \frac{C}{\varepsilon^{5/2}}.$$

and, by integration, as before,

$$|\partial_x^3 U(x,t)|_{L^\infty(\mathbb{R}_2^+)} \le \frac{C}{\varepsilon^{3/2}}.$$

Similarly we can estimate $\partial_x^4 U$ by working with the function $V_3 = \partial_t^3 \partial_x U$. We conclude that

$$|\partial_x^j U(x,t)|_{L^\infty(\mathbb{R}_2^+)} \le \frac{C}{\varepsilon^{j/2}} \quad \text{for } j = 0,1,2,3,4.$$

It follows that

$$|\partial_s^j \tilde{Q}_\varepsilon(s,t)| \le C \text{ in } \Omega.$$

The remaining inequalities in (2.20) follow from the differential equation (2.17). ∎

LEMMA 2.3. *For any $0 < T < \infty$, the following inequalities hold for some $\gamma > 0$:*

$$|\partial_t^k \partial_s^l \tilde{Q}_\varepsilon(s,t)| \le Ce^{-\gamma s^2} \text{ in } \Omega_T \cap \{s > 0\} \quad (0 \le 2k + l \le 4), \tag{2.24}$$

where C is a constant independent of ε.

Proof. Recall that $\tilde{Q}_\varepsilon(s,0) = 0$ if $s > A_0$, so that also $\partial_s^l \tilde{Q}_\varepsilon(s,0) = 0$ if $s > A_0$. We can then compare $\partial_s^l \tilde{Q}_\varepsilon$ with a solution of (2.17) of the form

$$\frac{C}{\sqrt{t}} e^{-\delta s^2/t} \text{ in } G \equiv \{s > A_0,\ t > 0\} \left(\text{where } \delta = \frac{1}{4\sigma^2}\right).$$

The difference

$$V = \frac{C}{\sqrt{t}} e^{-\delta s^2/t} - \partial_s^l \tilde{Q}_\varepsilon(s,t)$$

is a bounded solution of (2.17) in G, with $V(s,0) = 0$ if $s > A_0$ and $V(A_0,t) > 0$ (< 0) if C is positive and large (negative and large in absolute value). By the maximum principle for parabolic equations in an unbounded domain (see, e.g., [7, Chap. 2, Thm. 9]), we conclude that $V > 0$ $(V < 0)$ in G in the case where C was positive (negative), so that the inequality (2.24) with $k = 0$ follows. The remaining estimates follow from (2.17). □

Consider the function

$$W = Q_\varepsilon - \tilde{Q}_\varepsilon \text{ in } \Omega_T,\ 0 < T < \infty.$$

It satisfies the equation

$$\partial_t W - \sigma^2 \partial_s^2 W + K\sqrt{\varepsilon}\partial_{ts}^2 W + \frac{\varepsilon}{k}\partial_s^2 W = F_\varepsilon, \tag{2.25}$$

where

$$F_\varepsilon = -K\sqrt{\varepsilon}\partial_{ts}^2 \tilde{Q}_\varepsilon - \frac{\varepsilon}{k}\partial_t^2 \tilde{Q}_\varepsilon. \tag{2.26}$$

Clearly

$$W \equiv 0 \text{ on } L_0 \cup \Gamma_T \tag{2.27}$$

and, by (2.9), (2.15),

$$\partial_t W = -\partial_t \tilde{Q}_\varepsilon,\ \partial_s W = -\partial_s \tilde{Q}_\varepsilon = -\frac{\sqrt{\varepsilon}}{v}\partial_t \tilde{Q}_\varepsilon \text{ on } \Gamma_T. \tag{2.28}$$

Also, by (2.14),

$$\partial_t Q_\varepsilon = \partial_t q + v\partial_x q = \frac{v - v_2}{\sqrt{\varepsilon}} q_0'(s) \text{ on } L_0,$$

so that

$$\partial_t W = -\frac{k_2(v_2 - v_1)}{k\sqrt{\varepsilon}} q_0'(s) \text{ on } L_0. \tag{2.29}$$

From Lemmas 2.2 and 2.3 we deduce that

$$\iint_{\Omega_T} F_\varepsilon^2 \le \iint_{\Omega_T \cap \{s<0\}} C_0\varepsilon + \iint_{\Omega_T \cap \{s>0\}} C_0\varepsilon e^{-2\gamma s^2} \le C\sqrt{\varepsilon},$$

where C_0, C are constants (independent of ε). Hence, for any function $Z \in L^2(\Omega_T)$ and any small $\eta > 0$,

$$\left|\iint_{\Omega_T} F_\varepsilon Z\right| \le \eta \iint_{\Omega_T} Z^2 + C\sqrt{\varepsilon}, \tag{2.30}$$

where C is a constant which depends on η and T.

LEMMA 2.4. *For any $0 < T < \infty$, there exists a constant $C = C(T)$ such that*

$$\iint_{\Omega_T} W_t^2 ds dt + \sup_{0\le t\le T} \int_{s_t}^{\infty} W_s^2(s,t) ds \le C, \tag{2.31}$$

$$\iint_{\Omega_T} (W^2 + W_s^2) ds dt \le C\sqrt{\varepsilon}. \tag{2.32}$$

Proof. If we multiply (2.25) by W_t and integrate over Ω_T, we get

$$\iint_{\Omega_T} W_t^2 - \sigma^2 \iint_{\Omega_T} W_t W_{ss} + K\sqrt{\varepsilon} \iint_{\Omega_T} W_t W_{ts} + \frac{\varepsilon}{k} \iint_{\Omega_T} W_t W_{tt} = -\iint_{\Omega_T} W_t F_\varepsilon. \tag{2.33}$$

We proceed to evaluate terms on the left-hand side of (2.33). By integration by parts we obtain

$$\begin{aligned} -\iint_{\Omega_T} W_t W_{ss} &= -\iint_{\Omega_T} [\partial_s(W_t W_s) - W_{ts} W_s] \\ &= \int_{\Gamma_T} W_t W_s dt + \frac{1}{2}\int_{s_T}^{\infty} W_s^2(s,T) ds - \frac{1}{2}\int_{L_0\cup\Gamma_T} W_s^2 ds, \end{aligned}$$

and, by (2.28),

$$\int_{\Gamma_T} W_t W_s dt = \frac{\sqrt{\varepsilon}}{v} \int_{\Gamma_T} \tilde{Q}_{\varepsilon,t}^2 dt$$

and

$$\int_{L_0\cup\Gamma_T} W_s^2 ds = \int_{\Gamma_T} W_s^2 ds = \int_{\Gamma_T} \tilde{Q}_{\varepsilon,s}^2 ds = \frac{\sqrt{\varepsilon}}{v} \int_{\Gamma_T} \tilde{Q}_{\varepsilon,t}^2 dt,$$

where in the last equation we used also (2.16). Hence,

$$-\sigma^2 \iint_{\Omega_T} W_t W_{ss} = \frac{1}{2}\sigma^2 \int_{s_T}^{\infty} W_s^2(s,T) ds - \frac{1}{2}\sigma^2 \frac{\sqrt{\varepsilon}}{v} \int_{\Gamma_T} \tilde{Q}_{\varepsilon,t}^2 dt. \tag{2.34}$$

Next,

$$\begin{aligned} K\sqrt{\varepsilon} \iint_{\Omega_T} W_t W_{ts} &= \frac{1}{2} K\sqrt{\varepsilon} \iint_{\Omega_T} \partial_s(W_t^2) \\ &= -\frac{1}{2} K\sqrt{\varepsilon} \int_{\Gamma_T} W_t^2 dt = -\frac{1}{2} K\sqrt{\varepsilon} \int_{\Gamma_T} \tilde{Q}_{\varepsilon,t}^2 dt. \end{aligned} \tag{2.35}$$

Consider finally

$$\varepsilon \iint_{\Omega_T} W_t W_{tt} = \frac{\varepsilon}{2} \int_{s_T}^{\infty} W_t^2(s,T) ds - \frac{\varepsilon}{2} \int_{L_0\cup\Gamma_T} W_t^2 ds.$$

Since

$$-\frac{\varepsilon}{2} \int_{\Gamma_T} W_t^2 ds = -\frac{\varepsilon}{2} \int_{\Gamma_T} \tilde{Q}_{\varepsilon,t}^2 ds = -\frac{\sqrt{\varepsilon}}{2v} \int_{\Gamma_T} \tilde{Q}_{\varepsilon,t}^2 dt,$$

and

$$-\frac{\varepsilon}{2}\int_{L_0} W_t^2 ds = \mathcal{O}(1)$$

by (2.29) and the fact that (by (2.17), (2.18)) $\tilde{Q}_{\varepsilon,t} = 0$ on $L_0 \cap \{s > A_0\}$, we obtain

$$\varepsilon \iint_{\Omega_T} W_t W_{tt} = \frac{\varepsilon}{2}\int_{s_T}^{\infty} W_t^2(s,T)ds - \frac{\sqrt{\varepsilon}}{2v}\int_{\Gamma_T} \tilde{Q}_{\varepsilon,t}^2 dt + \mathcal{O}(1). \tag{2.36}$$

Substituting (2.34)–(2.36) into (2.33) and using also (2.30) with $Z = W_t$, $\eta = \frac{1}{2}$, and the inequality $|\tilde{Q}_{\varepsilon,t}| < C$ on Γ_T, we obtain

$$\iint_{\Omega_T} W_t^2 + \frac{1}{2}\sigma^2\int_{s_T}^{\infty} W_s^2(s,T)ds + \frac{\varepsilon}{k}\int_{s_T}^{\infty} W_t^2(s,T)ds \le C\sqrt{\varepsilon}\int_{\Gamma_T} dt + \mathcal{O}(1).$$

This implies $\iint_{\Omega_T} W_t^2 < C_1(T)$, and $\int_{s_t}^{\infty} W_s^2(s,t)ds < C_2(t)$ for $t \in [0,T]$ and some continuous function C_2, and (2.31) follows, with $C(T) = C_1(T) + \sup_{0\le t\le T} C_2(t)$.

To prove (2.32) we multiply (2.25) by W and integrate over Ω_T. By integration by parts we obtain

$$\begin{aligned}\frac{1}{2}\int_{s_T}^{\infty} W^2(s,T)ds + \sigma^2\iint_{\Omega_T} W_s^2 - K\sqrt{\varepsilon}\iint_{\Omega_T} W_s W_t \\ + \frac{\varepsilon}{k}\int_{s_T}^{\infty}(WW_t)(s,T)ds - \frac{\varepsilon}{k}\iint_{\Omega_T} W_t^2 = \iint_{\Omega_T} WF_\varepsilon.\end{aligned} \tag{2.37}$$

By Lemma 2.3 and (2.31),

$$\left|K\sqrt{\varepsilon}\iint_{\Omega_T} W_s W_t\right| \le \frac{1}{2}\sigma^2\iint_{\Omega_T} W_s^2 + \frac{K^2\varepsilon}{2\sigma^2}\iint_{\Omega_T} W_t^2 \le \frac{1}{2}\sigma^2\iint_{\Omega_T} W_s^2 + C\varepsilon.$$

Also, by (2.30),

$$\left|\iint_{\Omega_T} WF_\varepsilon\right| \le \eta\iint_{\Omega_T} W^2 + C\sqrt{\varepsilon}, \quad C = C(\eta).$$

Hence

$$\begin{aligned}\frac{1}{2}\int_{s_T}^{\infty} W^2(s,T)ds + \frac{1}{2}\sigma^2\iint_{\Omega_T} W_s^2 + \frac{\varepsilon}{2k}\int_{s_T}^{\infty}(W^2)_t(s,T)ds \\ \le \eta\iint_{\Omega_T} W^2 + C\sqrt{\varepsilon}.\end{aligned} \tag{2.38}$$

Integrating both sides with respect to T, $0 < T < T_0$, and choosing $\eta = \frac{1}{4T_0}$, we obtain

$$\iint_{\Omega_{T_0}} W^2 + \int_0^{T_0}\left(\iint_{\Omega_T} W_s^2\right)dT + \varepsilon\int_{s_{T_0}}^{\infty} W^2(s,T_0)ds \le C\sqrt{\varepsilon}, \tag{2.39}$$

where C depends on T_0. Finally, if we use the inequality

$$\int_0^{T_0}\left(\int_0^T f(t)dt\right)dT = \int_0^{T_0}(T_0 - t)f(t)dt \ge \delta_0\int_0^{T_0-\delta_0} f(t)dt$$

with $f(t) = \int_{s_t}^{\infty} W_s^2(s,t)ds$ in (2.39), we obtain the estimate (2.32). ∎

We extend the function $\tilde{Q}_\varepsilon(s,t)$ by 1 into the domain $\{s < -vt/\sqrt{\varepsilon},\ t > 0\}$, and wish to estimate $\tilde{Q}_\varepsilon - Q_0$, where Q_0 is the bounded solution of

$$\partial_t Q_0 - \sigma^2 \partial_s^2 Q_0 = 0 \text{ in } \mathbb{R}^2_+, \tag{2.40}$$

$$Q_0(s,0) = q_0(s) \text{ if } s > 0, \quad Q_0(s,0) = 1 \text{ if } s < 0. \tag{2.41}$$

LEMMA 2.5. *The following inequality holds:*

$$\int_{-\infty}^{\infty} (\tilde{Q}_\varepsilon - Q_0)^2(s,T)ds + \int_0^T \int_{-\infty}^{\infty} (\partial_s \tilde{Q}_\varepsilon - \partial_s Q_0)^2 ds dt \le C\sqrt{\varepsilon}, \tag{2.42}$$

where $0<T<\infty$ and C is a constant independent of T and ε.

Proof. We first estimate $\tilde{Q}_\varepsilon - Q_0$ on Γ_T. For this purpose we represent the function $V \equiv 1 - Q_0$ in the form

$$V(s,t) = \int_0^{\infty} \frac{e^{-\frac{(s-\zeta)^2}{4\sigma^2 t}}}{\sigma\sqrt{4\pi t}} \phi(\zeta) d\zeta,\ \ \phi(\zeta) = 1 - q_0(\zeta)$$

and compute

$$V\left(-\frac{vt}{\sqrt{\varepsilon}}, t\right) = \int_0^{\infty} \frac{e^{-(-\frac{vt}{\sqrt{\varepsilon}} - \zeta)^2/(4\sigma^2 t)}}{\sigma\sqrt{4\pi t}} \phi(\zeta) d\zeta.$$

Substituting

$$\xi = \left(\frac{vt}{\sqrt{\varepsilon}} + \zeta\right) \frac{1}{\sqrt{t}}$$

we obtain

$$\left| V\left(-\frac{vt}{\sqrt{\varepsilon}}, t\right) \right| \le C \int_{v\sqrt{t/\varepsilon}} e^{-\frac{\xi^2}{4\sigma^2}} d\xi \le C e^{-\frac{\alpha t}{\varepsilon}} \quad (\text{for some } \alpha > 0).$$

Similarly we obtain

$$\left| \partial_s V\left(-\frac{vt}{\sqrt{\varepsilon}}, t\right) \right| \le (C/\sqrt{t}) e^{-\frac{\alpha t}{\varepsilon}}.$$

Consider the function $R = \tilde{Q}_\varepsilon - Q_0\varepsilon$. It satisfies the equation

$$\partial_t R - \sigma^2 \partial_s^2 R = 0 \text{ in } \Omega_T, \tag{2.43}$$

and, by the last two estimates,

$$|R|_{\Gamma_T} \le C e^{-\frac{\alpha t}{\varepsilon}}, \quad |\partial_s R|_{\Gamma_T} \le \frac{C}{\sqrt{t}} e^{-\frac{\alpha t}{\varepsilon}}. \tag{2.44}$$

If we multiply (2.43) by R and integrate over Ω_T, we obtain, after integration by parts,

$$\frac{1}{2}\int_{s_T}^{\infty} R^2(s,T)ds - \sigma^2 \iint_{\Omega_T} (\partial_s R)^2 = \frac{1}{2}\int_{\Gamma_T} R^2 ds - \sigma^2 \int_{\Gamma_T} R \cdot \partial_s R dt.$$

Using (2.44) we find that each of the two integrals on the right-hand side is bounded by $C\sqrt{\varepsilon}$, and thus we derive the estimate (2.42). ∎

Combining the estimates (2.32), (2.42) we obtain the following theorem.

THEOREM 2.6. *Consider the system* (2.1)–(2.4) *under the assumptions* (2.5) *and* (2.6). *Then we can write* q *in the form*

$$q(x,t) = Q_\varepsilon\left(\frac{x-vt}{\sqrt{\varepsilon}}, t\right), \tag{2.45}$$

where $Q_\varepsilon(s,t)$ *(extended by* 1 *for* $s < -vt/\sqrt{\varepsilon}$*) converges to the solution* $Q_0(s,t)$ *of* (2.40), (2.41) *as* $\varepsilon \to 0$ *in the following sense:*

$$\sup_{0\le t\le T}\int_{-\infty}^{\infty}(Q_\varepsilon - Q_0)^2(s,t)ds + \int_0^T\int_{-\infty}^{\infty}(\partial_s Q_\varepsilon - \partial_s Q_0)^2 dsdt \le C\sqrt{\varepsilon} \tag{2.46}$$

for any $0 < T < \infty$, *where* C *is a constant which may depend on* T.

Remark 2.1. The proof of Theorem 2.6 extends to the case where the first two conditions on $q_0(s)$ are replaced by the weaker condition that $\partial_s^j q_0(s)$ are continuous functions for $s \ge 0$ and they belong to $L^2(\mathbb{R})$ for $0 \le j \le 4$.

Remark 2.2. Theorem 2.6 extends, with essentially the same proof, to the case where $v_1 > 0$, provided $v_1 \ne v_2$ and we prescribe the boundary condition $p(x,t) \equiv \frac{k_2}{k_1}$ at $x = 0$.

Remark 2.3. Consider the case where $q_0 \equiv 0$, and take for simplicity $v_1 = 0$. Then the initial and boundary data form a function which is discontinuous at $(0,0)$, so that the proof of Theorem 2.6 cannot be extended to this case. If we introduce an approximating system by changing the initial data,

$$q(x,0) = q_{0\delta}\left(\frac{x}{\sqrt{\varepsilon}}\right), \quad q_{0\delta}(0) = 1,\ q'_{0\delta} \le 0,\ q_{0\delta}(s) = 0 \text{ if } s \ge \delta, \tag{2.47}$$

then for the corresponding solution (p_δ, q_δ) we have the following result.

THEOREM 2.7. *The following inequality holds:*

$$\sup_{T>0}\int_0^\infty [(p_\delta(x,T) - p(x,T))^2 + (q_\delta(x,T) - q(x,T))^2]dx \le C\delta, \tag{2.48}$$

where C *is a constant independent of the function* $q_{0\delta}$.

Proof. Set $\tilde{p} = p_\delta - p,\ \tilde{q} = q_\delta - q$. Then

$$\partial_t\tilde{p} = -k_1\tilde{p} + k_2\tilde{q},\ (\partial_t + v_1\partial_x)\tilde{q} = k_1\tilde{p} - k_2\tilde{q}.$$

Multiplying the first equation by $k_2\tilde{p}$ and the second equation by $k_1\tilde{q}$ and adding, we obtain the inequality

$$k_2\partial_t\tilde{p}^2 + k_1((\partial_t + v_1\partial_x)\tilde{q})^2 \le 0.$$

Integrating over $0 < x < \infty,\ 0 < t < T$, the inequality (2.48) easily follows. ∎

Theorem 2.7 combined with Theorem 2.6 suggests that, if $q_0(s) \equiv 0$, then $q(x,t) - Q_0(\frac{x-vt}{\sqrt{\varepsilon}}, t)$ converges to zero in some sense when Q_0 is the solution of (2.40), (2.41) with $q_0(s) \equiv 0$. This situation will be considered, for more general dynamical systems, in section 3.

3. The case $n > 2$. Let $K = (k_{ij})$ be an $n \times n$ matrix satisfying the following conditions:

$$k_{ij} \geq 0 \text{ if } i \neq j. \tag{3.1}$$

$$\sum_{i=1}^{n} k_{ij} = 0 \left(\text{so that } k_{jj} = -\sum_{\substack{i=1 \\ i\neq j}}^{n} k_{ij}\right). \tag{3.2}$$

$$\text{For any indices } i_0 \neq i_1 \text{ there are } j_1, j_2, \dots, j_m \text{ such that} \\ j_1 = i_0,\ j_m = i_1, \text{ and } k_{j_l j_{l+1}} > 0 \text{ for } l = 1, \dots, m-1. \tag{3.3}$$

As proved in [11], under conditions (3.1)–(3.3), the null space of the matrix (k_{ij}) is one-dimensional, and it is generated by a vector $(\lambda_1, \dots, \lambda_n)$ with positive components. For simplicity we take $\sum_{j=1}^{n} \lambda_j = 1$. For later reference we write

$$\sum_{j=1}^{n} k_{ij}\lambda_j = 0 \text{ for } 1 \leq i \leq n. \tag{3.4}$$

In this section we consider a collection of populations undergoing transitions

$$p_i \underset{k_{ji}}{\overset{k_{ij}}{\rightleftharpoons}} p_j \text{ for } i, j = 1, \dots, n$$

with the dynamics given by (1.2), in the special case (1.4), or, more generally, for the case when $v_j \geq 0$ for all j, and

$$p_i(x,0) = \lambda_i q_0\left(\frac{x}{\sqrt{\varepsilon}}\right),\ 1 \leq i \leq n, \\ p_j(0,t) = \lambda_j \text{ if } v_j > 0. \tag{3.5}$$

Here we have the following:

$$\text{either } q_0 \equiv 0, \text{ or } q_0(s) \text{ is of class } C^n, \text{ has compact} \\ \text{support, } q_0(0) = 1, \text{ and } q_0^{(k)}(0) = 0 \text{ for all } k = 1, \dots, n. \tag{3.6}$$

THEOREM 3.1. *Let the matrix (k_{ij}) satisfy conditions* (3.1)–(3.3)*, and consider the system* (1.2) *with $v_j \geq 0$ for all $1 \leq j \leq n$, $v_n > 0$, and with the initial and boundary conditions* (3.5)*, where q_0 is a function satisfying* (3.6)*. Then, for some constants $v > 0$ and $\sigma^2 > 0$, the following holds:*

$$p_j(x,t) = Q_{\varepsilon,j}\left(\frac{x - vt}{\sqrt{\varepsilon}}, t\right) \quad (1 \leq j \leq n), \tag{3.7}$$

where, as $\varepsilon \to 0$,

$$Q_{\varepsilon,j} \to \lambda_j Q_0 \text{ weakly in } L^r(\mathbb{R}^2_+) \tag{3.8}$$

for any $1 < r < \infty$, and $Q_0(s,t)$ is the bounded solution of (2.40), (2.41).

Proof. For clarity we first prove the theorem for $j = n$, $q_0(0) = 1, q_0 \in C^n$, and set $Q_{\varepsilon,j} = Q_\varepsilon$. As in section 2 one can prove that the functions $p_j(x,t)$ belong to

$C^n(\bar{D})$. Then, by algebraic elimination, as in [5, p. 14], it follows that each of the functions p_j satisfies the equation

$$\det\left(\lambda_{ij} - \delta_{ij}\varepsilon(\partial_t + v_j\partial_x)\right) w = 0. \tag{3.9}$$

We shall henceforth use this equation for the function $p_n(x,t)$. Introducing the function Q_ε as in (3.7), the calculations in [9], [11] show that

$$\partial_t Q_\varepsilon - \sigma^2\partial_s^2 Q_\varepsilon = P(\partial_s, \partial_t)Q_\varepsilon, \tag{3.10}$$

where $v = \sum_{j=1}^n \lambda_j v_j$, $-\sigma^2 = \frac{a_{02}}{a_{10}}$, where a_{kl} are defined by $\det(k_{ij} - \gamma\delta_{ij} - \lambda(v_i - v)\delta_{ij}) = \sum_{0\le k+l\le n} a_{kl}\gamma^k\lambda^l$, and

$$P(\partial_s, \partial_t) = \sum_{l+k\le n} \beta_{lk}\varepsilon^{l+\frac{k}{2}-1}\partial_t^l\partial_s^k, \tag{3.11}$$

where the β_{lk} are constants depending only on k_{ij} and v_j.

Set $p_i(x,t) = P_i(\frac{x-vt}{\sqrt{\varepsilon}}, t)$. Then

$$\varepsilon\left(\partial_t + \frac{v}{\sqrt{\varepsilon}}\partial_s\right)P_i + \frac{v_i}{\sqrt{\varepsilon}}\partial_s P_i = \sum_{j=1}^n k_{ij}P_j. \tag{3.12}$$

LEMMA 3.2. *There holds*

$$\partial_t^l\partial_x^k p_j(0,t) = 0 \text{ for all } 0 \le k+l \le n,\ t > 0,$$

or, equivalently,

$$\partial_t^l\partial_s^k P_{j,\varepsilon}\left(\frac{-vt}{\sqrt{\varepsilon}}, t\right) = 0 \text{ for all } 0 \le k+l \le n,\ t > 0.$$

Proof. Let $A = \{i : v_i = 0\},\ B = \{i : v_i > 0\}$. We break (1.2) at $x = 0$ into two subsystems:

$$\partial_t p_i(0,t) = \sum_{j\in A} k_{ij}p_j(0,t) + \sum_{j\in B} k_{ij}p_j(0,t) \text{ for } i \in A, \tag{3.13}$$

$$\partial_t p_i(0,t) + v_i\partial_x p_i(0,t) = \sum_{j=1}^n k_{ij}p_j(0,t) \text{ for } i \in B. \tag{3.14}$$

Since $p_j(0,t) \equiv \lambda_j$ if $j \in B$, and $p_i(0,0) = \lambda_i$ if $i \in A$, using (3.4) we deduce that the unique solution of the ODE system (3.13) is $p_i(0,t) \equiv \lambda_i$ for all $i \in A$. Then $p_j(0,t) \equiv \lambda_j$ for all $j = 1, \dots, n$ and (3.14) gives

$$\partial_x p_i(0,t) = 0 \text{ for } i \in B. \tag{3.15}$$

We now apply ∂_x to (1.2) and deduce analogously to (3.13), (3.14) that

$$\partial_t\partial_x p_i(0,t) = \sum_{j\in A} k_{ij}\partial_x p_j(0,t) + \sum_{j\in B} k_{ij}\partial_x p_j(0,t) \text{ for } i \in A, \tag{3.16}$$

$$\partial_t\partial_x p_i(0,t) + v_i\partial_x\partial_x p_i(0,t) = \sum_{j=1}^n k_{ij}\partial_x p_j(0,t) \text{ for } i \in B. \tag{3.17}$$

Note that, by (3.15), we have $\sum_{j\in B} k_{ij}\partial_x p_j(0,t)=0$; also, $\partial_x p_i(0,0)=0$ for all $i\in A$. Then the unique solution of the ODE system (3.16) is $\partial_x p_i(0,0)\equiv 0$ for all $i\in A$, so that, upon recalling (3.15), $\partial_x p_i(0,0)\equiv 0$ for all $i=1,\dots,n$. By differentiating (3.15) in t we obtain $\partial_t\partial_x p_i(0,t)\equiv 0$, and using (3.17) we conclude that $\partial_x^2 p_i(0,t)\equiv 0$ for all $i\in B$.

In the same way, if we differentiate (1.2) twice in x, we can conclude that $\partial_x^2 p_j(0,t)\equiv 0$ for all $1\le j\le n$, and similarly $\partial_x^k p_j(0,t)\equiv 0$ for all $1\le k\le n$, $1\le j\le n$. This implies the statement of the lemma. ∎

LEMMA 3.3. *There holds*

$$\partial_t^l\partial_s^k P_{j,\varepsilon}(s,0)=O(\varepsilon^{\frac{1}{2}-l}) \text{ for all } l\ge 1,\ k+l\le n,\ s>0.$$

Proof. By (3.12), at $t=0$,

$$\partial_t P_m(s,0)=\frac{v-v_m}{\sqrt{\varepsilon}}\partial_s P_m(s,0)+\frac{1}{\varepsilon}\sum_{j=1}^n k_{mj}P_j(s,0)=\frac{v-v_m}{\sqrt{\varepsilon}}\lambda_m q_0'(s),$$

i.e., $\partial_t P_m(s,0)=C_1^1\varepsilon^{-\frac{1}{2}}q_0'(s)$ for some constant C_1^1. Next, applying ∂_t to (3.12), we obtain

$$\begin{aligned}\partial_t^2 P_m(s,0)&=\partial_t\left(\frac{v-v_m}{\sqrt{\varepsilon}}\partial_s P_m(s,0)+\frac{1}{\varepsilon}\sum_{j=1}^n k_{mj}P_j\right)(s,0)\\&=C_1^2\varepsilon^{-\frac{3}{2}}q_0'(s)+C_2^2\varepsilon^{-1}q_0''(s)\end{aligned}$$

for some constants C_1^2 and C_2^2. Similarly, for any $l\le n$ we obtain

$$\partial_t^l P_m(s,0)=\sum_{j=1}^l C_j^l\varepsilon^{l-\frac{j}{2}}q_0^{(j)}(s)$$

for some constants C_j^l, which implies the statement of the lemma. ∎

We return to the proof of Theorem 3.1 for p_n, in the case $q_0(0)=1, q_0\in C^n$. As in the proof of the maximum principle (Lemma 2.1) one can prove that

$$0\le p_j(x,t)\le\lambda_j. \tag{3.18}$$

We extend the function $Q_\varepsilon(s,t)$ by 1 into $\{-\infty<s<-\frac{vt}{\sqrt{\varepsilon}},t>0\}$. By (3.18) with $j=n$, any sequence $\varepsilon'\to 0$ has a subsequence $\varepsilon''\to 0$ such that $Q_{\varepsilon''}\to\bar{Q}$ in $L^r(\mathbb{R}_2^+)$ for any $0<r<\infty$, where $\mathbb{R}_2^+=\{(s,t)\in\mathbb{R}_2: t>0\}$

Take any smooth function ϕ with compact support $K\subset\mathbb{R}_2^+$. If we multiply ϕ by the left-hand side of (3.10) and perform integration by parts, we obtain

$$\iint_K\phi(\partial_t Q_\varepsilon-\sigma^2\partial_s^2 Q_\varepsilon)=\iint_K Q_\varepsilon(-\partial_t-\sigma^2\partial_s^2)\phi, \tag{3.19}$$

provided ε is sufficiently small so that K stays to the right of Γ_∞. Similarly, from the right-hand side of (3.10) we get

$$\iint_K\phi P(\partial_s,\partial_t)Q_\varepsilon=\iint_K Q_\varepsilon P^*(\partial_s,\partial_t)\phi, \tag{3.20}$$

where P^* is the adjoint of the differential operator P defined in (3.11). As $\varepsilon = \varepsilon'' \to 0$ the right-hand side of (3.20) converges to zero. Hence, by (3.10), the same is true of each of the two sides of (3.19), so that

$$\iint_K \bar{Q}(-\partial_t - \sigma^2 \partial_s^2)\phi = 0.$$

It follows that $\bar{Q}$ is a weak solution of the heat equation

$$\partial_t \bar{Q} - \sigma^2 \partial_s^2 \bar{Q} = 0 \text{ in } \mathbb{R}_2^+. \tag{3.21}$$

By regularity of weak solutions we conclude that $\bar{Q}$ is a smooth solution of (3.21).

Next let ϕ be a smooth function with compact support K in $\{(s,t) : s > 0, -1 < t < \infty\}$. If we multiply the right-hand side of (3.10) by ϕ and integrate by parts, we obtain

$$\iint_K Q_\varepsilon P^*(\partial_s, \partial_t)\phi - \int_{K \cap \{t=0\}} \phi \left(\sum_{l+k \le n} \beta_{lk} \varepsilon^{l+\frac{k}{2}-1} \partial_t^{l-1} \partial_s^k \right) Q_\varepsilon,$$

and by Lemma 3.3 the last integral converges to zero as $\varepsilon \to 0$; the first integral also converges to zero, as in the previous case. We conclude from (3.10) that

$$\iint_K \phi(\partial_t Q_\varepsilon - \sigma^2 \partial_s^2 Q_\varepsilon) \to 0 \text{ as } \varepsilon \to 0.$$

By integration by parts, the left-hand side is equal to

$$\iint_K Q_\varepsilon(-\partial_t - \sigma^2 \partial_s^2)\phi - \int_{K \cap \{t=0\}} \lambda_n q_0 \phi.$$

Hence, as $\varepsilon = \varepsilon'' \to 0$, we get

$$\iint_K \bar{Q}(-\partial_t - \sigma^2 \partial_s^2)\phi - \int_{K \cap \{t=0\}} \lambda_n q_0 \phi = 0.$$

This means that $\bar{Q}$ takes the initial data $\lambda_n q_0(s)$ in a weak sense and, by regularity results, also in the classical sense.

Finally, take a smooth function ϕ with compact support K in $\{(s,t) : s < 0, -1 < t < \infty\}$. We proceed as in the previous case, but with the function $R_\varepsilon = Q_\varepsilon - \lambda_n$. By Lemma 3.3, when integrating by parts, we do not get any boundary integrals on Γ_∞. Hence, after going to the limit with $\varepsilon = \varepsilon'' \to 0$, we find that $\bar{Q} - \lambda_n$ takes the initial value 0 on $\{(s,t) : s < 0, t = 0\}$. We have thus proved that $\bar{Q} = Q_0$, and this completes the proof of the theorem for p_n in the case $q_0(0) = 1, q_0 \in C^n$.

Consider next the case $q_0 \equiv 0$. Since the solution is not continuous, the preceding proof cannot be applied directly. Instead, we approximate the problem by introducing initial data as in Theorem 1, but with

$$q_0'(s) \le 0, \ q_0(s) \equiv 0 \text{ if } s > \delta.$$

We denote the corresponding solutions by $p_{j,\delta}$ and set $q_{\delta,\varepsilon} = p_{j,\delta}$. We introduce a function $Q_{\delta,\varepsilon}(s,t)$ by

$$q_{\delta,\varepsilon}(x,t) = Q_{\delta,\varepsilon}(s,t), \ s = \frac{x - vt}{\sqrt{\varepsilon}}.$$

Then any sequence $(\delta', \varepsilon') \to 0$ has a subsequence $(\delta'', \varepsilon'') \to 0$ such that $Q_{\delta'', \varepsilon''} \to \bar{Q}$ weakly in $L^r(\mathbb{R}^+_2)$, and, as before, $\bar{Q}$ coincides with the solution of the heat equation (1.7).

The proof of Theorem 3.1 for any p_j is the same as for p_n, since Lemmas 3.2 and 3.3 hold for any $1 \le j \le n$. □

REFERENCES

[1] J. J. Blum and M. C. Reed, *The transport of organelles in axons*, Math. Biosci., 90 (1988), pp. 233–245.

[2] J. J. Blum and M. C. Reed, *A model for slow axonal transport and its application to neurofilamentous neuropathies*, Cell Motility Cytoskeleton, 12 (1989), pp. 53–65.

[3] E. A. Brooks, *Probabilistic methods for a linear reaction-hyperbolic system with constant coefficients*, Ann. Appl. Probab., 9 (1999), pp. 719–731.

[4] A. Brown, L. Wang, and P. Jung, *Stochastic simulation of neurofilament transport in axons: The "stop and go" hypothesis*, Mol. Biol. Cell, 16 (2005), pp. 4243–4255.

[5] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2, John Wiley and Sons, 1962.

[6] G. Craciun, A. Brown, and A. Friedman, *A dynamical system model of neurofilament transport in axons*, J. Theoret. Biol., 237 (2005), pp. 316–322.

[7] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Hills, NJ, 1964.

[8] A. Friedman and G. Craciun, *A model of intracellular transport of particles in an axon*, J. Math. Biol., 51 (2005), pp. 217–246.

[9] M. Pinsky, *Differential equations with a small parameter and the central limit theorem for functions defined on a finite Markov chain*, Z. Wahrsch. Verw. Gebiete, 9 (1967), pp. 101–111.

[10] M. C. Reed and J. J. Blum, *Theoretical Analysis of radioactivity profiles during fast axonal transport: Effects of deposition and turnover*, Cell Motility Cytoskeleton, 6 (1986), pp. 620–627.

[11] M. C. Reed, S. Venakides, and J. J. Blum, *Approximate traveling waves in linear reaction-hyperbolic equations*, SIAM J. Appl. Math., 50 (1990), pp. 167–180.

[12] L. Wang and A. Brown, *Rapid intermittent movement of axonal neurofilaments observed by fluorescence photobleaching*, Mol. Biol. Cell, 12 (2001), pp. 3257–3267.

[13] L. Wang, C.-L. Ho, D. Sun, R. K. H. Liem, and A. Brown, *Rapid movement of axonal neurofilaments interrupted by prolonged pauses*, Nat. Cell Biol., 2 (2000), pp. 137–141.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 759–781

LINEAR PROBLEMS RELATED TO ASYMPTOTIC STABILITY OF SOLITONS OF THE GENERALIZED KdV EQUATIONS*

YVAN MARTEL†

Abstract. This paper is concerned with the classification of global and uniformly localized solutions of linear problems related to the dynamics of the generalized Korteweg–de Vries equations in a neighborhood of solitons. In this context, we give a simpler and more general proof of a classification result introduced by Martel and Merle [*J. Math. Pures Appl.*, 79 (2000), pp. 339–425], [*Arch. Ration. Mech. Anal.*, 157 (2001), pp. 219–254]. We also consider the case of a nonlinear equation similar to the Korteweg–de Vries equation which was introduced by Peregrine [*J. Fluid Mechanics*, 27 (1967), pp. 815–827] and Benjamin, Bona, and Mahony [*Philos. Trans. Roy. Soc. London Ser.*, 272 (1972), pp. 47–78].

Key words. generalized KdV equations, solitons, asymptotic stability

AMS subject classifications. Primary, 35Q53; Secondary, 35Q51, 35B40

DOI. 10.1137/050637510

1. Introduction. This paper is concerned with the classification of global and uniformly localized solutions of linear problems related to the dynamics of the generalized Korteweg–de Vries equations (gKdV equations) in a neighborhood of solitons. We also consider the case of a nonlinear dispersive equation similar to the KdV equation which was introduced by Peregrine [22] and Benjamin, Bona, and Mahony [2]; see below in this section and section 4.

It is well known that the gKdV equations

$$\partial_t \psi + \partial_x(\partial_x^2 \psi + \psi^p) = 0, \quad t, x \in \mathbb{R}, \tag{1}$$

for $p \geq 2$ integer, have explicit traveling wave solutions. Let

$$Q(x) = \left(\frac{p+1}{2\cosh^2\left(\frac{p-1}{2}x\right)} \right)^{\frac{1}{p-1}} \tag{2}$$

be the unique H^1 positive solution (up to translations) of

$$Q'' + Q^p = Q \quad \text{on } \mathbb{R}.$$

Then, for any $c > 0$, $x_0 \in \mathbb{R}$, the functions

$$R_{c,x_0}(t,x) = Q_c(x - x_0 - ct), \quad \text{where} \quad Q_c(x) = c^{\frac{1}{p-1}} Q(\sqrt{c}x),$$

are solutions of the gKdV equations (1). We call these solutions *solitons*.

Let us recall some general results concerning the solutions of (1).

First, we recall that the local Cauchy problem for (1) is well-posed in H^1 for any $p \geq 2$ integer. More precisely, Kenig, Ponce, and Vega [10] proved that for any

*Received by the editors August 4, 2005; accepted for publication (in revised form) March 8, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/sima/38-3/63751.html
†Mathématiques, Université de Versailles-Saint-Quentin-en-Yvelines, 45, av. des Etats-Unis, 78035 Versailles Cedex, France (martel@math.uvsq.fr).

$\psi_0 \in H^1(\mathbb{R})$ there exists a unique (in a suitable sense) maximal H^1 solution $\psi(t)$ of (1) satisfying $\psi(0) = \psi_0$. Moreover, the following quantities are invariant for such H^1 solutions:

$$\int_{\mathbb{R}} \psi^2(t) = \int_{\mathbb{R}} \psi^2(0), \tag{3}$$

$$\int_{\mathbb{R}} \left[\frac{1}{2}(\partial_x \psi(t))^2 - \frac{1}{p+1}\psi^{p+1}(t)\right] = \int_{\mathbb{R}} \left[\frac{1}{2}(\partial_x \psi(0))^2 - \frac{1}{p+1}\psi^{p+1}(0)\right]. \tag{4}$$

It follows that if $1 < p < 5$, then H^1 solutions are globally defined and uniformly bounded in H^1.

Second, we review results on the large time dynamics of (1) in a neighborhood of the solitons $R_{c,x_0}(t,x)$. We refer to Miura [19] and Schuur [23] for results on the KdV and modified KdV equations ($p = 2$, 3) which are specific to the integrability theory and the inverse scattering transform. We recall that the stability problem in the energy space H^1 is completely solved. Note that by the scaling and translation invariances of the equation, it is enough to study the stability of the solution $R(t,x) = Q(x-t)$. The following results are well known:

• For $1 < p < 5$, the solitons are stable in H^1, which means that the following result holds.

Stability of solitons (see [26]). *Let $p = 2$, 3, or 4. Let $\psi_0 \in H^1(\mathbb{R})$, and let $\psi(t)$ be the global H^1 solution of (1) satisfying $\psi(0) = \psi_0$. For all $\epsilon > 0$, there exists $\delta > 0$ such that if $\|\psi_0 - Q\|_{H^1} \le \delta$, then for all $t \in \mathbb{R}$, there exists $x(t) \in \mathbb{R}$ such that*

$$\|\psi(t, . + x(t)) - Q\|_{H^1} \le \epsilon.$$

See Benjamin [1], Bona [3], and Weinstein [26].

• For $p \ge 5$, the solitons are not stable in H^1. See Bona, Souganidis, and Strauss [4] for the case $p > 5$. For $p = 5$, the following stronger statement holds true: *there exist initial data ψ_0 arbitrarily close to Q in H^1 such that the corresponding solution of (1) blows up in the H^1 norm in finite time* (see Merle [17], Martel and Merle [14], and the references therein).

The next natural question concerns the asymptotic completeness of the family of solitons. First results related to this question are due to Pego and Weinstein [21], who have proved the asymptotic completeness of the family of solitons in some weighted spaces for (1) with $p = 2$ and 3.

Later, Martel and Merle [13] proved the asymptotic stability of solitons of the gKdV equations in the energy space for $p = 2$, 3, and 4. More precisely, they proved the following.

Asymptotic stability of solitons (see [13]). *Let $p = 2$, 3, or 4. Let $\psi_0 \in H^1(\mathbb{R})$, and let $\psi(t)$ be the global H^1 solution of (1) satisfying $\psi(0) = \psi_0$. There exists $\delta > 0$ such that if*

$$\|\psi_0 - Q\|_{H^1} < \delta, \tag{5}$$

then there exists c^+ close to 1, and for all $t \in \mathbb{R}^+$, there exist $x(t) \in \mathbb{R}$ such that

$$\|\psi(t) - Q_{c^+}(. - x(t))\|_{H^1(x>t/10)} \to 0 \quad \text{as } t \to +\infty, \tag{6}$$

$$\text{and} \quad x'(t) \to c^+ \quad \text{as } t \to +\infty. \tag{7}$$

See also Martel, Merle, and Tsai [16] for the case of multisoliton solutions.

Note that in (6) the convergence as $t \to +\infty$ holds for $x \geq t/10$, which is a large region around the soliton $Q_{c^+}(. - x(t))$ since $x'(t)$ is close to 1. We refer to [15] for the optimality of this result in H^1. Note that a similar result had been previously proved by the same techniques for $p = 5$; see Theorem 2 in [12] (the statement for $p = 5$ is weaker due to the instability of solitons and to a special degeneracy of the critical case).

The proofs of these results in [12] and [13] proceed into three independent steps:

1. First, the problem of asymptotic stability is reduced to a rigidity property of (1) close to the solitons. More precisely, we show that asymptotic stability is true if the following nonlinear Liouville-type theorem holds.

NONLINEAR LIOUVILLE PROPERTY (see [13]). *Let $p = 2$, 3, or 4. Let $\psi_0 \in H^1(\mathbb{R})$, and let $\psi(t)$ be the global H^1 solution of (1) satisfying $\psi(0) = \psi_0$. There exists $\delta > 0$ such that if*

$$\|\psi_0 - Q\|_{H^1} < \delta, \tag{8}$$

and if, for some function $t \in \mathbb{R} \mapsto x(t) \in \mathbb{R}$,

$$\forall \epsilon > 0,\ \exists A > 0,\ \forall t \in \mathbb{R}, \quad \int_{|x|>A} \psi^2(t, x + x(t))dx < \epsilon \tag{9}$$

is satisfied, then there exist c_1 close to 1 and $x_1 \in \mathbb{R}$ such that $\psi(t, x) \equiv R_{c_1,x_1}(t, x)$.

This result is an illustration of the rigidity of the gKdV equations around the solitons. See Theorem 1 in [12] and Theorem 2 in [13].

2. The second step is to prove the nonlinear Liouville property. This is done by reducing the proof to a similar property on a linear problem. Indeed, since we want to prove a result in a neighborhood of the function Q by passing to the limit ($\delta \to 0$) on renormalized problems, we show that it is sufficient to classify a related linearized problem. See a precise statement (linear Liouville property) in Theorem 1 below (see also Corollary 1 in section 3.4).

3. The last step is to prove the linear Liouville property. For $p = 2$, 3, 4, and 5, this property was proved in [12] and [13] by using the positivity of the quadratic form (15), under orthogonality conditions. The study of this quadratic form required at some point numerical calculations.

With respect to the first papers [12] and [13], some of the arguments, especially in the first step, have been simplified. Exponential decay properties of solutions that are uniformly localized in the L^2 norm can be proved more easily using monotonicity properties. See, for example, Laurent and Martel [11] for a simple approach of smoothness and decay for localized solutions.

In a third paper on the subject [15], the proof of asymptotic stability has been further simplified for $p = 2$, 3, and 4. Indeed the property of positivity of the quadratic form (15) could be used to determine a Liapunov functional for the original gKdV problem. This Liapunov functional implies directly the asymptotic stability result and the use of the linear or nonlinear Liouville property is not necessary. However, this approach does not seem to work in the critical case $p = 5$ and still relies on the study of the quadratic form (15) below. Moreover it was unclear whether such a problem could be solved by the same technique for $p > 5$, or even in general for $p \in (1, 5)$ not integer.

In this paper, we present a simpler approach to prove the linear Liouville property (i.e., step 3) which is successful for any real value of $p > 1$. Since there is no restriction

on p, it is now clear that this question is not related to the stability problem. Let

$$\mathcal{L}u = -\partial_x^2 u + u - pQ^{p-1}u = -\partial_x^2 u + u - \frac{p(p+1)}{2\cosh^2\left(\frac{p-1}{2}x\right)}u. \tag{10}$$

Our main result is the following theorem.

THEOREM 1 (linear Liouville property). *Let $p > 1$. Let $u \in C(\mathbb{R}, H^1(\mathbb{R})) \cap L^\infty(\mathbb{R}, H^1(\mathbb{R}))$ be a solution of*

$$\partial_t u = \partial_x (\mathcal{L}u) \quad \text{on } \mathbb{R}\times\mathbb{R}. \tag{11}$$

Assume that for two constants $C > 0$, $\sigma > 0$,

$$\forall t, x \in \mathbb{R}, \quad |u(t,x)| \le Ce^{-\sigma|x|}. \tag{12}$$

Then there exists a constant $b_0 \in \mathbb{R}$ such that for all $t \in \mathbb{R}$,

$$u(t) \equiv b_0 Q'. \tag{13}$$

Since $\mathcal{L}Q' = 0$ (see Lemma 1 below), it is clear that $u(t,x) = b_0 Q'(x)$ is a solution of (11). Theorem 1 claims that there is no other solution that is uniformly localized in the sense of (12).

Remark 1. Theorem 1 is not exactly the result required for the asymptotic stability result. See Corollary 1 in section 3.4 for a direct consequence of Theorem 1 that is the suitable form of the rigidity result to be used for the asymptotic stability.

Remark 2. Assumption (12) can be relaxed to the following condition:

$$\forall \epsilon > 0,\ \exists A > 0,\ \forall t \in \mathbb{R}, \quad \int_{|x|>A} u^2(t,x)dx < \epsilon. \tag{14}$$

Indeed, in this context (14) implies (12) and in fact a much stronger result: H^1 bounded solutions of (11) satisfying (14) are $C^\infty(\mathbb{R}\times\mathbb{R})$ and all their derivatives have exponential decay in x; see Lemma 4 in section 3.2. This is a consequence of smoothing properties of the gKdV equations observed by Kato [9] and of refined techniques introduced in [13] and [11].

Remark 3. Theorem 1 holds for all $p > 1$, and there is no difference in the proof for the various values of p. In particular, sub- or supercriticality of p (i.e., whether $1 < p < 5$ or $p \ge 5$) is not relevant in this problem.

As mentioned above, Theorem 1 (and Corollary 1) was already proved for $p = 5$ (see Theorem 3 in [12]) and for $p = 2$, 3, and 4 (see Proposition 5 in [13]). The proof presented here is simpler and does not require numerical computations. Moreover, it is a unified proof for all real values of $p > 1$.

Remark 4. From Theorem 1, we deduce that for $p \ge 6$ integer, asymptotic completeness of the family of solitons is true in the same sense as Theorem 2 in [12] for $p = 5$. The proof of this result requires some properties of stability of weak H^1 convergence through the flow of the gKdV equation (Lemma 17 in [12]). This property was checked in [12] for $p = 5$ and in [13] for $p = 2$, 3, and 4 using estimates of Kenig, Ponce, and Vega [10] used for proving the well-posedness of the Cauchy problem in H^s spaces. The same can be done for $p \ge 6$ integer.

Since $p > 1$ need not be an integer in Theorem 1, one can expect the asymptotic stability result to be true for all $p > 1$. This would require considering

$$\partial_t \psi + \partial_x(\partial_x^2\psi + |\psi|^{p-1}\psi) = 0$$

instead of (1) and adapting the methods of [10] to this situation.

We now sketch the proof of Theorem 1. Recall that the proof in [13] was based on the quantity $\int xu^2(t,x)dx$. A direct calculation gives for $u(t)$ a solution of (11):

$$-\frac{d}{dt}\int xu^2 = 3\int (\partial_x u)^2 + \int u^2 - p\int Q^{p-1}u^2 + p(p-1)\int xQ'Q^{p-2}u^2 = \overline{H}(u); \tag{15}$$

see Lemma 8 in [13]. When proving the positivity of the quadratic form $\overline{H}(u)$ under orthogonality conditions on $u(t)$, the term $\int Q^{p-1}u^2$ has the negative sign, but it is a classical potential, whose spectrum is explicit. The term $\int xQ'Q^{p-2}u^2$ also has a negative sign since $xQ'(x) \le 0$ for all $x \in \mathbb{R}$, and it is quite difficult to control.

In the present paper, the main idea is to consider the function $v(t,x) = \mathcal{L}u(t,x)$ instead of $u(t,x)$. It satisfies

$$\partial_t v = \mathcal{L}(\partial_x v).$$

Then, we have

$$-\frac{d}{dt}\int xv^2 = 3\int (\partial_x v)^2 + \int v^2 - p\int Q^{p-1}v^2 - p(p-1)\int xQ'Q^{p-2}v^2,$$

and thus the last term now has the good sign, which implies easily the positivity result for some values of p. In fact, we consider in this paper the quantity $\int v^2(t,x)g(x)dx$ for a suitable choice of function $g(x)$ (see Lemma 2) which allows us to prove at once the result for any $p > 1$.

Since we argue on $v(t,x)$ rather than on $u(t,x)$, we need to obtain some regularity on $v(t,x)$. In section 3.2, we prove that under the assumptions of Theorem 1, $u(t,x)$ is completely smooth. The main argument of Theorem 1 is presented in section 3.3.

We now turn to another nonlinear equation, introduced by Peregrine [22] and Benjamin, Bona, and Mahony [2]:

$$(1-\partial_y^2)\partial_s\psi + \partial_y(\psi + \psi^2) = 0, \quad s, y \in \mathbb{R}. \tag{16}$$

Equation (16) is not scaling invariant, but it still has a two-parameter family of traveling wave solutions: for any $c > 1$ and any $y_0 \in \mathbb{R}$,

$$S_{c,y_0}(s,y) = \varphi(y - cs - y_0) \tag{17}$$

is a solution of (16), where φ_c is an H^1 solution of

$$-c\partial_y^2\varphi_c + (c-1)\varphi_c - \varphi_c^2 = 0 \quad \text{on } \mathbb{R}. \tag{18}$$

Note that we have the following expression for φ_c:

$$\varphi_c(y) = (c-1)Q\left(\sqrt{\frac{c-1}{c}}y\right), \tag{19}$$

where Q is defined in (2).

Equation (16) has properties similar to the gKdV equation: it is globally well-posed in $H^1(\mathbb{R})$ (the proof is straightforward, unlike for the gKdV equation), and H^1

solutions have two invariant quantities: for all $s \in \mathbb{R}$,

$$\begin{aligned} &\int (\psi^2 + (\partial_y \psi)^2)(s) = \int (\psi^2 + (\partial_y \psi)^2)(0), \\ &\int \left(\frac{1}{2}\psi^2 + \frac{1}{3}\psi^3\right)(s) = \int \left(\frac{1}{2}\psi^2 + \frac{1}{3}\psi^3\right)(0). \end{aligned} \tag{20}$$

It was proved by Weinstein [27] that all traveling waves S_{c,y_0} are stable in H^1, by the same techniques as for the gKdV equation in [26], using only the invariant quantities (20).

The first result on asymptotic stability of traveling waves for (16) was given by Miller and Weinstein [18] in the spirit of the work of Pego and Weinstein [21] on gKdV. Next, Mizumachi [20] and El Dika [5], [6] gave independently the first result of asymptotic stability of S_{c,y_0} in the energy space. Their proof is in the spirit of [13], with several structural differences. It was first proved in [20] and [6] that asymptotic stability in H^1 is equivalent to a linear Liouville property. Then, their proof differs from the one of the KdV equation since Mizumachi and El Dika both apply a spectral result of Miller and Weinstein [18] to classify the linearized problem around φ_c. In fact, due to the use of Miller and Weinstein's result, the classification was obtained with the restriction $c \in (1, +\infty) \setminus E$, where E is an unknown set with no accumulation point (recall that (16) is not scaling invariant). Results for the case of several traveling waves were obtained by El Dika and Martel [8].

In the present paper, applying the strategy of the proof of Theorem 1, we are able to prove the linear Liouville property related to (16) for any $c > 1$. The linear equation related to the traveling waves S_{c,y_0} is

$$(1 - \partial_y^2)\partial_s \omega = \partial_y(-c\partial_y^2 \omega + (c-1)\omega - 2\varphi_c \omega). \tag{21}$$

We change the variable

$$x = \sqrt{\frac{c-1}{c}}\, y, \quad t = \frac{(c-1)^{3/2}}{c^{1/2}}\, s, \quad \text{and} \quad \omega(s,y) = u(t,x).$$

Then $u(t,x)$ satisfies

$$\left(1 - \frac{c-1}{c}\partial_x^2\right)\partial_t u = \partial_x(-\partial_x^2 u + u - 2Qu).$$

Let $\lambda = \frac{c-1}{c} \in (0,1)$. Note that $\lambda = 0$ corresponds to the limit case of the KdV equation, and $\lambda = 1$ corresponds to $c \to +\infty$. These limit cases, which make no sense for (16), are covered by our result. We claim the following theorem.

THEOREM 2 (linear Liouville property related to (16)). *Let $p = 2$. Let $\lambda \in [0,1]$. Let $u \in C(\mathbb{R}, H^1(\mathbb{R})) \cap L^\infty(\mathbb{R}, H^1(\mathbb{R}))$ be a solution of*

$$(1 - \lambda\partial_x^2)\partial_t u = \partial_x(\mathcal{L}u) \quad \text{on } \mathbb{R} \times \mathbb{R}. \tag{22}$$

Assume that for two constants $C > 0$, $\sigma > 0$,

$$\forall t, x \in \mathbb{R}, \quad \int (u^2 + (\partial_x u)^2)(t,x) e^{\sigma|x|} dx \le C. \tag{23}$$

Then there exists a constant $b_0 \in \mathbb{R}$ such that for all $t \in \mathbb{R}$,

$$u(t) \equiv b_0 Q'. \tag{24}$$

This result implies that the asymptotic stability results stated for (16) in [6] and [8] hold for all $c > 1$ (see [6, Theorem 1] for $p = 2$ and [8, Theorem 2] for $p = 2$ concerning the case of several traveling waves).

Remark 5. A condition of the type (14) on u and $\partial_x u$ is sufficient. See Remark 2 and El Dika [7].

Remark 6. Considering the generalized form of (16),

$$(1 - \partial_y^2)\partial_s \psi + \partial_y(\psi + \psi^p) = 0 \tag{25}$$

for $p > 1$ integer, the problem of asymptotic stability can be reduced to the classification of the following linear equation:

$$(1 - \lambda\partial_x^2)\partial_t u = \partial_x(\mathcal{L}u) \quad \text{on } \mathbb{R} \times \mathbb{R}, \tag{26}$$

where $\mathcal{L}u = -\partial_x^2 u + u - pQ^{p-1}u$. In the present paper, we consider only $p = 2$, and our proof does not seem to adapt immediately to the other cases. Note that $p = 3$ with the restriction $c \notin E$, for some set E without accumulation point, is treated by Miller and Weinstein [18] and thus also in [20] and [6].

The proof of Theorem 2 is given in section 4.

2. Preliminaries. We first gather in Lemma 1 some well-known properties of the operator $\mathcal{L}$. Recall that $\mathcal{L}$ is a classical operator (see Titchmarsh [24]). Then, in Lemma 2, we introduce the function $g(x)$ to be used in the proof of Theorem 1.

LEMMA 1 (properties of $\mathcal{L}$). *Let $p > 1$. The operator $\mathcal{L}$ defined in* (10) *satisfies the following properties:*

(i) *First eigenfunction:* $\mathcal{L}Q^{\frac{p+1}{2}} = -\frac{1}{4}(p-1)(p+3)Q^{\frac{p+1}{2}}$.
(ii) *Second eigenfunction:* $\mathcal{L}Q' = 0$, *and the kernel of* $\mathcal{L}$ *is* $\{\lambda Q', \lambda \in \mathbb{R}\}$.
(iii) *For any* L^2 *function* $h(x)$ *orthogonal to* Q' *for the* L^2 *scalar product, there exists a unique* $f(x) \in H^2(\mathbb{R})$ *orthogonal to* Q' *such that* $\mathcal{L}f = h$.
(iv) $\mathcal{L}(\frac{2}{p-1}Q + xQ') = -2Q$.
(v) *Positivity under orthogonality: for any function* $w \in H^1(\mathbb{R})$,

$$\int w(x)Q^{\frac{p+1}{2}}(x)dx = 0 \Rightarrow \int \left\{(\partial_x w)^2(x) + w^2(x) - pQ^{p-1}(x)w^2(x)\right\} dx \geq 0.$$

Proof. We recall that

$$Q'' = Q - Q^p \quad \text{and} \quad (Q')^2 = Q^2 - \frac{2}{p+1}Q^{p+1} \tag{27}$$

by integration. Thus,

$$\frac{d^2}{dx^2}Q^{\frac{p+1}{2}} = \frac{p+1}{2}\left[\frac{p-1}{2}Q'^2Q^{\frac{p-3}{2}} + Q''Q^{\frac{p-1}{2}}\right] = \left(\frac{p+1}{2}\right)^2 Q^{\frac{p+1}{2}} - pQ^{p-1}Q^{\frac{p+1}{2}},$$

and so $\mathcal{L}Q^{\frac{p+1}{2}} = -\left[\left(\frac{p+1}{2}\right)^2 - 1\right]Q^{\frac{p+1}{2}} = -\frac{1}{4}(p-1)(p+3)Q^{\frac{p+1}{2}}$.

The property $\mathcal{L}Q' = 0$ is easily checked. Moreover, the fact that the spectrum of $\mathcal{L}$ is restricted to $\{\lambda Q', \lambda \in \mathbb{R}\}$ was proved by ODE techniques (see Weinstein [25, Proposition 2.8(b)]). The third property is a direct consequence of the structure of $\mathcal{L}$ as well as the Lax–Milgram theorem. Property (iv) is obtained by direct calculations.

Finally, it follows from the fact that $Q'(x)$ has only one zero on $\mathbb{R}$ that 0 is the second eigenvalue of $\mathcal{L}$. Therefore, if $w \in H^1$ is orthogonal to $Q^{\frac{p+1}{2}}$, then

$$(\mathcal{L}w, w) = \int \left\{(\partial_x w)^2(x) + w^2(x) - pQ^{p-1}(x)w^2(x)\right\} dx \geq 0. \qquad \square$$

We introduce a suitable function $g(x)$ to be used in the proof of Theorem 1, and we list some of its properties.

LEMMA 2 (properties of the function $g(x)$). *Define*

$$g(x) = -\left(\frac{p+1}{p-1}\right)\frac{Q'(x)}{Q(x)}.$$

Then $g(x)$ satisfies the following properties for all $x \in \mathbb{R}$:

$$g'(x) = Q^{p-1}(x) = \frac{p+1}{2\cosh^2\left(\frac{p-1}{2}x\right)}, \quad \left(\frac{g''(x)}{g'(x)}\right)^2 = (p-1)^2\left(1-\frac{2}{p+1}Q^{p-1}(x)\right),$$
$$\frac{g'''(x)}{g'(x)} = (p-1)^2\left(1-\frac{3}{p+1}Q^{p-1}(x)\right), \quad g^2(x) = \left(\frac{p+1}{p-1}\right)^2\left(1-\frac{2}{p+1}Q^{p-1}(x)\right),$$

and

$$(Q^{p-1}g)' = -(p+1)Q^{p-1} + 3Q^{2p-2}, \quad |g(x)| \le \frac{p+1}{p-1}.$$

Proof. By (27) and direct calculations

$$g' = -\left(\frac{p+1}{p-1}\right)\frac{Q''Q-(Q')^2}{Q^2} = Q^{p-1},$$

$$\frac{g''}{g'} = (p-1)\frac{Q'}{Q}, \quad \left(\frac{g''}{g'}\right)^2 = (p-1)^2\left(\frac{Q'}{Q}\right)^2 = (p-1)^2\left(1-\frac{2}{p+1}Q^{p-1}\right).$$

Next,

$$g''' = (p-1)(p-2)(Q')^2Q^{p-3} + (p-1)Q''Q^{p-2} = (p-1)^2Q^{p-1}\left(1-\frac{3}{p+1}Q^{p-1}\right).$$

We also have

$$(Q^{p-1}g)' = -\frac{p+1}{p-1}(Q'Q^{p-2})' = -\frac{p+1}{(p-1)^2}g'''$$

and

$$g^2(x) = \left(\frac{p+1}{p-1}\right)^2\frac{(Q')^2}{Q^2} = \left(\frac{p+1}{p-1}\right)^2\left(1-\frac{2}{p+1}Q^{p-1}\right).$$

The last identity also implies $|g(x)| \le \frac{p+1}{p-1}$. □

3. Linear problem related to the gKdV equations. In this section, we prove Theorem 1. In sections 3.1 and 3.2, we recall some technical arguments for the proof. The main argument of the proof is presented in section 3.3.

3.1. The H^1 Cauchy problem. For the sake of completeness, we claim a well-posedness result for (11) in $H^1(\mathbb{R})$.

LEMMA 3 (H^1 well-posedness of the Cauchy problem for (11)). *Let $u_0 \in H^1(\mathbb{R})$. There exists one and only one function $u \in C(\mathbb{R}, H^1(\mathbb{R}))$ that satisfies (11) in the sense of distributions and such that $u(0) = u_0$. Moreover, if $u_0 \in H^s(\mathbb{R})$, for some*

$s > 1$, then $u \in C(\mathbb{R}, H^s(\mathbb{R}))$ and if (u_0^n) is a sequence of $H^s(\mathbb{R})$, $s \geq 2$, satisfying $u_0^n \to u_0$ in $H^1(\mathbb{R})$, then for all $T > 0$, the corresponding solutions u^n of (11) satisfy $u^n \to u$ in $C([-T, T], H^1(\mathbb{R}))$.

Proof. The existence proof is given in [12, Lemma 9], for $p = 5$. The same argument applies to any $p > 1$.

For uniqueness in the class of functions $C(\mathbb{R}, H^1(\mathbb{R}))$ satisfying (11) in the sense of distribution, since the equation is linear in $u(t)$, we only have to prove that a solution $u(t)$ in this class satisfying $u(0, x) \equiv 0$ is necessarily $u(t, x) \equiv 0$.

To do this, we apply an energy method not directly to $u(t)$ but to $v(t) \in C(\mathbb{R}, H^3(\mathbb{R}))$ defined for any $t \in \mathbb{R}$, by $v(t) = \left(1 - \partial_x^2\right)^{-1} u(t)$. Indeed, $v(t)$ satisfies

$$\partial_t v = -\partial_x^3 v + \partial_x v - p \left(1 - \partial_x^2\right)^{-1} [\partial_x (Q^{p-1} \left(1 - \partial_x^2\right) v)].$$

The second member is in L^2; thus we can take the L^2 scalar product by $v(t)$ and integrate by parts to obtain

$$\frac{d}{dt} \int v^2(t) = p \int Q^{p-1} [\left(1 - \partial_x^2\right) v] \left(1 - \partial_x^2\right)^{-1} \partial_x v.$$

By further integrations by parts, we obtain

$$\begin{aligned} \frac{d}{dt} \int v^2(t) = -\frac{p}{2} \int (Q^{p-1})' v^2 - p \int (Q^{p-1})'' v \left(1 - \partial_x^2\right)^{-1} \partial_x v \\ - 2p \int (Q^{p-1})' v \left(1 - \partial_x^2\right)^{-1} \partial_x^2 v. \end{aligned}$$

Since $\| \left(1 - \partial_x^2\right)^{-1} \partial_x v \|_{L^2} + \| \left(1 - \partial_x^2\right)^{-1} \partial_x^2 v \|_{L^2} \leq C \| v \|_{L^2}$, and by the properties of Q, we obtain by the Cauchy–Schwarz inequality $\left| \frac{d}{dt} \int v^2(t) \right| \leq C \int v^2(t)$. Thus, $v(0) = 0$ implies by the Gronwall inequality that $v(t) = 0$ and thus $u(t) = 0$. This proves the uniqueness statement. The persistence of regularity in H^s and continuous dependence results are easily checked. ∎

3.2. Smoothness and exponential decay. By applying arguments of [11] (see also [13]), we claim that a solution of (11) which is uniformly localized in the L^2 norm is necessarily completely smooth and exponentially decaying.

LEMMA 4 (smoothness and exponential decay). *Let $u(t)$ be an H^1 solution of (11), bounded in H^1 and satisfying*

$$\forall \epsilon > 0, \ \exists A > 0, \ \forall t \in \mathbb{R}, \quad \int_{|x| > A} u^2(t, x) dx < \epsilon. \tag{28}$$

Then $u \in C^\infty(\mathbb{R} \times \mathbb{R})$. Moreover, there exists $\sigma > 0$ and, for all $k \in \mathbb{N}$, there exists $C_k > 0$ such that

$$\forall t, x \in \mathbb{R}, \quad \left| \frac{\partial^k}{\partial x^k} u(t, x) \right| \leq C_k e^{-\sigma |x|}. \tag{29}$$

Remark 7. From the proof of Lemma 4, we can choose $\frac{1}{\sigma} = \max(2, \frac{1}{p-1})$. Note also that for the proof of Theorem 1, we only need to check $u \in H^3(\mathbb{R})$ and (29) for $k = 0$, 1, and 2.

Proof. The proof of Lemma 4 follows from arguments used in the proof of Theorem 1 in [11], but it is in fact simpler since the equation is linear and since the potential

Q^{p-1} has both smoothness and decay properties. Recall that Theorem 1 in [11] concerns similar properties for L^2 localized solutions of the nonlinear equation (1). We present in this proof formal calculations on $u(t)$ that can be justified rigorously by using Lemma 3 and some usual regularizing arguments.

Let $K = \max(2, \frac{1}{p-1})$. For $x \in \mathbb{R}$, let

$$\phi(x) = \frac{2}{\pi} \arctan(\exp(x/K)),$$

so that $\lim_{+\infty} \phi = 1$, $\lim_{-\infty} \phi = 0$, and, for all $x \in \mathbb{R}$, $\phi(-x) = 1 - \phi(x)$. Note also that by direct calculations

$$\phi'(x) = \frac{1}{K\pi \cosh(x/K)}, \quad \phi'''(x) \le \frac{1}{K^2}\phi'(x) \le \frac{1}{4}\phi'(x). \tag{30}$$

Let $x_0 > 0$, $t_0 \in \mathbb{R}$. We define, for all $t < t_0$,

$$I_{x_0,t_0}(t) = \int u^2(t,x)\, \phi\left(x - \frac{1}{2}(t_0 - t) - x_0\right) dx.$$

We first prove the following property of $I_{x_0,t_0}(t)$.

LEMMA 5. *There exists $\theta_0 > 0$ such that for all $x_0 > 0$, $t_0 \in \mathbb{R}$,*

$$\forall t \le t_0, \quad I_{x_0,t_0}(t_0) - I_{x_0,t_0}(t) \le \theta_0 \exp\left(-\frac{x_0}{K}\right).$$

Proof. Let $R_0 > 1$ to be chosen later, and let $x_0 > 2R_0$. Define $\tilde{x} = x - \frac{1}{2}(t_0 - t) - x_0$. Note that $\tilde{x} \le x$. By direct calculations, using the equation of $u(t)$, we have

$$\begin{aligned}
\frac{d}{dt} I_{x_0,t_0}(t) &= -3\int (\partial_x u)^2 \phi'(\tilde{x}) - \frac{1}{2}\int u^2 \phi'(\tilde{x}) + \int u^2 \phi'''(\tilde{x}) \\
&\quad + p\int \left[-(Q^{p-1})'\phi(\tilde{x}) + Q^{p-1}\phi'(\tilde{x})\right] u^2 \\
&\le -3\int (\partial_x u)^2 \phi'(\tilde{x}) - \frac{1}{4}\int u^2 \phi'(\tilde{x}) + p\int \left[-(Q^{p-1})'\phi(\tilde{x}) + Q^{p-1}\phi'(\tilde{x})\right] u^2
\end{aligned}$$

by using (30). By the properties of Q and ϕ, we have

$$\left|-(Q^{p-1})'\phi(\tilde{x}) + Q^{p-1}\phi'(\tilde{x})\right| \le Ce^{-(p-1)|x|}\phi(\tilde{x}).$$

We consider three cases depending on x:

- If $x < R_0$, then $\tilde{x} < -\frac{1}{2}(t_0 - t) - x_0 + R_0 < 0$ and so
$$e^{-(p-1)|x|}\phi(\tilde{x}) \le C\phi(\tilde{x}) \le Ce^{\frac{\tilde{x}}{K}} \le C\exp\left(-\frac{1}{2K}(t_0 - t) - \frac{x_0}{K} + \frac{R_0}{K}\right).$$
- If $R_0 < x < \frac{1}{2}(t_0 - t) + x_0$, then $\tilde{x} < 0$ and so
$$e^{-(p-1)|x|}\phi(\tilde{x}) \le e^{-(p-1)R_0}\phi(\tilde{x}) \le Ce^{-(p-1)R_0}\phi'(\tilde{x}),$$
since $\phi(y) \le C\phi'(y)$ for $y < 0$.
- If $x > \frac{1}{2}(t_0 - t) + x_0 > 0$, then
$$\begin{aligned}
e^{-(p-1)|x|}\phi(\tilde{x}) &\le Ce^{-(p-1)x} \le C\exp\left(-(p-1)\left[\frac{1}{2}(t_0 - t) + x_0\right]\right) \\
&\le C\exp\left(-\left[\frac{1}{2K}(t_0 - t) + \frac{x_0}{K}\right]\right),
\end{aligned}$$
since $K \ge 1/(p-1)$.

Therefore, we have, for all $x \in \mathbb{R}$,

$$\left|-(Q^{p-1})'\phi(\tilde{x}) + Q^{p-1}\phi'(\tilde{x})\right| \leq Ce^{\frac{R_0}{K}} \exp\left(-\frac{1}{2K}(t_0 - t) - \frac{x_0}{K}\right) + C'e^{-(p-1)R_0}\phi'(\tilde{x}).$$

We choose R_0 large enough so that

$$pC'e^{-(p-1)R_0} \int u^2\phi'(\tilde{x}) \leq \frac{1}{8} \int u^2\phi'(\tilde{x}).$$

The value of R_0 being fixed, since $\int u^2(t)$ is bounded, we obtain

$$\frac{d}{dt} I_{x_0,t_0}(t) \leq C \exp\left(-\frac{1}{2K}(t_0 - t) - \frac{x_0}{K}\right).$$

By integrating this estimate between t_0 and some $t < t_0$, we obtain the desired result for $x_0 > 2R_0$. The result for $0 < x_0 < 2R_0$ is clear by possibly taking a larger value of θ_0. □

We continue the proof of Lemma 4. By the proof of this result, proceeding exactly as in [11, pp. 162–164], and using (28), we first obtain

$$I_{x_0,t_0}(t_0) \leq \theta_0 \exp\left(-\frac{x_0}{K}\right).$$

Next, dividing by $\exp(-x_0/K)$, passing to the limit as $x_0 \to +\infty$, and using the properties of ϕ, we deduce that for some constant θ_0',

$$\forall t_0 \in \mathbb{R}, \quad \int u^2(t_0, x) \exp\left(\frac{x}{K}\right) dx + \int_{t_0-1}^{t_0} \int (\partial_x u)^2(s, x) \exp\left(\frac{x}{K}\right) dxds \leq \theta_0'.$$

Proceeding in the solution $u(t)$ of the linear problem (11) exactly as for the solution of (1) in [11, pp. 165–169], using the functionals

$$I^{(k)}_{x_0,t_0}(t) = \int (\partial_x^k u)^2(t, x)\, \phi\left(x - \frac{1}{2}(t_0 - t) - x_0\right) dx$$

for $k \geq 1$ integer, we find by iteration on k that for a constant $\theta_k > 0$, for any $t_0 \in \mathbb{R}$,

$$\int (\partial_x^k u)^2(t_0, x) \exp\left(\frac{x}{K}\right) dx + \int_{t_0-1}^{t_0} \int (\partial_x^{k+1} u)^2(s, x) \exp\left(\frac{x}{K}\right) dxds \leq \theta_k.$$

Since $u(-t, -x)$ is also a solution of (11) satisfying (28), we have as well

$$\int (\partial_x^k u)^2(t_0, x) \exp\left(\frac{|x|}{K}\right) dx + \int_{t_0-1}^{t_0} \int (\partial_x^{k+1} u)^2(s, x) \exp\left(\frac{|x|}{K}\right) dxds \leq \theta_k.$$

It follows that $u(t) \in \cap_{s \leq 0} H^s(\mathbb{R})$, and thus by the equation of $u(t)$ and the smoothness of $Q(x)$, we obtain $u \in C^\infty(\mathbb{R} \times \mathbb{R})$. Finally, by the Gagliardo–Nirenberg inequality

$$\|w\|^2_{L^\infty(x>R_0)} \leq 2\, \|w\|_{L^2(x>R_0)} \|\partial_x w\|_{L^2(x>R_0)},$$

we obtain pointwise estimates (29), as required. We refer to [11] for more details. The only difference in the proof is the method used to estimate the nonlinear terms, which we have detailed in the proof of Lemma 5 for I_{x_0,t_0}. The same estimates are applied to $\frac{d}{dt} I^{(k)}_{x_0,t_0}$. □

3.3. Main argument of the proof of Theorem 1. Let $u(t)$ be an H^1 bounded solution of

$$\partial_t u = \partial_x \left(\mathcal{L} u\right) \quad \text{on } \mathbb{R}\times\mathbb{R},$$

satisfying

$$\forall t, x \in \mathbb{R}, \quad |u(t,x)| \le C e^{-\sigma|x|}.$$

By Lemma 4 we have $u \in C^\infty(\mathbb{R}\times\mathbb{R})$ and

$$\forall k \in \mathbb{N},\ \forall t \in \mathbb{R},\ \forall x \in \mathbb{R}, \quad \left|\frac{\partial^k u}{\partial x^k}(t,x)\right| \le C_k e^{-\sigma|x|}. \tag{31}$$

Let $\tilde v(t,x) = \mathcal{L}u(t,x)$. The equation of $\tilde v(t)$ is

$$\partial_t \tilde v = \mathcal{L}(\partial_t u) = \mathcal{L}\left(\partial_x \tilde v\right). \tag{32}$$

Note also that $\int \tilde v Q' = \int u \mathcal{L} Q' = 0$. Now, we modify $\tilde v(t)$ to obtain a second orthogonality condition by setting

$$v(t,x) = \tilde v(t,x) - \left(\frac{\int \tilde v(t,x) Q^p(x) dx}{\int Q^{p+1}(x) dx}\right) Q(x),$$

so that

$$\int v(t,x) Q^p(x) dx = \int v(t,x) Q'(x) dx = 0.$$

Moreover, since $\mathcal{L}Q' = 0$, $v(t,x)$ satisfies the equation

$$\partial_t v = \mathcal{L}\left(\partial_x v\right) + \delta(t) Q, \tag{33}$$

where

$$\delta(t) = -\frac{1}{\int Q^{p+1}} \frac{d}{dt} \int \tilde v(t,x) Q^p(x) dx.$$

Note that $v \in C^\infty(\mathbb{R}\times\mathbb{R})$ and by (31) and the properties of $Q(x)$ there exist constants σ, $C > 0$ such that

$$\forall t \in \mathbb{R},\ \forall x \in \mathbb{R}, \quad |v(t,x)| \le C e^{-\sigma|x|}. \tag{34}$$

We multiply the equation of $v(t)$ by $2\, v(t,x) g(x)$, where $g(x)$ is defined in Lemma 2, and we integrate over $\mathbb{R}$; we obtain

$$\frac{d}{dt}\int v^2(t,x) g(x) dx = 2\int v \partial_t v g = 2\int \mathcal{L}\left(\partial_x v\right) v g + 2\,\delta(t) \int Q v g = 2\int \mathcal{L}\left(\partial_x v\right) v g.$$

The term in $\delta(t)$ disappears since $\int Q v(t) g = -\left(\frac{p+1}{p-1}\right)\int v(t) Q' = 0$, by the definition of the function $g(x)$ and orthogonality of $v(t)$. By integration by parts, we have

$$\begin{aligned} 2\int \mathcal{L}\left(\partial_x v\right) v g &= -2\int (\partial_x^3 v) v g + 2\int (\partial_x v) v g - 2p \int Q^{p-1} (\partial_x v) v g \\ &= 2\int (\partial_x^2 v)(\partial_x v) g + 2\int (\partial_x^2 v) v g' - \int v^2 g' + p\int (Q^{p-1} g)' v^2 \\ &= -3\int (\partial_x v)^2 g' - \int v^2 \left[-g''' + g' - p(Q^{p-1} g)'\right]. \end{aligned} \tag{35}$$

We now set $w(t,x) = v(t,x)\sqrt{g'(x)} = v(t,x)Q^{\frac{p-1}{2}}(x)$. We have

$$\partial_x w = \partial_x v\sqrt{g'} + \frac{1}{2}\frac{g''}{\sqrt{g'}}v,$$

and so by integration by parts,

$$\int (\partial_x w)^2 = \int (\partial_x v)^2 g' + \frac{1}{4}\int v^2\frac{(g'')^2}{g'} - \frac{1}{2}\int v^2 g''',$$

which also reads

$$\int (\partial_x v)^2 g' = \int (\partial_x w)^2 - \frac{1}{4}\int w^2\left(\frac{g''}{g'}\right)^2 + \frac{1}{2}\int w^2\left(\frac{g'''}{g'}\right).$$

Thus,

$$\frac{d}{dt}\int v^2(t,x)g(x)dx = -3\int (\partial_x w(t,x))^2 dx - \int w^2(t,x)A(x)dx,$$

where

$$A(x) = 1 + \frac{1}{2}\frac{g'''(x)}{g'(x)} - \frac{3}{4}\left(\frac{g''(x)}{g'(x)}\right)^2 - p\frac{(Q^{p-1}g)'(x)}{g'(x)}.$$

Using Lemma 2, we obtain the following expression for $A(x)$:

$$\begin{aligned} A(x) &= 1 + \frac{1}{2}(p-1)^2 - \frac{3(p-1)^2}{2(p+1)}Q^{p-1} - \frac{3}{4}(p-1)^2 + \frac{3(p-1)^2}{2(p+1)}Q^{p-1} \\ &\quad + p(p+1) - 3pQ^{p-1} \\ &= \frac{3}{4}(p+1)^2 - 3pQ^{p-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} -\frac{1}{3}\frac{d}{dt}\int v^2(t,x)g(x)dx &= \int (\partial_x w)^2 + \left(\frac{p+1}{2}\right)^2\int w^2 - p\int Q^{p-1}w^2 \\ &= (\mathcal{L}w, w) + \frac{(p-1)(p+3)}{4}\int w^2 \end{aligned} \tag{36}$$

(where $(.,.)$ denotes the L^2 scalar product). In the right-hand member, we recognize a nonnegative quadratic form related to the operator $\mathcal{L}$. Note that

$$\int v(t,x)Q^p(x)dx = 0 \quad \text{is equivalent to} \quad \int w(t,x)Q^{\frac{p+1}{2}}(x)dx = 0.$$

By Lemma 1, we thus have $(\mathcal{L}w, w) \geq 0$,

$$-\frac{1}{3}\frac{d}{dt}\int v^2(t,x)g(x)dx \geq \frac{1}{4}(p-1)(p+3)\int w^2(t,x)dx. \tag{37}$$

Since $|g(x)| \leq \frac{p+1}{p-1}$, the function $\int v^2(t,x)g(x)dx$ is uniformly bounded in time, and so

$$\int_{-\infty}^{+\infty}\int w^2(t,x)dxdt < +\infty.$$

It follows that for some sequence $t_n \to +\infty$, we have

$$\int w^2(t_n, x)dx \to 0 \quad \text{as } n \to +\infty.$$

Thus, from (34), we obtain

$$\int v^2(t_n, x)dx \to 0 \quad \text{as } n \to +\infty.$$

Similarly, we have for a sequence $s_n \to +\infty$

$$\int v^2(-s_n, x)dx \to 0 \quad \text{as } n \to +\infty.$$

Using (37) integrated over $(-s_n, t_n)$ and the bound on $g(x)$, we obtain, for all $n \in \mathbb{N}$,

$$\int_{-s_n}^{t_n} \int w^2(t, x)dxdt \leq C \left(\int \left[v^2(t_n, x) + v^2(-s_n, x) \right] dx \right).$$

By passing to the limit as $n \to +\infty$, we obtain

$$\int_{-\infty}^{+\infty} \int w^2(t, x)dxdt = 0,$$

giving us $w(t, x) \equiv 0$, and thus $v(t, x) \equiv 0$. Therefore, $\mathcal{L}u(t, x) = \tilde{v}(t, x) = \gamma(t)Q(x)$, and so by Lemma 1,

$$u(t, x) = -\frac{1}{2}\gamma(t)\left(\frac{2}{p-1}Q + xQ' \right)(x) + b(t)Q'(x) = a(t)\left(\frac{2}{p-1}Q + xQ' \right)(x) + b(t)Q'(x).$$

But from the equation of $u(t, x)$, the functions $a(t)$, $b(t)$ satisfy the following relations:

$$a'(t) = 0, \quad b'(t) = -2a(t).$$

Since $u(t)$ is uniformly bounded, the only possible solutions $u(t)$ are for $b(t) \equiv b_0$ constant, and $a(t) \equiv 0$. Thus, $u(t) \equiv b_0 Q'$.

3.4. Classification for a more general problem. As a direct corollary of Theorem 1, we obtain the following classification result.

COROLLARY 1. *Let $p > 1$. Let $u \in C(\mathbb{R}, H^1(\mathbb{R})) \cap L^\infty(\mathbb{R}, H^1(\mathbb{R}))$ be a solution of*

$$\partial_t u = \partial_x (\mathcal{L}u) + \alpha(t) \left(\frac{2}{p-1}Q + xQ' \right) + \beta(t)Q' \quad \text{on } \mathbb{R} \times \mathbb{R}, \tag{38}$$

where $\alpha(t)$ and $\beta(t)$ are two continuous and bounded functions. Assume that for two constants $C > 0$, $\sigma > 0$,

$$\forall t, x \in \mathbb{R}, \quad |u(t, x)| \leq Ce^{-\sigma|x|}. \tag{39}$$

Then for all $t \in \mathbb{R}$,

$$u(t) \equiv a(t) \left(\frac{2}{p-1}Q + xQ' \right) + b(t)Q' \tag{40}$$

for some C^1 bounded functions $a(t)$ and $b(t)$ satisfying

$$a'(t) = \alpha(t), \quad b'(t) = -2\,a(t) + \beta(t). \tag{41}$$

Since by direct calculations

$$\mathcal{L}\left(\frac{2}{p-1}Q + xQ'\right) = -2Q \quad \text{and} \quad \mathcal{L}Q' = 0$$

(see Lemma 1), it is easily checked that (40)–(41) indeed define a solution of (38). Corollary 1 means that there is no other H^1 bounded solution of (38) satisfying (39).

Problem (38) is the one that really appears in the proof of asymptotic stability in [12], [13].

Proof. Let $u(t)$ be a solution of (38) as in the statement of Corollary 1, where $\alpha(t)$, $\beta(t)$ are bounded. We have to consider separately the cases $p \neq 5$ and $p = 5$, because of some structural differences.

Case $p \neq 5$. Since $p \neq 5$, we have, by integrating by parts,

$$\int \left(\frac{2}{p-1}Q + xQ'\right) Q = \frac{5-p}{2(p-1)} \int Q^2 \neq 0. \tag{42}$$

Multiplying the equation of $u(t)$ by Q and integrating on $\mathbb{R}$, using $\mathcal{L}Q' = 0$, we obtain by (42)

$$\frac{5-p}{2(p-1)} \left(\int Q^2\right) \alpha(t) = \frac{d}{dt} \int u(t)Q.$$

Let

$$\tilde{u}(t) = u(t) - \frac{2(p-1)}{5-p} \frac{1}{\int Q^2} \left(\int u(t)Q\right) \left(\frac{2}{p-1}Q + xQ'\right).$$

Then $\tilde{u}(t)$ satisfies, by $\mathcal{L}(\frac{2}{p-1}Q + xQ') = -2Q$,

$$\begin{aligned} \partial_t \tilde{u} - \partial_x(\mathcal{L}\tilde{u}) &= \partial_t u - \alpha(t) \left(\frac{2}{p-1}Q + xQ'\right) - \partial_x(\mathcal{L}u) - \frac{4(p-1)}{5-p} \frac{Q'}{\int Q^2} \left(\int u(t)Q\right) \\ &= \tilde{\beta}(t)Q', \end{aligned} \tag{43}$$

where $\tilde{\beta}(t) = \beta(t) - \frac{p-1}{5-p} \frac{4}{\int Q^2} \int u(t)Q$ is bounded. Moreover, for all $t \in \mathbb{R}$, $\int \tilde{u}(t)Q = 0$.

Now let

$$J(t) = \int_{\mathbb{R}} \tilde{u}(t,y) \left(\int_0^y \left(\frac{2}{p-1}Q + xQ'\right) dx\right) dy.$$

Then, since $\mathcal{L}(\frac{2}{p-1}Q + xQ') = -2Q$, we have

$$J'(t) = 2 \int \tilde{u}(t)Q - \tilde{\beta}(t) \int Q \left(\frac{2}{p-1}Q + xQ'\right), \text{ and so } \frac{5-p}{2(p-1)} \left(\int Q^2\right) \tilde{\beta}(t) = -J'(t).$$

Since $\mathcal{L}Q' = 0$, it follows that $w(t,x) = \tilde{u}(t) - \frac{2(p-1)}{5-p} \frac{1}{\int Q^2} J(t)Q'$ satisfies

$$\partial_t w - \partial_x(\mathcal{L}w) = 0,$$

and thus $w(t)$ satisfies the assumptions of Theorem 1. It follows that $w(t,x) = b_0 Q'(x)$, for some constant b_0, and

$$u(t,x) = \tilde{u}(t,x) + a(t)\left(\frac{2}{p-1}Q + xQ'\right) = a(t)\left(\frac{2}{p-1}Q + xQ'\right) + b(t)Q',$$

where $a(t)$ and $b(t)$ are bounded and satisfy

$$a'(t) = \alpha(t), \quad b'(t) = -2a(t) + \beta(t).$$

Case $p = 5$. Since $p = 5$, we have $\int (\frac{2}{p-1}Q + xQ')Q = 0$, and we need to modify the previous argument. In fact, the argument that we present now for $p = 5$ works also for any $p \neq 3$. Indeed, since $p \neq 3$, we have

$$\int \left(\frac{2}{p-1}Q + xQ'\right) = \frac{3-p}{p-1}\int Q \neq 0.$$

By integrating the equation of $u(t)$, we deduce

$$\frac{d}{dt}\int u(t) = \alpha(t)\frac{3-p}{p+1}\int Q.$$

Thus, we set

$$\tilde{u}(t) = u(t) - \frac{p+1}{3-p}\frac{1}{\int Q}\left(\int u(t)\right)\left(\frac{2}{p-1}Q + xQ'\right),$$

so that $\tilde{u}(t)$ satisfies (43) for some $\tilde{\beta}(t)$ and $\int \tilde{u}(t) = 0$.

Observe now that $\mathcal{L}(1) = 1 - pQ^{p-1}$. Let $j \in H^1(\mathbb{R})$ be such that $\int j(x)Q'(x)dx = 0$ and $\mathcal{L}j = pQ^{p-1}$ (the existence and uniqueness of $j(x)$ is a consequence of Lemma 1). Then $\mathcal{L}(1+j) = 1$. Multiplying the equation of $\tilde{u}(t)$ by $\int_0^x (1+j)(y)dy$ and integrating, we obtain the following:

$$\text{let} \quad J(t) = \int_{\mathbb{R}} \tilde{u}(t,x)\left(\int_0^x (1+j)(y)dy\right)dx; \quad \text{then} \quad J'(t) = -\tilde{\beta}(t)\int Q(1+j)(x).$$

But $\int Q(1+j) = -\frac{1}{2}\int \mathcal{L}(\frac{2}{p-1}Q + xQ')(1+j) = -\frac{1}{2}\int (\frac{2}{p-1}Q + xQ') = -\frac{1}{2}\frac{3-p}{p-1}\int Q$, so that by setting

$$w(t,x) = \tilde{u}(t,x) - \frac{2(p-1)}{3-p}\frac{Q'}{\int Q}J(t),$$

we obtain $\partial_t w - \partial_x(\mathcal{L}w) = 0$, and w satisfies the assumptions of Theorem 1. We finish as for the case $p \neq 5$. □

4. Linear problem related to (16). In this section, we prove Theorem 2. The proof follows closely that for the KdV case (i.e., Theorem 1 with $p = 2$). First, we observe that by the techniques developed by El Dika [7], if $u(t)$ is an H^1 solution of

$$(1 - \lambda\partial_x^2)\partial_t u = \partial_x(\mathcal{L}u) \quad \text{on } \mathbb{R}\times\mathbb{R},$$

satisfying

$$\forall t, x \in \mathbb{R}, \quad \int (u^2 + (\partial_x u)^2)(t,x)e^{\sigma_0|x|}dx \leq C,$$

then $u \in C^\infty(\mathbb{R}\times\mathbb{R})$ and satisfies

$$\forall t,x\in\mathbb{R},\quad \left|\frac{\partial^k}{\partial x^k}u(t,x)\right| \le C_k e^{-\sigma|x|}. \tag{44}$$

Using [7], the proof is similar to that of Lemma 4 and so we omit it.

We restrict ourselves to the case $p=2$ and consider any $\lambda\in[0,1]$ (which means that for the original nonlinear problem, we consider any speed $c>1$).

Let

$$v(t,x)=\mathcal{L}u(t,x)\quad\text{and}\quad \tilde{h}(t,x)=\left(1-\lambda\partial_x^2\right)^{-1}v(t,x), \tag{45}$$

i.e., for any $t\in\mathbb{R}$, $\tilde{h}(t)$ is solution of $\left(1-\lambda\partial_x^2\right)\tilde{h}(t)=v(t)$. Then $v(t,x)$ and $\tilde{h}(t,x)$ satisfy

$$\partial_t v=\mathcal{L}(\partial_x\tilde{h}),\quad \partial_t\tilde{h}=\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x\tilde{h}).$$

We set $R(x)=\left(1-\lambda\partial_x^2\right)Q(x)=(1-\lambda)Q(x)+\lambda Q^2(x)$, by the equation of $Q(x)$. We have

$$\int vQ'=\int u\mathcal{L}Q'=0\quad\text{and so}\quad \int\tilde{h}R'=0.$$

We modify $\tilde{h}$ in order to get another orthogonality condition. Indeed, setting

$$h(t,x)=\tilde{h}(t,x)-\frac{\int\tilde{h}(t,x)R(x)dx}{\int R(x)Q(x)dx}Q(x)$$

(note that $\int R(x)Q(x)dx=(1-\lambda)\int Q^2+\lambda\int Q^3\neq 0$), we have, for all $t\in\mathbb{R}$,

$$\int h(t,x)R(x)dx=\int h(t,x)R'(x)dx=0.$$

Moreover, since $\mathcal{L}Q'=0$, $h(t,x)$ satisfies the equation

$$\partial_t h=\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)+\delta(t)Q,$$

where

$$\delta(t)=-\frac{1}{\int RQ}\frac{d}{dt}\int\tilde{h}(t,x)R(x)dx.$$

In contrast to the KdV case, we will need an expression of $\delta(t)$ in terms of h. For this, we take the L^2 scalar product of the equation of h by $R(x)$. We obtain

$$0=\frac{d}{dt}\int h(t,x)R(x)dx=\int\left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]R+\delta(t)\int QR.$$

Since $\left(1-\lambda\partial_x^2\right)^{-1}R=Q$ and $\mathcal{L}Q=-Q^2$, we obtain

$$\delta(t)=-\frac{1}{(1-\lambda)\int Q^2+\lambda\int Q^3}\int h(Q^2)'. \tag{46}$$

From (44) it follows that

$$\forall t \in \mathbb{R}, \quad \int v^2(t,x)\cosh(\sigma x)dx \le C, \tag{47}$$

where we choose $0 < \sigma < \frac{1}{2}$. Using $v(t) = \left(1 - \lambda\partial_x^2\right)\tilde{h}(t)$ and integrating by parts, we obtain

$$\forall t \in \mathbb{R}, \quad \int \left(\lambda^2(\partial_x^2\tilde{h}(t))^2 + 2\lambda(\partial_x\tilde{h}(t))^2 + (1-\lambda\sigma^2)\tilde{h}^2(t)\right)\cosh(\sigma x)dx \le C, \tag{48}$$

and so by the properties of Q,

$$\forall t \in \mathbb{R}, \quad \int \left[\lambda(\partial_x h(t))^2 + h^2(t)\right] e^{\sigma|x|}dx \le C. \tag{49}$$

We now choose the functional of $h(t,x)$ that will allow us to obtain $h = 0$. Set

$$K(t) = \lambda\int (\partial_x h)^2(t)g + \int h^2(t)g - \lambda\int h^2(t)g'',$$

where the function $g(x)$ is defined in Lemma 2. We claim the following lemma.

LEMMA 6. *For all $t \in \mathbb{R}$,*

$$K'(t) \le -H(t), \quad \textit{where } H(t) = -2\int \mathcal{L}(\partial_x h)hg - 2\int(\partial_x h)^2 g'. \tag{50}$$

Proof. We have by direct calculations using the equation of $h(t,x)$ and integration by parts,

$$\begin{aligned}\frac{d}{dt}\int(\partial_x h)^2 g &= 2\int \partial_x\left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right](\partial_x hg) + 2\delta(t)\int Q'(\partial_x hg)\\ &= -2\int\left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right](\partial_x^2 hg + \partial_x hg') - 2\delta(t)\int(Q'g)'h,\end{aligned}$$

$$\frac{d}{dt}\int h^2 g = 2\int\left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]hg + 2\delta(t)\int Qhg,$$

and

$$\frac{d}{dt}\int h^2 g'' = 2\int\left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]hg'' + 2\delta(t)\int Qhg''.$$

Thus,

$$\begin{aligned}K'(t) = 2\int&\left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right](-\lambda\partial_x^2 hg - \lambda\partial_x hg' - \lambda hg'' + hg)\\ &+ 2\delta(t)\int h\left(-\lambda(Q'g)' + Qg - \lambda Qg''\right).\end{aligned}$$

We set $V = -\lambda(Q'g)' + Qg - \lambda Qg''$. Moreover, we note that

$$-\lambda\partial_x^2 hg - 2\lambda\partial_x hg' - \lambda hg'' + hg = \left(1-\lambda\partial_x^2\right)(hg).$$

Thus, we obtain

$$
\begin{aligned}
K'(t) &= 2\int \left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]\left(-\lambda\partial_x^2 hg - 2\lambda\partial_x hg' - \lambda hg'' + hg\right)\\
&\quad + 2\lambda\int \left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]\partial_x hg' + 2\delta(t)\int hV \qquad (51)\\
&= 2\int \mathcal{L}(\partial_x h)hg + 2\lambda\int \left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]\partial_x hg' + 2\delta(t)\int hV.
\end{aligned}
$$

The first term in (51), i.e., $2\int \mathcal{L}(\partial_x h)hg$, has already been developed for the KdV equation (see (35)). We consider the second term in (51). Since $\mathcal{L}h = -\partial_x^2 h + h - 2Qh = -\partial_x^2 h + h - 2g'h$, we have

$$
\begin{aligned}
&2\lambda\int \left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]\partial_x hg'\\
&= 2\int (\partial_x h)^2 g' - 2(1-\lambda)\int \left[\left(1-\lambda\partial_x^2\right)^{-1}(\partial_x h)\right]\partial_x hg'\\
&\quad - 4\lambda\int \left[\left(1-\lambda\partial_x^2\right)^{-1}(\partial_x hg')\right]\partial_x hg'.
\end{aligned}
$$

For $t\in\mathbb{R}$, we denote by $a(t,x)$ and $b(t,x)$ the functions satisfying, respectively,

$$
\left(1-\lambda\partial_x^2\right)a(t,x) = h(t,x) \quad \text{and} \quad \left(1-\lambda\partial_x^2\right)b(t,x) = \partial_x h(t,x)g'(x).
$$

In other words, $a(t) = \left(1-\lambda\partial_x^2\right)^{-1}h(t)$ and $b(t) = \left(1-\lambda\partial_x^2\right)^{-1}(\partial_x h(t)g')$. Using these functions, we obtain

$$
\begin{aligned}
\int \left[\left(1-\lambda\partial_x^2\right)^{-1}(\partial_x h)\right]\partial_x hg' &= \int \partial_x a\left[\left(1-\lambda\partial_x^2\right)(\partial_x a)\right]g'\\
&= \int \left[(\partial_x a)^2 + \lambda(\partial_x^2 a)^2\right]g' - \frac{\lambda}{2}\int (\partial_x a)^2 g'''
\end{aligned}
$$

and

$$
\int \left(1-\lambda\partial_x^2\right)^{-1}(\partial_x hg')\partial_x hg' = \int b\left(1-\lambda\partial_x^2\right)b = \int \left[b^2 + \lambda(\partial_x b)^2\right].
$$

Thus, we have

$$
\begin{aligned}
&2\lambda\int \left[\left(1-\lambda\partial_x^2\right)^{-1}\mathcal{L}(\partial_x h)\right]\partial_x hg'\\
&= 2\int (\partial_x h)^2 g' - 2(1-\lambda)\int \left(\left[(\partial_x a)^2 + \lambda(\partial_x^2 a)^2\right]g' - \frac{\lambda}{2}(\partial_x a)^2 g'''\right)\\
&\quad - 4\lambda\int \left[b^2 + \lambda(\partial_x b)^2\right].
\end{aligned}
$$

By Lemma 2 for $p=2$, we have $g''' = g' - (g')^2 \le g'$, and thus by $\lambda \le 1$, we obtain

$$
\int \left(\left[(\partial_x a)^2 + \lambda(\partial_x^2 a)^2\right]g' - \frac{\lambda}{2}(\partial_x a)^2 g'''\right) \ge 0.
$$

It follows that

$$
K'(t) \le 2\int \mathcal{L}(\partial_x h)hg + 2\int (\partial_x h)^2 g' + 2\delta(t)\int hV.
$$

We now consider the term $2\delta(t)\int hV$, where $V = -\lambda(Q'g)' + Qg - \lambda Qg''$. We use Lemma 2 with $p = 2$ to compute V:

$$V = 3\lambda\left(\frac{(Q')^2}{Q}\right)' - 3Q' - \lambda QQ' = 3\lambda\left(Q - \frac{2}{3}Q^2\right)' - 3Q' - \frac{\lambda}{2}(Q^2)' = -3R' + \frac{\lambda}{2}(Q^2)'.$$

Thus, by (46) and $\int hR' = 0$, we obtain

$$2\delta(t)\int hV = -\frac{\lambda}{(1-\lambda)\int Q^2 + \lambda\int Q^3}\left(\int h(Q^2)'\right)^2 \le 0.$$

Therefore, we finally obtain

$$K'(t) \le 2\int \mathcal{L}(\partial_x h)hg + 2\int (\partial_x h)^2 g',$$

which finishes the proof of Lemma 6. □

We continue the proof of Theorem 2. By (35), we have

$$H(t) = \int (\partial_x h)^2 g' + \int h^2\left[-g''' + g' - 2(Qg)'\right].$$

We set $w(t,x) = h(t,x)\sqrt{g'(x)} = h(t,x)\sqrt{Q(x)}$. Note first that

$$\int hR = 0 \quad \Leftrightarrow \quad \int w\left[(1-\lambda)Q^{\frac{1}{2}} + \lambda Q^{\frac{3}{2}}\right] = 0.$$

Since

$$\int (\partial_x h)^2 g' = \int (\partial_x w)^2 - \frac{1}{4}\int w^2\left(\frac{g''}{g'}\right)^2 + \frac{1}{2}\int w^2\left(\frac{g'''}{g'}\right),$$

we obtain

$$H(t) = \int (\partial_x w)^2 + \int B(x)w^2,$$

where, by Lemma 2 applied to $p = 2$,

$$\begin{aligned} B(x) &= -\frac{1}{4}\left(\frac{g''}{g'}\right)^2 - \frac{1}{2}\left(\frac{g'''}{g'}\right) + 1 - 2\frac{(Qg)'}{g'} \\ &= -\frac{1}{4}\left(1 - \frac{2}{3}Q\right) - \frac{1}{2}(1-Q) + 1 + 2(3-3Q) = \frac{25}{4} - \frac{16}{3}Q. \end{aligned} \tag{52}$$

Now, we have to check the positivity of

$$H(t) = \int (\partial_x w)^2 + \frac{25}{4}\int w^2 - \frac{16}{3}\int Qw^2,$$

under the following orthogonality condition on w:

$$\int w\left[(1-\lambda)Q^{\frac{1}{2}} + \lambda Q^{\frac{3}{2}}\right] = 0.$$

The expression of H in w is related to the operator $-\partial_x^2+\frac{25}{4}-\frac{16}{3}Q$, which is a classical operator (see Titchmarsh [24]). Let us recall some computations related to Q^α for $\alpha>0$:

$$-(Q^\alpha)''=-\alpha^2Q^\alpha+\alpha\left(\frac{2\alpha+1}{3}\right)Q^{\alpha+1}. \tag{53}$$

Indeed, we have

$$\begin{aligned}-(Q^\alpha)''&=-\alpha\left(Q''Q^{\alpha-1}+(\alpha-1)(Q')^2Q^{\alpha-2}\right)\\&=-\alpha\left(\alpha Q^\alpha-\left(1+\frac{2(\alpha-1)}{3}\right)Q^{\alpha+1}\right)=-\alpha^2Q^\alpha+\alpha\left(\frac{2\alpha+1}{3}\right)Q^{\alpha+1}.\end{aligned}$$

This leads us to compare $H(t)$ to a quadratic form for which the first eigenfunction will be explicit and simple. Indeed, we decompose $H(t)$ as

$$H(t)=\int(\partial_xw)^2+\frac{25}{4}\int w^2-5\int Qw^2-\frac{1}{3}\int Qw^2,$$

and we observe that since $|Q(x)|\le\frac{3}{2}$, we have

$$H(t)\ge\int(\partial_xw)^2+\frac{23}{4}\int w^2-5\int Qw^2\equiv\tilde H(t).$$

The quadratic form $\tilde H(t)$ is related to the operator $\tilde{\mathcal{L}}=-\partial_x^2+\frac{23}{4}-5Q$, i.e., $\tilde H(t)=(\tilde{\mathcal{L}}w,w)$. By (53), we have $\tilde{\mathcal{L}}Q^{\frac{5}{2}}=-\frac{1}{2}Q^{\frac{5}{2}}$, which means that the first eigenvalue of $\tilde{\mathcal{L}}$ is $-\frac{1}{2}$. On the other hand, we check by elementary computation that $\tilde{\mathcal{L}}(Q^2)'=\frac{7}{4}(Q^2)'$, which means that the second eigenvalue of $\tilde{\mathcal{L}}$ is $\frac{7}{4}$. Moreover, $\tilde{\mathcal{L}}$ has a continuous spectrum, which is $[\frac{23}{4},+\infty)$. Thus, for odd functions, the operator is positive, and there is only one negative direction for even functions. We now want to check that there exists $\kappa>0$ such that, for any H^1 function w,

$$\int w\left[(1-\lambda)Q^{\frac{1}{2}}+\lambda Q^{\frac{3}{2}}\right]=0\quad\Rightarrow\quad(\tilde{\mathcal{L}}w,w)\ge\kappa\int w^2 \tag{54}$$

(where $(.,.)$ is the L^2 scalar product). Let $\tilde R=(1-\lambda)Q^{\frac{1}{2}}+\lambda Q^{\frac{3}{2}}$. Following Lemma E.1 of Weinstein [26], it suffices to check that $(\tilde R,Q^{\frac{5}{2}})>0$, which is clearly true, and $(\tilde{\mathcal{L}}^{-1}\tilde R,\tilde R)<0$. Using (53) it turns out that $\tilde{\mathcal{L}}^{-1}\tilde R$ can be computed explicitly.

Indeed, by (53) for $\alpha=\frac{3}{2}$, we have

$$\tilde{\mathcal{L}}Q^{\frac{3}{2}}=\frac{7}{2}Q^{\frac{3}{2}}-3Q^{\frac{5}{2}},$$

and since $\tilde{\mathcal{L}}Q^{\frac{5}{2}}=-\frac{1}{2}Q^{\frac{5}{2}}$, we deduce

$$Q^{\frac{3}{2}}=\tilde{\mathcal{L}}\left(\frac{2}{7}Q^{\frac{3}{2}}-\frac{12}{7}Q^{\frac{5}{2}}\right).$$

In a similar way, we have

$$\tilde{\mathcal{L}}Q^{\frac{1}{2}}=\frac{11}{2}Q^{\frac{1}{2}}-\frac{14}{3}Q^{\frac{3}{2}},$$

and so using $\tilde{\mathcal{L}}Q^{\frac{3}{2}}$ and $\tilde{\mathcal{L}}Q^{\frac{5}{2}}$, we obtain

$$Q^{\frac{1}{2}} = \tilde{\mathcal{L}}\left(\frac{2}{11}Q^{\frac{1}{2}} + \frac{8}{33}Q^{\frac{3}{2}} - \frac{16}{11}Q^{\frac{5}{2}}\right).$$

Thus, using

$$\int Q = \int Q^2, \quad \int Q^3 = \frac{6}{5}\int Q^2, \quad \int Q^4 = \frac{54}{35}\int Q^2 \tag{55}$$

(deduced from $Q'' = Q - Q^2$, $(Q')^2 = Q^2 - \frac{2}{3}Q^3$), we obtain

$$\begin{aligned}
(\tilde{\mathcal{L}}^{-1}Q^{\frac{3}{2}}, Q^{\frac{3}{2}}) &= \frac{2}{7}\int Q^3 - \frac{12}{7}\int Q^4 = -\frac{562}{245}\int Q^2 < 0,\\
(\tilde{\mathcal{L}}^{-1}Q^{\frac{1}{2}}, Q^{\frac{1}{2}}) &= \frac{2}{11}\int Q + \frac{8}{33}\int Q^2 - \frac{16}{11}\int Q^3 = -\frac{218}{165}\int Q^2 < 0,\\
(\tilde{\mathcal{L}}^{-1}Q^{\frac{3}{2}}, Q^{\frac{1}{2}}) &= (\tilde{\mathcal{L}}^{-1}Q^{\frac{1}{2}}, Q^{\frac{3}{2}}) = \frac{2}{7}\int Q^2 - \frac{12}{7}\int Q^3 = -\frac{62}{35}\int Q^2 < 0.
\end{aligned} \tag{56}$$

Since all the terms in (56) are negative, and $\tilde{R} = (1-\lambda)Q^{\frac{1}{2}} + \lambda Q^{\frac{3}{2}}$ with $\lambda \in [0,1]$, it is clear that $(\tilde{\mathcal{L}}^{-1}\tilde{R}, \tilde{R}) < 0$. Therefore, (54) is verified.

Using (54) and the previous calculations on $K'(t)$ (Lemma 6), we deduce

$$K'(t) \le -H(t) \le -\kappa \int w^2(t).$$

Integrating between $-\infty$ and $+\infty$, we deduce that $\int_{-\infty}^{+\infty}\int w^2(t,x)dxdt < +\infty$. Using this in the expression of $H(t)$, we also have $\int_{-\infty}^{+\infty}\int (\partial_x w)^2(t,x)dxdt < +\infty$. At this point, using (49), we obtain $w(t) \equiv 0$ on $\mathbb{R}\times\mathbb{R}$ exactly as in the proof of Theorem 1 in section 3.

By $w \equiv 0$ we obtain $h \equiv 0$, and thus $\tilde{h}(t,x) \equiv a(t)Q(x)$. Then $\partial_t \tilde{h} = 0$ implies $a'(t) = 0$, and so $\tilde{h}(t,x) = a_0 Q(x)$ for a constant $a_0 \in \mathbb{R}$. Finally, we obtain $\mathcal{L}u = a_0\left(1 - \lambda\partial_x^2\right)Q$, and so $u(t) = a_0\mathcal{L}^{-1}\left(\left(1-\lambda\partial_x^2\right)Q\right) + b(t)Q'$, and by the equation of $u(t)$, this implies $b'(t) = a_0$. But since $b(t)$ has to be bounded, this implies $a_0 = 0$ and $b(t) = b_0$ for some constant $b_0 \in \mathbb{R}$. Thus $u(t) \equiv b_0 Q'$, which is solution for any $b_0 \in \mathbb{R}$, and Theorem 2 is proved.

REFERENCES

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

[2] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans. Roy. Soc. London Ser. A, 272 (1972), pp. 47–78.

[3] J. L. Bona, *On the stability theory of solitary waves*, Proc. Roy. Soc. London Ser. A, 349 (1975), pp. 363–374.

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

[5] K. El Dika, *Stabilité asymptotique des ondes solitaires de l'équation de Benjamin–Bona–Mahony*, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 649–652.

[6] K. El Dika, *Asymptotic stability of solitary waves for the Benjamin–Bona–Mahony equation*, Discrete Contin. Dyn. Syst., 13 (2005), pp. 583–622.

[7] K. El Dika, *Smoothing effect of the generalized BBM equation for localized solutions moving to the right*, Discrete Contin. Dyn. Syst., 12 (2005), pp. 973–982.

[8] K. El Dika and Y. Martel, *Stability of N solitary waves for the generalized BBM equations*, Dyn. Partial Differ. Equ., 1 (2004), pp. 401–437.
[9] T. Kato, *On the Cauchy problem for the (generalized) Korteweg–de Vries equation*, in Studies in Applied Mathematics, Adv. Math. Suppl. Stud. 8, Academic Press, New York, 1983, pp. 93–128.
[10] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg–de Vries equation via the contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.
[11] C. Laurent and Y.Martel, *Smoothness and exponential decay of L^2-compact solutions of the generalized KdV equations*, Comm. Partial Differential Equations, 28 (2003), pp. 2093–2107.
[12] Y. Martel and F. Merle, *A Liouville theorem for the critical generalized Korteweg–de Vries equation*, J. Math. Pures Appl., 79 (2000), pp. 339–425.
[13] Y. Martel and F. Merle, *Asymptotic stability of solitons for subcritical generalized KdV equations*, Arch. Ration. Mech. Anal., 157 (2001), pp. 219–254.
[14] Y. Martel and F. Merle, *Blow up in finite time and dynamics of blow up solutions for the L^2-critical generalized KdV equations*, J. Amer. Math. Soc., 15 (2002), pp. 617–664.
[15] Y. Martel and F. Merle, *Asymptotic stability of solitons of the subcritical gKdV equations revisited*, Nonlinearity, 18 (2005), pp. 55—80.
[16] Y. Martel, F. Merle, and T.-P. Tsai, *Stability and asymptotic stability in the energy space of the sum of N solitons for subcritical gKdV equations*, Comm. Math. Phys., 231 (2002), pp. 347–373.
[17] F. Merle, *Existence of blow-up solutions in the energy space for the critical generalized Korteweg–de Vries equation*, J. Amer. Math. Soc., 14 (2001), pp. 555–578.
[18] J. R. Miller and M. I. Weinstein, *Asymptotic stability of solitary waves for the regularized long-wave equation,* Comm. Pure Appl. Math., 49 (1996), pp. 399–441.
[19] R. M. Miura, *The Korteweg–de Vries equation: A survey of results*, SIAM Rev., 18 (1976), pp. 412–459.
[20] T. Mizumachi, *Asymptotic stability of solitary wave solutions to the regularized long-wave equation*, J. Differential Equations, 200 (2004), pp. 312–341.
[21] R. L. Pego and M. I. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys., 164 (1994), pp. 305–349.
[22] D. H. Peregrine, *Long waves on a beach*, J. Fluid Mech., 27 (1967), pp. 815–827.
[23] P. C. Schuur, *Asymptotic Analysis of Solitons Problems*, Lecture Notes in Math. 1232, Springer-Verlag, Berlin, 1986.
[24] E. C. Titchmarsh, *Eigenfunction Expansions Associated with Second-Order Differential Equations*, Clarendon Press, Oxford, 1946.
[25] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.
[26] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.
[27] M. I. Weinstein, *Existence and dynamic stability of solitary wave solutions of equations arising in long wave propagation*, Comm. Partial Differential Equations, 12 (1987), pp. 1133–1173.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 782–794

ON GLOBAL WELL-POSEDNESS OF THE LAGRANGIAN AVERAGED EULER EQUATIONS*

THOMAS Y. HOU† AND CONGMING LI‡

Abstract. We study the global well-posedness of the Lagrangian averaged Euler equations in three dimensions. We show that a necessary and sufficient condition for the global existence is that the bounded mean oscillation of the stream function is integrable in time. We also derive a sufficient condition in terms of the total variation of certain level set functions, which guarantees the global existence. Furthermore, we obtain the global existence of the averaged two-dimensional (2D) Boussinesq equations and the Lagrangian averaged 2D quasi-geostrophic equations in finite Sobolev space in the absence of viscosity or dissipation.

Key words. well-posedness, nonlinearity partial differential equations, Euler equations

AMS subject classifications. 76B03, 35L60, 35M10

DOI. 10.1137/050625783

1. Introduction.

The question of global existence for the three-dimensional (3D) incompressible Euler equations is a very challenging open question. The main difficulty is to understand the effect of vortex stretching, which is absent in the two-dimensional (2D) Euler equations. As part of the effort to understand the vortex stretching effect for 3D flows, various simplified model equations have been proposed in the literature. Amongst these models, the 2D Boussinesq system and the quasi-geostrophic equations are two of the most commonly used because they share a similar vortex stretching effect as that in the 3D incompressible flow. An interesting recent development is the Lagrangian averaged Euler equations [15, 16]. This work was originally motivated by the development of a one-dimensional (1D) shallow water theory [3]. The averaged Euler models have been used to study the average behavior of the 3D Euler and Navier–Stokes equations and used as a turbulent closure model (see, e.g., [5]). The theoretical and computational aspects of the Lagrangian averaged Euler and Navier–Stokes equations have been studied by several authors [4, 5, 16, 22, 21, 13]. However, the global existence of the 3D Lagrangian averaged Euler equations is still open, although the Lagrangian averaged Navier–Stokes equations have been shown to have global existence [21, 13].

In this paper, we consider the global existence of the 3D Lagrangian averaged Euler equations and the corresponding 2D Lagrangian averaged Boussinesq equations and the averaged 2D quasi-geostrophic equations in the absence of viscosity or dissipation. The 3D Lagrangian averaged Euler equations have been derived by Holm, Marsden, and Ratiu [15, 16] (see page 1458 of [21]) in the following form:

$$\partial_t u + (u_\alpha \cdot \nabla)u + (\nabla u_\alpha)^T \cdot u = -\nabla p; \tag{1}$$

*Received by the editors March 2, 2005; accepted for publication (in revised form) February 7, 2006; published electronically August 7, 2006.

http://www.siam.org/journals/sima/38-3/62578.html

†Applied and Computational Mathematics, 217-50, Caltech, Pasadena, CA 91125 (hou@acm.caltech.edu). This author's research was supported in part by the National Science Foundation under FRG grant DMS-0353838 and ITR grant ACI-0204932.

‡Department of Applied Mathematics, University of Colorado, Boulder, CO 80309 (cli@colorado.edu). This author's research was supported in part by the National Science Foundation under grant DMS-0401174.

here the notation is different from that in [21]. Our u_α corresponds to the original u and our u corresponds to $(1-\alpha^2\triangle)u$ in [21].

We will adopt the vorticity formulation [21]:

$$\partial_t\omega + (u_\alpha\cdot\nabla)\omega = \nabla u_\alpha\cdot\omega, \tag{2}$$

where u, ω, the α-averaged velocity u_α, and the divergence-free vector stream function ψ are related by

$$-\triangle\psi = \omega, \quad u = \nabla\times\psi, \tag{3}$$

$$u_\alpha = (1-\alpha^2\triangle)^{-1}u. \tag{4}$$

One of the important properties of the averaged Euler equations is the following identity (see (3.3) on page 1457 of [21]; recall that u_α is called u in [21]):

$$\frac{1}{2}\frac{d}{dt}\int_{R^3}\left(|u_\alpha|^2+\alpha^2|\nabla u_\alpha|^2\right)dx = 0.$$

This conservation property gives a priori bound on the H^1 norm of u_α:

$$\|u_\alpha\|_{H^1}\le C_\alpha. \tag{5}$$

The above reformulation gives a clear physical interpretation of the Lagrangian averaged Euler equations. The vorticity is convected by the α-averaged velocity field. If one discretizes the averaged Lagrangian Euler equations by the point vortex method, i.e., to approximate the initial vorticity by a collection of point vortices (Dirac delta functions), then the resulting numerical approximation is a vortex blob method with α being the vortex blob size [23, 14].

With the above interpretation of the averaged Lagrangian Euler equations, we can clearly apply the same averaging principle to other fluid dynamics equations. For example, if we apply the same Lagrangian averaging principle to the density equation, we would obtain the following Lagrangian averaged 2D Boussinesq equations:

$$\omega_t + u\cdot\nabla\omega = \rho_{x_1}, \tag{6}$$

$$\rho_t + u_\alpha\cdot\nabla\rho = 0\,, \tag{7}$$

where $u_\alpha = (1-\alpha^2\triangle)^{-1}u$ and u is related to the vorticity ω through the usual vorticity stream function formulation; see (3). We refer the reader to [25] for the derivation and discussions of the physical applications for the Boussinesq equations. Note that we replace the velocity by the averaged velocity only in the density equation but not in the vorticity equation. We remark that the global well-posedness of the viscous 2D Boussinesq equations has been obtained in [17].

Similarly, we can derive the Lagrangian averaged 2D quasi-geostrophic equations as follows:

$$\theta_t + u_\alpha\cdot\nabla\theta = 0, \tag{8}$$

$$u = \nabla^\perp\psi, \quad (-\triangle)^{1/2}\psi = \theta, \tag{9}$$

$$u_\alpha = (1-\alpha^2\triangle)^{-1/2}u\,, \tag{10}$$

where $\nabla^\perp \psi = (\partial_{x_2}\psi, -\partial_{x_1}\psi)$ and $(-\triangle)^{1/2}$ is defined as

$$(-\triangle)^{1/2}\psi \equiv \int e^{2\pi i x\cdot\xi}(2\pi|\xi|)\hat{\psi}(\xi)\, d\xi,$$

with $\hat{\psi}(\xi)$ being the Fourier transform of ψ. We refer the reader to [9] for derivation and discussions of the quasi-geostrophic equation. Note that we use a weaker averaged velocity field for the 2D quasi-geostrophic equation. The exponent 1/2 in the averaging operator corresponds to the critical case in the corresponding dissipative quasi-geostrophic equations [8].

In this paper, we prove that a necessary and sufficient condition for the global existence is that the bounded mean oscillation (BMO) norm (see [18]) of the stream function is integrable in time. This is an analogue of the well-known Beale–Kato–Majda condition [1] for the 3D Euler equations. For some recent results on the 3D Euler equations that explore the geometric properties of the Euler flow, we refer the reader to [10, 11]. Moreover, using a level formulation, we derive a sufficient condition for the global existence. The nonblow-up condition we obtain is expressed in terms of the total variation of a level set function; see (53) in section 3 for the precise definition. Assume that the initial vorticity can be expressed in the form $\omega(0,x) = \omega_0(\phi_0,\psi_0)\nabla\phi_0 \times \nabla\psi_0$ for some smooth and bounded level set functions ϕ_0 and ψ_0. Let ϕ and ψ be the level set functions satisfying

$$\begin{aligned} &\phi_t + (u_\alpha \cdot \nabla)\phi = 0, \quad \phi(0,x) = \phi_0(x),\\ &\psi_t + (u_\alpha \cdot \nabla)\psi = 0, \quad \psi(0,x) = \psi_0(x). \end{aligned}$$

Then vorticity can be expressed in terms of these two level set functions:

$$\omega = \omega_0(\phi,\psi)\nabla\phi \times \nabla\psi.$$

Moreover, if the total variation of either ϕ or ψ is integrable in time, then there is no finite time blow-up of the 3D averaged Euler equations. This result has a geometric interpretation. In particular, it excludes the possibility of a finite number of isolated singularities when vorticity is considered as a 1D function by fixing the other two variables. If there is a finite time singularity, the 1D restriction of vorticity must be highly oscillatory at the singularity time, and the singularities are dense in the singular region. We remark that global well-posedness of a family of Navier–Stokes alpha-like models has been recently studied by Olson and Titi [24].

Application of the same argument to the corresponding 2D models gives much sharper existence results. In particular, we prove the global existence of the Lagrangian averaged 2D Boussinesq equations and the averaged 2D quasi-geostrophic equations in finite Sobolev spaces without any assumption on the solution itself.

The rest of the paper is organized as follows. In section 2, we prove the necessary and sufficient condition for the 3D Lagrangian averaged Euler equations and prove the global existence for the averaged 2D Boussinesq equations and the averaged 2D quasi-geostrophic equations. In section 3 we present some results for the global existence of the 3D Lagrangian averaged Euler equations using a novel level set formulation.

2. Main results and proofs. In this section, we present three results. The first result is a necessary and sufficient condition for the global existence of the averaged Euler equations. The second result is the global existence of the averaged 2D Boussinesq equations. The third result is the global existence of the averaged 2D

quasi-geostrophic equations. We begin by stating our first result for the 3D averaged Euler equations. Our result uses the BMO norm. Before we state our existence result, we remind the reader of the definition of the BMO norm:

$$\|f\|_{BMO} = \sup_{x\in R^3}\sup_{r>0}\frac{1}{|B_r|}\int |f-\bar{f}|dx,$$

where $\bar{f} = \frac{1}{|B_r|}\int_{B_r} f(y)dy$, $B_r = \{y \in R^3, \quad |y-x| \le r\}$, and $|B_r|$ is the volume of B_r.

THEOREM 1. *Assume that $\omega_0 \in H^m(R^3)$, $m \ge 0$. Then for any $\alpha > 0$, the solution of the Lagrangian averaged 3D Euler equations (2)–(4) has a unique global solution in $H^m(R^3)$ satisfying*

$$\|\omega(t)\|_{H^m} \le C(T)\|\omega_0\|_{H^m} \quad \textit{for} \quad 0 \le t \le T$$

if we have

$$\int_0^T \|\psi\|_{BMO}dt < \infty \tag{11}$$

for any $T > 0$. Conversely, if the maximal time T of the existence of classical solutions is finite, then we necessarily have

$$\int_0^T \|\psi\|_{BMO}dt = \infty. \tag{12}$$

Proof. The proof relies on the following estimate obtained by Kozono and Taniuchi [19]:

$$\|f\|_\infty \le C\left(1 + \|f\|_{BMO}(1 + \log(\|f\|_{W^{s,p}} + e)\right) \tag{13}$$

for all $f \in W^{s,p}$ with $1 < p < \infty$ and $s > n/p$, where n is the space dimension. We remark that inequalities of similar type for various functional spaces have been obtained earlier; see, e.g., [28].

Another useful result is the following embedding estimate in the BMO norm:

$$\|Rf\|_{BMO} \le C\|f\|_{BMO} \tag{14}$$

for any Riesz-type operator R (see [26, 27]).

It follows from (2)–(4) that

$$u_\alpha = (1-\alpha^2\triangle)^{-1}\nabla \times \psi. \tag{15}$$

This implies that

$$\nabla u_\alpha = \tilde{R}\psi, \tag{16}$$

where $\tilde{R} = \nabla(1-\alpha^2\triangle)^{-1}\nabla\times$ is a Riesz-type operator.

Now applying the embedding estimate (14) to (16), we obtain

$$\|\nabla u_\alpha\|_{BMO} \le C\|\psi\|_{BMO}. \tag{17}$$

Using estimates (13) and (17), we get

$$\|\nabla u_\alpha\|_\infty \le C\,(1+\|\nabla u_\alpha\|_{BMO}\log(\|\nabla u_\alpha\|_{W^{1,4}}+e)) \tag{18}$$

$$\le C\,(1+\|\psi\|_{BMO}\log(\|\omega\|_{L^2}+e))\,,$$

where we have used $u=\nabla\times(-\triangle)^{-1}\omega$ and the Sobolev embedding estimate

$$\|\nabla u_\alpha\|_{W^{1,p}} \le C(\|u_\alpha\|_{H^3}) \le C(\|u_\alpha\|_{H^1}+\|\omega\|_{L^2})$$

for $p\in[2,6]$ and the fact that $\|u_\alpha\|_{H^1}$ is bounded from (5).

Next, we perform an energy estimate for the vorticity equation. Multiplying both sides of the vorticity equation (2) by ω and integrating over R^3, we get

$$\frac{1}{2}\frac{d}{dt}\int_{R^3}|\omega|^2dx+\int_{R^3}(u_\alpha\cdot\nabla\omega)\cdot\omega dx=\int_{R^3}(\nabla u_\alpha\omega)\cdot\omega dx. \tag{19}$$

Note that using integration by parts, we have

$$\int_{R^3}(u_\alpha\cdot\nabla\omega)\cdot\omega dx=\frac{1}{2}\int_{R^3}(u_\alpha\cdot\nabla)|\omega|^2dx=-\frac{1}{2}\int_{R^3}(\nabla\cdot u_\alpha)|\omega|^2dx=0, \tag{20}$$

since $\nabla\cdot u_\alpha=0$.

On the other hand, we obtain by using estimate (18)

$$\begin{aligned}\left|\int_{R^3}(\nabla u_\alpha\omega)\cdot\omega dx\right| &\le \|\nabla u_\alpha\|_\infty\int_{R^3}|\omega|^2dx\\ &\le C\,(1+\|\psi\|_{BMO}\log(\|\omega\|_{L^2}+e))\,\|\omega\|_{L^2}^2.\end{aligned} \tag{21}$$

Putting together estimates (19)–(21), we get

$$\frac{1}{2}\frac{d}{dt}\|\omega\|_{L^2}^2 \le C\,(1+\|\psi\|_{BMO}\log(\|\omega\|_{L^2}+e))\,\|\omega\|_{L^2}^2. \tag{22}$$

The Gronwall inequality then implies that

$$\|\omega(t)\|_{L^2}\le C(T)\quad\text{for}\quad 0\le t\le T, \tag{23}$$

since $\int_0^T\|\psi\|_{BMO}dt<\infty$ by our assumption (11). Moreover,

$$\|\nabla u_\alpha(t)\|_\infty \le C||u_\alpha||_{H^3}\le C(||u_\alpha||_{H^1}+||\omega||_{L^2})\le C(T). \tag{24}$$

Using (2) and (24), we can easily show that

$$\|\omega(t)\|_\infty\le\|\omega_0\|_\infty\exp\left(\int_0^T\|\nabla u_\alpha\|_\infty dt\right)\le C(T)\quad\text{for}\quad 0\le t\le T. \tag{25}$$

Now it is a standard exercise to obtain energy estimates in high-order Sobolev norms [20]

$$\frac{d}{dt}\|\omega\|_{H^m}\le C_m(\|\nabla u_\alpha\|_\infty+\|\omega\|_\infty)\|\omega\|_{H^m}. \tag{26}$$

Since $\|\nabla u_\alpha(t)\|_\infty$ and $\|\omega(t)\|_\infty$ are bounded for $0 \le t \le T$, we obtain the desired estimate for $\|\omega\|_{H^m}$ up to time T.

Now if the maximal time T of the existence of classical solutions is finite, then we must have

$$\int_0^T \|\psi\|_{BMO} dt = \infty,$$

since if $\int_0^T \|\psi\|_{BMO} dt < \infty$, the above argument would imply that $\|\omega(t)\|_{H^m} \le C(T)\|\omega_0\|_{H^m}$ for $t \le T$, which is a contradiction. This completes the proof.

Next, we prove the global existence of the averaged Boussinesq equations (6)–(7).

THEOREM 2. *Assume that $\omega_0 \in H^m(R^2)$ and $\rho_0 \in H^{m+1}(R^2)$ for $m \ge 0$. Then for any $\alpha > 0$, the Lagrangian averaged 2D Boussinesq equations* (6)–(7) *have a unique global solution in $H^m(R^2)$ satisfying*

$$\|\omega(t)\|_{H^m} + \|\rho(t)\|_{H^{m+1}} \le C(T)(\|\omega_0\|_{H^m} + \|\rho_0\|_{H^{m+1}}), \quad 0 \le t \le T,$$

for any $T > 0$.

Proof. First, a standard energy estimate shows that $\|u\|_{L^2}$ is bounded since $\|\rho\|_{L^2}$ is conserved in time and bounded.

Let $W = \nabla^\perp \rho$. Then W satisfies the following evolution equation:

$$W_t + (u_\alpha \cdot \nabla) W = \nabla u_\alpha \cdot W. \tag{27}$$

For any odd integer $p > 2$, we multiply (6) by ω^{p-1} and (27) by $|W|^{p-2}W$, respectively, and integrate over R^2. Upon using integration by parts for the convection terms and exploring the incompressibility of the velocity fields, u and u_α, we obtain

$$\begin{aligned} \frac{1}{p}\frac{d}{dt}\int_{R^2} (|\omega|^p + |W|^p)\, dx &\le (1 + \|\nabla u_\alpha\|_\infty) \int_{R^2} |W|^p dx + \int_{R^2} |\omega|^p dx \\ &\le (1 + \|\nabla u_\alpha\|_\infty) \left(\int_{R^2} |\omega|^p dx + \int_{R^2} |W|^p dx \right), \end{aligned} \tag{28}$$

where we have used Yang's inequality to obtain

$$\begin{aligned} \int_{R^2} \omega^{p-1} |\rho_{x_1}| dx &\le \frac{p-1}{p} \int_{R^2} |\omega|^p dx + \frac{1}{p} \int_{R^2} |\rho_{x_1}|^p dx \\ &\le \int_{R^2} |\omega|^p dx + \int_{R^2} |W|^p dx. \end{aligned}$$

Using estimates (13), we get

$$\begin{aligned} \|\nabla u_\alpha\|_\infty &\le C\,(1 + \|\nabla u_\alpha\|_{BMO} \log(\|\nabla u_\alpha\|_{W^{1,p}} + e)) \\ &\le C\,(1 + \|\nabla u_\alpha\|_{BMO} \log(\|\omega\|_{L^p} + e))\,, \end{aligned} \tag{29}$$

where we have used $u = \nabla^\perp(-\triangle)^{-1}\omega$ and the Sobolev embedding estimate

$$\|\nabla u_\alpha\|_{W^{1,p}} \le C(\|u_\alpha\|_{H^1} + \|\omega\|_{L^p})$$

for $p \ge 2$ and the fact that $\|u_\alpha\|_{H^1} \le C\|u\|_{L^2}$ is bounded.

On the other hand, we obtain by using the John–Nirenberg-type estimate (definition of BMO) in two dimensions

$$\|\nabla u_\alpha\|_{BMO} \le C\|\nabla u_\alpha\|_{H^1} \le C\|u\|_{L^2} \le C, \tag{30}$$

where we have used the fact that $\|u\|_{L^2}$ is bounded. Therefore, we obtain by combining (28), (29), and (30) that

$$\frac{d}{dt}\left(\|\omega\|_{L^p}^p + \|W\|_{L^p}^p\right) \le \left(\|\omega\|_{L^p}^p + \|W\|_{L^p}^p\right)\left(1 + \log(\|\omega\|_{L^p}^p + \|W\|_{L^p}^p + e)\right). \tag{31}$$

The Gronwall inequality then implies that

$$\|\omega(t)\|_{L^p} + \|W(t)\|_{L^p} \le C(T) \quad \text{for} \quad 0 \le t \le T. \tag{32}$$

Using (29), (30), and (32), we get

$$\|\nabla u_\alpha(t)\|_\infty \le C(T) \quad \text{for} \quad 0 \le t \le T. \tag{33}$$

It follows from (27) and (33) that

$$\|W(t)\|_\infty \le C(T) \quad \text{for} \quad 0 \le t \le T, \tag{34}$$

which in turns implies that

$$\|\omega(t)\|_\infty \le C(T) \quad \text{for} \quad 0 \le t \le T. \tag{35}$$

Now it is a standard exercise to show [20] that

$$\frac{d}{dt}\left(\|\omega\|_{H^m} + \|W\|_{H^m}\right) \le C(T)(\|\nabla u_\alpha\|_\infty + \|W\|_\infty)(\|\omega\|_{H^m} + \|W\|_{H^m}).$$

The theorem now follows from (34)–(35) and the Gronwall inequality. This completes the proof of the theorem.

Next, we prove the global existence of the averaged 2D quasi-geostrophic equations (8)–(10).

THEOREM 3. *Assume that $\theta_0 \in H^{m+1}(R^2)$ for $m \ge 0$. Then for any $\alpha > 0$, the solution of the Lagrangian averaged 2D quasi-geostrophic equations* (8)–(10) *has a unique global solution in $H^{m+1}(R^2)$ satisfying*

$$\|\theta(t)\|_{H^{m+1}} \le C(T)\|\theta_0\|_{H^{m+1}}), \quad 0 \le t \le T,$$

for any $T > 0$.

Proof. Again, we can perform a standard energy estimate to show that $\|\theta\|_{L^p}$ is bounded by $\|\theta_0\|_{L^p}$ (including $p = \infty$, which can be obtained via the so-called maximum principle).

Let $\omega = \nabla^\perp \theta$. Then ω satisfies the following evolution equation:

$$\omega_t + (u_\alpha \cdot \nabla)\omega = \nabla u_\alpha \cdot \omega. \tag{36}$$

Thus, ω shares the similar vortex stretching term as the 3D Euler equation. Now using an argument similar to our energy estimate for (27), we can obtain

$$\frac{1}{p}\frac{d}{dt}\int_{R^2} |\omega|^p dx \le \|\nabla u_\alpha\|_\infty \int_{R^2} |\omega|^p dx. \tag{37}$$

Note that

$$\nabla u_\alpha = \nabla(1-\alpha^2\triangle)^{-1/2}(-\triangle)^{-1/2}\nabla^\perp\theta \equiv R\theta$$

for some Riesz-type operator R. Using the embedding estimates

$$\|\nabla u_\alpha\|_{BMO} \le C\|\theta\|_{BMO}$$

and

$$\|\nabla u_\alpha\|_{W^{1,p}} \le C\|\theta\|_{W^{1,p}} \le C(\|\theta\|_{L^p} + \|\nabla\theta\|_{L^p}) \quad \text{for } 1 < p < \infty,$$

we obtain using (13) that

$$\begin{aligned}\|\nabla u_\alpha\|_\infty &\le C\left(1 + \|\nabla u_\alpha\|_{BMO}\log(\|\nabla u_\alpha\|_{W^{1,p}} + e)\right)\\ &\le C\left(1 + \|\theta\|_{BMO}\log(\|\theta\|_{L^p} + \|\omega\|_{L^p} + e)\right)\\ &\le C\left(1 + \|\theta\|_\infty \log(\|\omega\|_{L^p} + e)\right)\\ &\le C\left(1 + \log(\|\omega\|_{L^p} + e)\right). \qquad (38)\end{aligned}$$

Substituting (38) into (37) gives

$$\frac{d}{dt}\|\omega\|_{L^p}^p \le C\left(1 + \log(\|\omega\|_{L^p} + e)\right)\|\omega\|_{L^p}^p. \qquad (39)$$

The Gronwall inequality then implies that

$$\|\omega(t)\|_{L^p} \le C(T)\|\omega_0\|_{L^p}, \quad 0 \le t \le T, \qquad (40)$$

which, together with (38), gives

$$\|\nabla u_\alpha\|_\infty \le C(T), \quad 0 \le t \le T. \qquad (41)$$

Now it follows from (41) and (36) that

$$\|\omega\|_\infty \le C(T), \quad 0 \le t \le T. \qquad (42)$$

Now it is a standard exercise to show [20] that

$$\frac{d}{dt}\|\omega\|_{H^m} \le C(T)(\|\omega\|_\infty + \|\nabla u_\alpha\|_\infty)\|\omega\|_{H^m}.$$

The theorem now follows from (41)–(42) and the Gronwall inequality. This completes the proof of the theorem.

3. The level set formulation for the 3D Euler equations. In this section, we will present a level set formulation for the 3D Euler equations and show how they can be used to obtain a sufficient condition to guarantee the global existence of the averaged Euler equations.

We consider the 3D Euler equations in the vorticity form:

$$\partial_t\omega + (u\cdot\nabla)\omega = \nabla u\cdot\omega, \quad \omega(0,x) = \omega_0(x), \qquad (43)$$

where $\omega = \nabla\times u$ and u is divergence-free.

Let $X(t,\alpha)$ be the Lagrangian flow map satisfying

$$\frac{d}{dt}X(t,\alpha) = u(t,X(t,\alpha)), \quad X(0,\alpha) = \alpha. \tag{44}$$

Since u is divergence-free, we know that the determinant of the Jacobian matrix $\frac{\partial X}{\partial \alpha}$ is identically equal to one. It is well known that vorticity along the Lagrangian trajectory has the following analytical expression [6]:

$$\omega(t,X(t,\alpha)) = \frac{\partial X}{\partial \alpha}\omega_0(\alpha). \tag{45}$$

Let $\theta(t,x)$ be the inverse map of $X(t,\alpha)$, i.e., $X(t,\theta(t,x)) \equiv x$. Then it is easy to show that θ satisfies the following evolution equation:

$$\theta_t + (u\cdot\nabla)\theta = 0, \quad \theta(0,x) = x. \tag{46}$$

Let $\theta = (\theta_1,\theta_2,\theta_3)$ and $\omega_0 = (\omega_0^{(1)},\omega_0^{(2)},\omega_0^{(3)})$. Using (45) and the fact that $X_\alpha\theta_x = I$ and $|\theta_x| = 1$, we can show that

$$\omega(t,x) = \omega_0^{(1)}(\theta)\nabla\theta_2\times\nabla\theta_3 + \omega_0^{(2)}(\theta)\nabla\theta_3\times\nabla\theta_1 + \omega_0^{(3)}(\theta)\nabla\theta_1\times\nabla\theta_2. \tag{47}$$

Note that $\theta_j (j = 1,2,3)$ are level set functions convected by the flow velocity u. In general, one can show that if the initial vorticity $\omega(0,x) = \omega_0(\phi_0,\psi_0)\nabla\phi_0\times\nabla\psi_0$ and the level set functions ϕ and ψ satisfy

$$\phi_t + (u\cdot\nabla)\phi = 0, \quad \phi(0,x) = \phi_0(x), \tag{48}$$

$$\psi_t + (u\cdot\nabla)\psi = 0, \quad \psi(0,x) = \psi_0(x), \tag{49}$$

then the vorticity at a later time can be expressed in terms of these two level set functions and their gradients:

$$\omega(t,x) = \omega_0(\phi,\psi)\nabla\phi\times\nabla\psi. \tag{50}$$

This level set formulation has been considered by Deng, Hou, and Yu in their study of the 3D Euler equations [12]. The special case when $\omega_0 = 1$ is also known as the Clebsch representation [7]. In this case, the velocity field has the form

$$u = \nabla p + \phi\nabla\psi$$

for some potential function p.

It is easy to see that the above level set formulation of vorticity for the 3D Euler equations also applies to the 3D Lagrangian averaged Euler equations. The only change is that the level set functions now satisfy

$$\phi_t + (u_\alpha\cdot\nabla)\phi = 0, \quad \phi(0,x) = \phi_0(x), \tag{51}$$

$$\psi_t + (u_\alpha\cdot\nabla)\psi = 0, \quad \psi(0,x) = \psi_0(x). \tag{52}$$

Now we state a sufficient condition for the global existence of the Lagrangian averaged Euler equations in terms of the property of the level set functions defined by (51)–(52). Before we state our result, we first introduce a definition of the total variation of a level set function, ϕ, as follows:

$$\|\phi\|_{TVx_1} = \sup_{x_2,x_3}\int_{-\infty}^{\infty}\left|\frac{\partial}{\partial x_1}\phi(x_1,x_2,x_3)\right|dx_1. \tag{53}$$

We can define $\|\phi\|_{TVx_2}$ and $\|\phi\|_{TVx_3}$ similarly and let $\|\phi\|_{TV} = \sum_{i=1}^{3} \|\phi\|_{TVx_i}$.

THEOREM 4. *Assume that $\omega(0,x) = \omega_0(\phi_0,\psi_0)\nabla\phi_0 \times \nabla\psi_0$ is the form of the initial vorticity, with ω_0, ϕ_0, and ψ_0 being smooth and bounded. Moreover, we assume that ϕ and ψ satisfy* (51)–(52) *such that either $\int_0^T \|\phi\|_{TV}dt < \infty$ or $\int_0^T \|\psi\|_{TV}dt < \infty$ for any $T > 0$. Then the averaged 3D Euler equations have a unique smooth global solution satisfying*

$$\|\omega(t)\|_{H^m} \le C(T)\|\omega(0)\|_{H^m}), \quad 0 \le t \le T,$$

for any $T > 0$.

Remark. As we mentioned before, the above result has a clear geometric interpretation. It implies that if the 1D restriction of the level set function ϕ or ψ has a total variation which is integrable in time, then there is no finite time blow-up. This excludes the possibility of a finite number of isolated singularities when vorticity is considered as a 1D function by fixing the other two variables. In particular, if there is a finite time singularity, the 1D restriction of vorticity must be highly oscillatory at the singularity time, and the singularities are dense in the singular region.

Proof. Recall that

$$u_\alpha = (1-\alpha^2\triangle)^{-1}\nabla \times (-\triangle)^{-1}\omega.$$

Thus we have

$$\nabla u_\alpha = (1-\alpha^2\triangle)^{-1}R\omega,$$

where $R = \nabla\nabla\times(-\triangle)^{-1}$ is a Riesz operator.

First, we consider the special case when $\omega_0 \equiv 1$. In this case, we have $\omega = \nabla\phi\times\nabla\psi$ for all times. Without loss of generality, we may assume that $\int_0^T \|\psi\|_{TV}dt < \infty$. We can further rewrite $\omega = \nabla\times(\phi\nabla\psi)$.

Let $B(y)$ be the integral kernel of the operator $(1-\alpha^2\triangle)^{-1}R$ in R^3. We set $x = 0$ and omit the reference to time. Then we can express

$$\begin{aligned}|\nabla u_\alpha(0)| &= \left|\int_{R^3} B(y)\omega(y)dy\right| \\ &= \left|\int_{R^3} \nabla B(y) \times (\phi(y)\nabla\psi(y))dy\right|.\end{aligned}$$

One can show that

$$|\nabla B(y)| \le \frac{C_\alpha}{|y|^2(1+|y|^{\frac{1}{4}})}. \tag{54}$$

Let B_ϵ denote the ball centered at the origin with radius $\epsilon < 1$. Note that the level set functions ϕ and ψ are bounded for all times. Let $p > 3$, q be the conjugate of p satisfying $\frac{1}{p}+\frac{1}{q} = 1$. Further, we denote $r = \frac{3-2q}{q}$. If y_i is one of the three components of y in R^3, we denote by y' the remaining 2D vector excluding y_i. Then we have

$$\begin{aligned}|\nabla u_\alpha(0)| &= \left|\int_{B_\epsilon} + \int_{|y|\ge\epsilon} \nabla B(y) \times (\phi(y)\nabla\psi(y))dy\right| \\ &\le \|\phi\|_\infty \left(\epsilon^{\frac{3-2q}{q}}\|\nabla\psi\|_{L^p} + \sum_{i=1}^{3}\int_{|y'|^2+|y_i|^2\ge\epsilon^2} \frac{dy'}{(|y_i|^2+|y'|^2)(1+|y|^{\frac{1}{4}})} \int \left|\frac{\partial\psi}{\partial y_i}\right| dy_i\right)\end{aligned}$$

$$\leq \|\phi\|_\infty \left(\epsilon^r \|\nabla\psi\|_{L^p} + \|\psi\|_{TV} \int_{R^2} \frac{dy'}{(\epsilon^2 + |y'|^2)(1+|y|^{\frac{1}{4}})} \right.$$
$$\left. + \|\psi\|_{TV} \int_{|y'|\geq\epsilon} \frac{dy'}{|y'|^2(1+|y'|^{\frac{1}{4}})} \right)$$
$$\leq \|\phi\|_\infty \left(\epsilon^r \|\nabla\psi\|_{L^p} + \|\psi\|_{TV} \log\frac{1}{\epsilon} \right),$$

where we have used the Hölder inequality in the estimate for the inner part. Note that $\|\phi\|_\infty \leq \|\phi_0\|_\infty$. By setting $\epsilon^r(e + \|\nabla\psi\|_{L^p}) = 1$, we obtain

$$|\nabla u_\alpha(0)| \leq C\left(1 + \|\psi\|_{TV} \log(\|\nabla\psi\|_{L^p} + e)\right). \tag{55}$$

Differentiating (52) with respect to x, we obtain

$$(\nabla\psi)_t + (u_\alpha \cdot \nabla)(\nabla\psi) + \nabla u_\alpha \nabla\psi = 0. \tag{56}$$

Performing the energy estimate to (56), we get

$$\frac{\partial}{\partial t}\|\nabla\psi\|_{L^p} \leq \|\nabla u_\alpha\|_\infty \|\nabla\psi\|_{L^p}$$
$$\leq C\left(1 + \|\psi\|_{TV} \log(\|\nabla\psi\|_{L^p} + e)\right) \|\nabla\psi\|_{L^p}.$$

The Gronwall inequality then implies that

$$\|\nabla\psi\|_{L^p} \leq C(T), \tag{57}$$

provided that $\int_0^T \|\psi\|_{TV} dt < \infty$. Substituting (57) back into (55), we conclude that

$$\int_0^T \|\nabla u_\alpha\|_\infty dt \leq C \int_0^T \|\psi\|_{TV} dt \leq C(T). \tag{58}$$

The bound on $\int_0^T \|\nabla u_\alpha\|_\infty dt$ immediately gives the maximum bound on $\nabla\psi$ from (56). Similarly, we obtain the maximum bound for $\nabla\phi$. Combining the maximum estimates for $\nabla\psi$ and $\nabla\phi$, we obtain the maximum bound for vorticity ω. Then it is a standard argument to prove the energy estimate for ω in the H^m norm using

$$\frac{\partial}{\partial t}\|\omega\|_{H^m} \leq C(\|\nabla u_\alpha\|_\infty + \|\omega\|_\infty)\|\omega\|_{H^m}.$$

It remains to comment on the more general case when $\omega_0 \neq 1$. Note that

$$\omega_0(\phi,\psi)\nabla\phi\times\nabla\psi = \nabla(\omega_0\phi)\times\nabla\psi - \phi(\omega_0)_\phi \nabla\phi\times\nabla\psi$$
$$= \nabla\times(\omega_0\phi\nabla\psi) - \phi(\omega_0)_\phi\nabla\phi\times\nabla\psi.$$

Define

$$h(\phi,\psi) = \int_0^\phi s(\omega_0)_\phi(s,\psi)ds.$$

Then we have

$$\nabla h(\phi,\psi) = h_\phi\nabla\phi + h_\psi\nabla\psi$$
$$= \phi(\omega_0)_\phi\nabla\phi + h_\psi\nabla\psi.$$

This implies that

$$\begin{aligned}(\phi(\omega_0)_\phi \nabla\phi) \times \nabla\psi &= \nabla h \times \nabla\psi - h_\psi \nabla\psi \times \nabla\psi \\ &= \nabla \times (h\nabla\psi).\end{aligned}$$

Note that h is bounded since both ϕ and $(\omega_0)_\phi$ are bounded. Therefore, we can rewrite

$$\omega_0(\phi,\psi)\nabla\phi \times \nabla\psi = \nabla \times (\omega_0\phi\nabla\psi) - \nabla \times (h\nabla\psi),$$

with both $\omega_0\phi$ and h being bounded. Thus the previous argument for $\omega_0 = 1$ applies to the case when $\omega_0 \neq 1$. This completes the proof of the theorem.

REFERENCES

[1] J. T. Beale, T. Kato, and A. Majda, *Remarks on the breakdown of smooth solutions for the 3D Euler equations*. Comm. Math. Phys., 94 (1984), pp. 61–64.

[2] A. P. Calderon and A. Zygmund, *On the existence of certain singular integrals*, Acta Math., 88 (1952), pp. 85–139.

[3] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[4] S. Y. Chen, D. D. Holm, L. G. Margolin, and R. Zhang, *Direct numerical simulations of the Navier-Stokes alpha model*, Phys. D., 133 (1999), pp. 66–83.

[5] S. Y. Chen, C. Foias, D. D. Holm, E. J. Olson, E. S. Titi, and S. Wynne, *The Camassa-Holm equations as a closure model for turbulent channel and pipe flows*, Phys. Fluids, 11 (1999), pp. 2343–2353.

[6] A. J. Chorin and J. E. Marsden, *A Mathematical Introduction to Fluid Mechanics*, 3rd ed., Springer-Verlag, New York, 1993.

[7] A. Clebsch, *Uber die integration der hydrodynamicschen gleichungen*, J. Reine Angew. Math., 56 (1859), pp. 1–10.

[8] P. Constantin, D. Cordoba, and J. Wu, *On the critical dissipative quasi-geostrophic equation*, Indiana Univ. Math. J., 50 (2001), pp. 97–107.

[9] P. Constantin, A. Majda, and E. Tabak, *Formation of strong fronts in the 2D quasi-geostrophic thermal active scalar*, Nonlinearity, 7 (1994), pp. 1495–1533.

[10] P. Constantin, C. Fefferman, and A. Majda, *Geometric constraints on potentially singular solutions for the 3D Euler equation*, Comm. Partial Differential Equations, 21 (1996), pp. 559–571.

[11] J. Deng, T. Y. Hou, and X. Yu, *Geometric properties and the non-blow-up of the three-dimensional Euler equation*, Comm. Partial Differential Equations, 30 (2005), pp. 225–243.

[12] J. Deng, T. Y. Hou, and X. Yu, *A level set formulation for the 3D incompressible Euler equations*, Methods Appl. Anal., 12 (2005), pp. 427–440.

[13] C. Foias, D. D. Holm, and E. S. Titi, *The three dimensional viscous Camassa-Holm equations, and their relation to the Navier-Stokes equations and turbulence theory*, J. Dynam. Differential Equations, 14 (2002), pp. 1–35.

[14] O. H. Hald, *Convergence of vortex blob methods for Euler's equations*, III., SIAM J. Numer. Anal., 24 (1987), pp. 538–582.

[15] D. D. Holm, J. E. Marsden, and T. S. Ratiu, *Euler-Poincaré models of ideal fluids with nonlinear dispersion*, Phys. Rev. Lett., 349 (1998), pp. 4173–4177.

[16] D. D. Holm, J. E. Marsden, and T. S. Ratiu, *Euler-Poincaré equations and semidirect products with applications to continuum theories*, Adv. Math., 137 (1998), pp. 1–81.

[17] T. Y. Hou and C. Li, *Global well-posedness of the viscous Boussinesq equations*, Discrete Contin. Dyn. Syst., 12 (2005), pp. 1–12.

[18] F. John and L. Nirenberg, *On functions of bounded mean oscillation*, Comm. Pure Appl. Math., 14 (1961), pp. 415–426.

[19] H. Kozono and Y. Taniuchi, *Limiting case of the Sobolev inequality in BMO, with application to the Euler equations*, Comm. Math. Phys., 214 (2000), pp. 191–200.

[20] A. Majda and A. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge University Press, Cambridge, UK, 2002.

[21] J. Marsden and S. Shkoller, *Global well-posedness for the lagrangian averaged Navier-Stokes (LANS-α) equations on bounded domains*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 359 (2001), pp. 1449–1468.

[22] J. E. Marsden, T. Ratiu, and S. Shkoller, *The geometry and analysis of the averaged Euler equations and a new differmophism group*, Geom. Funct. Anal., 10 (2000), pp. 582–599.
[23] M. Oliver and S. Shkoller, *The vortex blob method as a second-grade non-Newtonian fluid*, Comm. Partial Differential Equations, 26 (2001), pp. 91–110.
[24] E. Olson and E. S. Titi, *Viscosity Versus Vorticity Stretching: Global Well-Posedness for a Family of Navier-Stokes-Alpha-Like Models*, preprint, 2005.
[25] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1987.
[26] J. Peetre, *On convolution operators leaving $L^{p,\lambda}$ spaces invariant*, Ann. Mat. Pura Appl. (4), 72 (1966), pp. 295–304.
[27] E. M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton Math. Ser., 43 1993, Princeton University Press, Princeton, NJ.
[28] M. E. Taylor, *Pseudodifferential Operators and Nonlinear PDE*, Birkhäuser, Boston, 1991.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 795–823

A FULLY NONLINEAR VERSION OF THE INCOMPRESSIBLE EULER EQUATIONS: THE SEMIGEOSTROPHIC SYSTEM*

G. LOEPER†

Abstract. The semigeostrophic equations are used in meteorology. They appear as a variant of the two-dimensional Euler incompressible equations in vorticity form, where the Poisson equation that relates the stream function and the vorticity field is just replaced by the fully nonlinear elliptic Monge–Ampère equation. This work gathers new results concerning the semigeostrophic equations: existence and stability of measure-valued solutions, existence and uniqueness of solutions under certain continuity conditions for the density, and convergence to the incompressible Euler equations.

Key words. semigeostrophic equations, Monge–Ampère equations, optimal transportation

AMS subject classifications. 35J60, 76U05, 76B03

DOI. 10.1137/050629070

1. Introduction.

The semigeostrophic equations are an approximation of the Euler equations of fluid mechanics used in meteorology to describe atmospheric flows. They are believed (see [12]) to be an efficient model in describing frontogenesis. Different versions (incompressible [1], shallow water [10], compressible [11]) of this model have been studied, and we will focus here on the incompressible two-dimensional (2-d) and three-dimensional (3-d) versions. The 3-d model describes the behavior of an incompressible fluid in a domain $\Omega \subset \mathbb{R}^3$. To the evolution in Ω is associated a motion in a "dual" space, described by the following nonlinear transport equation:

$$\begin{aligned} &\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0, \\ &\mathbf{v} = (\nabla \Psi(x) - x)^{\perp}, \\ &\det D^2 \Psi = \rho, \\ &\rho(t=0) = \rho^0. \end{aligned}$$

Here ρ^0 is a probability measure on $\mathbb{R}^3$, and for every $\mathbf{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$, $\mathbf{v}^{\perp}$ stands for $(-v_2, v_1, 0)$. The velocity field is recovered at each time step by solving a Monge–Ampère equation in the sense of the polar factorization of maps (see [3]), i.e., in the sense that Ψ is convex from $\mathbb{R}^3$ to $\mathbb{R}$ and satisfies $\nabla \Psi_{\#} \rho = \chi_\Omega \mathcal{L}^3$, where $\mathcal{L}^3$ is the Lebesgue measure of $\mathbb{R}^3$, and χ_Ω is the indicator function of Ω. It is imposed as a compatibility condition that Ω has Lebesgue-measure one. This model arises as an approximation of the primitive equations of meteorology, and we shall give a brief idea of the derivation of the model, although the reader interested in more details should refer to [12].

In this work we will deal with various questions related to the semigeostrophic (SG) system: existence and stability of measure-valued solutions, existence and uniqueness of smooth solutions, and finally convergence toward the incompressible Euler equations in 2-d. As stated in the title, we will exploit throughout the paper the

*Received by the editors April 12, 2005; accepted for publication (in revised form) February 13, 2006; published electronically August 7, 2006.
http://www.siam.org/journals/sima/38-3/62907.html
†Institut Camille Jordan, Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France (loeper@math.univ-lyon1.fr).

strong analogy with the 2-d incompressible Euler equations that we recall here:

$$\begin{aligned}
&\partial_t \omega + \nabla \cdot (\omega \mathbf{v}) = 0,\\
&\mathbf{v} = (\nabla \Phi)^{\perp},\\
&\Delta \Phi = \omega,\\
&\omega(t=0) = \omega^0.
\end{aligned}$$

We recognize clearly that the vorticity ω here plays the role of the density ρ in the SG system. One obtains the SG system just by replacing the Poisson equation $\Delta\Phi = \omega$ by the Monge–Ampère equation $\det(I + D^2\Phi) = \omega$. (However, note that the density ρ does not have a clear physical interpretation since it is a density in a dual space.) From this analogy, and inspired by the well-developed mathematical theory on the 2-d Euler equations (see [21] for instance), the goal of this paper is twofold:

The first goal is to the study of the initial value problem for the SG system. We will first establish a global existence result for weak measure-valued solutions, hence giving a framework for weak solutions that strictly contains the results obtained in previous works. We will also obtain local smooth solutions, trying to lower as much as possible the requirement on the initial data, and prove uniqueness in a certain class of smooth solutions. This well-posedness result for smooth initial data will be our main result.

The second goal is to give some rigorous mathematical justification of the derivation of the SG system from the 2-d Euler equations. In an attempt in this direction, we will show that in some asymptotic regime (namely, "small" solutions over a long time) the SG system and the 2-d Euler system are asymptotically close.

We will use a combination of various techniques: The SG system is a transport equation, and we will study it as such, using either the Eulerian or the Lagrangian point of view. Since the coupling between density and velocity field involves a Monge–Ampère equation, we will also rely on the regularity theory developed for this fully nonlinear elliptic equation, which is much more recent and far less known than the results on solutions to the Poisson equation. More originally we will use optimal transportation and the technique developed in [18] to show uniqueness of certain solutions. Note (though this is a coincidence) that optimal transportation will appear earlier in the present paper in the derivation of the SG system. Finally the proof of convergence toward the incompressible Euler equations will be made using modulated energy techniques, a general technique (also known as the "relative entropy method," documented in [13]) used for the asymptotic study of hyperbolic systems.

The paper is organized as follows: In the next paragraph we give a brief description of the derivation of the SG system from the Euler incompressible equations. To formulate rigorously the system, we then review the results concerning optimal transportation and polar factorization of maps, which are key concepts used throughout this paper (section 1.2). We are then able to formulate the SG system, both in its Lagrangian and Eulerian (or dual) forms (sections 1.3 and 1.4).

Section 1.5 is dedicated to a longer discussion on the results obtained and gives a sketch of some of the crucial arguments. This section closes the introduction. Each of the following sections is dedicated to the proof of one of the results.

Section 2 is devoted to the existence of a weak measure-valued solution, in section 3 we show existence of Dini continuous solutions, in section 4 we show uniqueness of solutions with Hölder continuous density, and in section 5 we show the convergence of solutions of SG equations toward solutions of the 2-d Euler incompressible equations.

All those results will be reviewed and discussed in greater detail in section 1.5, after we have derived the SG equations.

1.1. Derivation of the SG equations. We now give for the sake of completeness a brief and simplified idea of the derivation of the system, inspired by [1]; more complete arguments can be found in [12].

Lagrangian formulation. We start with the 3-d incompressible Euler equations with constant Coriolis parameter f in a domain Ω.

$$\begin{aligned}&\frac{D\mathbf{v}}{Dt} + f\mathbf{v}^\perp = \frac{1}{\rho}\nabla p - \nabla\varphi,\\ &\nabla\cdot\mathbf{v} = 0, \quad \frac{D\rho}{Dt} = 0,\\ &\mathbf{v}\cdot\partial\Omega = 0,\end{aligned}$$

where $\frac{D\cdot}{Dt}$ stands for $\partial_t + \mathbf{v}\cdot\nabla$, and we still use $\mathbf{v}^\perp = (-v_2, v_1, 0)$. The term $\nabla\varphi$ denotes the gravitational effects (here we will take $\varphi = gx_3$ with constant g), and the term $f\mathbf{v}^\perp$ is the Coriolis force due to rotation of the Earth. For large scale atmospheric flows, the Coriolis force $f\mathbf{v}^\perp$ dominates the advection term $\frac{D\mathbf{v}}{Dt}$ and renders the flow mostly 2-d. We use the hydrostatic approximation, $\partial_{x_3} p = -\rho g$, and restrict ourselves to the case $\rho \equiv 1$.

Keeping only the leading order terms leads to the geostrophic balance

$$\mathbf{v}_g = -f^{-1}\nabla^\perp p,$$

which defines $\mathbf{v}_g$, the geostrophic wind. Decomposing $\mathbf{v} = \mathbf{v}_g + \mathbf{v}_{ag}$, where the second component is the ageostrophic wind, a supposed small departure from the geostrophic balance, the SG system reads

$$\begin{aligned}\frac{D\mathbf{v}_g}{Dt} + f\mathbf{v}^\perp &= \nabla_H p,\\ \nabla\cdot\mathbf{v} &= 0,\end{aligned}$$

where $\nabla_H = (\partial_{x_1}, \partial_{x_2}, 0)$. Note, however, that the advection operator $\partial_t + \mathbf{v}\cdot\nabla$ still uses the full velocity $\mathbf{v}$. Introducing the potential

$$\Phi = \frac{1}{2}|x_H|^2 + f^{-2}p,$$

with $x_H = (x_1, x_2, 0)$, we obtain

$$\frac{D}{Dt}\nabla\Phi(t,x) = f(x - \nabla\Phi(t,x))^\perp.$$

We introduce the Lagrangian map $\mathbf{g} : \Omega\times\mathbb{R}^+ \mapsto \Omega$, giving the position at time t of the particle of fluid located at x_0 at time 0. The previous equation means that if for fixed $x \in \Omega$ we consider the trajectory in the "dual" space, defined by $X(t,x) = \nabla\Phi(t, \mathbf{g}(t,x))$, we have

$$\partial_t X(t,x) = f\left(\mathbf{g}(t,x) - X(t,x)\right)^\perp.$$

By rescaling the time, we can set $f = 1$. Under this form the system looks underdetermined: Indeed Φ is unknown; however, we have the condition $X(t,x) =$

$\nabla\Phi(t, \mathbf{g}(t,x))$. Moreover, the motion of the fluid being incompressible and contained in Ω, the map $\mathbf{g}(t,\cdot)$ must be measure preserving in Ω for each t, i.e.,

$$\mathcal{L}^3(\mathbf{g}(t)^{-1}(B)) = \mathcal{L}^3(B)$$

for each $B \subset \Omega$ measurable (where $\mathcal{L}^3$ denotes the Lebesgue measure of $\mathbb{R}^3$). We shall hereafter denote by $G(\Omega)$ the set of all such measure preserving maps. Then Cullen's stability criteria [12] asserts that the potential Φ should be convex for the system to be stable to small perturbations of a particle's positions in the x space. Indeed the convexity of Φ asserts that $\nabla\Phi$ minimizes some potential energy (the reader interested in a more detailed explanation of this variational principle should refer to [12]).

Hence, for each t, Φ must be a convex function such that

$$X(t,\cdot) = \nabla\Phi(t, \mathbf{g}(t,\cdot)),$$

with $\mathbf{g}(t,\cdot) \in G(\Omega)$.

In the next paragraph we shall see that, under very mild assumptions on X, this decomposition, called polar factorization, can only happen for a unique choice of $\mathbf{g}$ and $\nabla\Phi$. Now if Φ^* is the Legendre transform of Φ,

$$\Phi^*(y) = \sup_{x\in\Omega} x\cdot y - \Phi(x),$$

then $\nabla\Phi$ and $\nabla\Phi^*$ are inverse maps of each other, and the SG system then reads

$$\frac{DX}{Dt} = \left(\nabla\Phi^*(X(t)) - X(t)\right)^{\perp},$$

$$\nabla\Phi^*(t)\circ X(t) \in G(\Omega).$$

In this context, $X(t)$ is thus the dual trajectory to the physical trajectory $\mathbf{g}(t)$, and $(\nabla\Phi^*(X(t)) - X(t))^{\perp}$ is up to a multiplicative constant, the geostrophic wind at point $\mathbf{g}(t) = \nabla\Phi^*(X(t))$.

In the next paragraph, we review the results concerning the existence and uniqueness of the gradients $\nabla\Phi, \nabla\Phi^*$.

1.2. Polar factorization of vector valued maps. The polar factorization of maps has been discovered by Brenier in [3]. It was later extended to the case of general Riemannian manifolds by McCann in [23].

The Euclidean case. Let Ω be a fixed bounded domain of $\mathbb{R}^d$ of Lebesgue measure 1 and satisfying the condition $\mathcal{L}^d(\partial\Omega) = 0$. We consider a mapping $X \in L^2(\Omega;\mathbb{R}^d)$. We will also consider the push-forward of the Lebesgue measure of Ω by X, which we will denote by $X_{\#}\chi_\Omega\mathcal{L}^d = d\rho$ (or, in short, $X_{\#}dx$) and which is defined by

$$\forall f \in C_b^0(\mathbb{R}^d),\ \int_{\mathbb{R}^d} f(x)\ d\rho(x) = \int_\Omega f(X(x))\ dx.$$

Let $\mathcal{P}$ be the set of probability measures $\mathbb{R}^d$, and let $\mathcal{P}_a^2$ be the subset of $\mathcal{P}$ where the subscript a means absolutely continuous with respect to the Lebesgue measure (or equivalently that have a density in $L^1(\mathbb{R}^d)$), and the superscript 2 means with finite second moment, i.e., such that

$$\int_{\mathbb{R}^d} |x|^2 d\rho(x) < +\infty.$$

Note that for $X \in L^2(\Omega, \mathbb{R}^d)$, the measure $\rho = X_{\#}dx$ has necessarily a finite second moment, and thus belongs to $\mathcal{P}^2$.

THEOREM 1.1 (Brenier [3]). *Let Ω be as above, $X \in L^2(\Omega; \mathbb{R}^d)$, and $\rho = X_{\#}dx$.*

1. *There exists a unique up to a constant convex function, which will be denoted $\Phi[\rho]$, such that*

$$\forall f \in C_b^0(\mathbb{R}^d),\ \int_\Omega f(\nabla\Phi[\rho](x))\ dx = \int_{\mathbb{R}^d} f(x)d\rho(x).$$

2. *Let $\Psi[\rho]$ be the Legendre transform of $\Phi[\rho]$. If $\rho \in \mathcal{P}_a^2$, then $\Psi[\rho]$ is the unique up to a constant convex function satisfying*

$$\forall f \in C_b^0(\Omega),\ \int_{\mathbb{R}^d} f(\nabla\Psi[\rho](x))\ d\rho(x) = \int_\Omega f(x)dx.$$

3. *If $\rho \in \mathcal{P}_a^2$, X admits the following unique polar factorization:*

$$X = \nabla\Phi[\rho] \circ g,$$

with $\Phi[\rho]$ convex, g measure preserving in Ω.

Remark. $\Psi[\rho], \Phi[\rho]$ depend only on ρ, and are solutions (in some weak sense) in $\mathbb{R}^d$ and Ω, respectively, of the Monge–Ampère equations

$$\begin{aligned} \det D^2\Psi &= \rho, \\ \rho(\nabla\Phi)\ \det D^2\Phi &= 1. \end{aligned}$$

When Ψ and Φ are not in C_{loc}^2 these equations can be understood in the viscosity (or Alexandrov) sense or in the sense of Theorem 1.1, which is strictly weaker. For the consistency of the different weak formulations and regularity issues, the reader can refer to [8].

The periodic case. The polar factorization theorem has been extended to Riemannian manifolds in [23] (see also [9] for the case of the flat torus). In this case, we consider a mapping $X : \mathbb{R}^d \mapsto \mathbb{R}^d$ such that for all $\vec{p} \in \mathbb{Z}^d$, $X(\cdot + \vec{p}) = X + \vec{p}$. Then $\rho = X_{\#}dx$ is a probability measure on $\mathbb{T}^d$. We define $\Psi[\rho], \Phi[\rho]$ through the following.

THEOREM 1.2. *Let $X : \mathbb{R}^d \to \mathbb{R}^d$ be as above, with $\rho = X_{\#}dx$.*

1. *Up to an additive constant there exists a unique convex function $\Phi[\rho]$ such that $\Phi[\rho](x) - x^2/2$ is $\mathbb{Z}^d$-periodic (and thus $\nabla\Phi[\rho](x) - x$ is $\mathbb{Z}^d$-periodic), and*

$$\forall f \in C^0(\mathbb{T}^d),\ \int_{\mathbb{T}^d} f(\nabla\Phi[\rho](x))\ dx = \int_{\mathbb{T}^d} f(x)\ d\rho(x).$$

2. *Let $\Psi[\rho]$ be the Legendre transform of $\Phi[\rho]$. If ρ is Lebesgue integrable, then $\Psi[\rho]$ is the unique up to a constant convex function satisfying the statement that $\Psi[\rho](x) - x^2/2$ is $\mathbb{Z}^d$-periodic (and thus $\nabla\Psi[\rho](x) - x$ is $\mathbb{Z}^d$-periodic), and*

$$\forall f \in C^0(\mathbb{T}^d),\ \int_{\mathbb{T}^d} f(\nabla\Psi[\rho](x))\ d\rho(x) = \int_{\mathbb{T}^d} f(x)\ dx.$$

3. *If ρ is Lebesgue integrable, X admits the following unique polar factorization:*

$$X = \nabla\Phi[\rho] \circ g,$$

with g measure preserving from $\mathbb{T}^d$ into itself, and $\Phi[\rho]$ convex, $\Phi[\rho] - |x|^2/2$ periodic.

Remark 1. From the periodicity of $\nabla\Phi[\rho](x) - x, \nabla\Psi[\rho](x) - x$, for every f $\mathbb{Z}^d$-periodic, $f(\nabla\Psi[\rho]), f(\nabla\Phi[\rho])$ are well defined on $\mathbb{R}^d/\mathbb{Z}^d$.

Remark 2. Both in the periodic and nonperiodic cases, the definitions of $\Psi[\rho]$ and $\Phi[\rho]$ make sense if ρ is absolutely continuous with respect to the Lebesgue measure. If not, the definition and uniqueness of $\Phi[\rho]$ is still valid, as well as the property $\nabla\Phi[\rho]_{\#}\rho = \chi_\Omega \mathcal{L}^d$. The definition of $\Psi[\rho]$ as the Legendre transform of $\Phi[\rho]$ is still valid also, but then the expression $\int f(\nabla\Psi[\rho](x))\, d\rho(x)$ does not necessarily make sense since $\nabla\Psi[\rho]$ is not necessarily continuous, and hence not defined $d\rho$ almost everywhere. Moreover the polar factorization no longer holds.

Remark 3. We have (see [9]) the unconditional bound

$$\|\nabla\Psi[\rho](x) - x\|_{L^\infty(\mathbb{T}^d)} \leq \sqrt{d}/2,$$

which will be useful later on.

1.3. Lagrangian formulation of the SG system. From Theorems 1.1 and 1.2 the Lagrangian formulation of the SG equation then becomes

$$\frac{DX}{Dt} = [\nabla\Psi(X) - X]^\perp, \tag{1}$$

$$\Psi = \Psi[\rho], \quad \rho = X_{\#}dx. \tag{2}$$

1.4. Eulerian formulation in dual variables. In both cases (periodic and nonperiodic) we thus investigate the following system, which will be referred to as *SG in dual variables* (but we will only say SG hereafter): We look for a time-dependent probability measure $t \to \rho(t, \cdot)$ satisfying

$$\partial_t\rho + \nabla \cdot (\rho\mathbf{v}) = 0, \tag{3}$$

$$\mathbf{v}(t,x) = (\nabla\Psi[\rho(t)](x) - x)^\perp, \tag{4}$$

$$\rho(t=0) = \rho^0. \tag{5}$$

Global existence of weak solutions (which are defined below) of this system with L^p initial data for $p \geq 1$ has been shown in [1], [10], [20].

1.5. Results. In this work we deal with various mathematical problems related to this system: We first extend the notion of weak solutions that have been shown to exist for $\rho \in L^\infty(\mathbb{R}_+, L^q(\mathbb{R}^3))$, $q > 1$ (see [1], [10]), and then for $\rho \in L^\infty(\mathbb{R}_+, L^1(\mathbb{R}^3))$ (see [20]), to the more general case of bounded measures. The question of existence of measure-valued solutions was raised and left unanswered in those papers, and we show here existence of global solutions to the Cauchy problem with initial data a bounded compactly supported measure, and we show the weak stability/compactness of these *weak measure* solutions.

Then we show existence of continuous solutions; more precisely, we show local existence of solutions with Dini continuous (see (12)) density. For these solutions, the velocity field is then C^1 and the Lagrangian system (1)–(2) is defined everywhere.

This proof relies heavily on the available regularity results on solutions to the Monge–Ampère equation (Theorem 3.1). Note that the Dini condition is the lowest condition known on the right-hand side of the Poisson equation that enforces C^2 regularity for the solution. Our result is not totally satisfactory since it does not provide existence of a global smooth solution, which is the case for the 2-d incompressible Euler equation. The reason for this more powerful result is that for the Poisson equation

$$\Delta\Phi = \omega,$$

ω bounded implies that $\nabla\Phi$ is log-Lipschitz. This continuity is slightly weaker than the Lipschitz continuity, but allows one to define a Hölder continuous flow (see [21]). Moreover, the flow being incompressible, this implies (when $d = 2$) that the vorticity is just transported along the streamlines. The construction of global smooth solutions can then be achieved only by using those two arguments.

For the SG system, solutions to

$$\det D^2\Psi = \rho$$

are only $C^{1,\alpha}$ when ρ is merely bounded. This is not enough to build a continuous flow and prevents us from obtaining the same results as for Euler.

We also show uniqueness in the class of Hölder continuous solutions (a subclass of Dini continuous solutions). This proof uses in a crucial way the optimal transportation of measures by convex gradients and its regularity properties, and can be adapted to give a new proof of uniqueness for solutions of the 2-d Euler equation with bounded vorticity, but also for a broad class of nonlinearly coupled system. The typical application is a density evolving through a transport equation where the velocity field depends on the gradient of a potential. The potential is obtained by solving an elliptic equation, where the density appears in the right-hand side. Well-known examples of such cases are the Vlasov–Poisson and Euler–Poisson systems (see [18]). We point out that the results of existence and uniqueness obtained here are all obtained by working in a purely Lagrangian framework.

Finally, in the 2-d case, we study the convergence of the system to the Euler incompressible equations; this convergence is expected for ρ close to 1, since formally expanding $\Psi = x^2/2 + \epsilon\psi$, and linearizing the determinant around the identity matrix, we get

$$\rho = \det D^2\Psi = 1 + \epsilon\Delta\psi + O(\epsilon^2),$$

and the Monge–Ampère equation turns into the Poisson equation

$$\Delta\psi = \frac{\rho - 1}{\epsilon} =: \mu.$$

We then perform the change of time scale $t \to t/\epsilon$ and consider now $\mu^\epsilon(t) := \mu(\frac{t}{\epsilon})$. Then μ^ϵ solves

$$\begin{aligned} \partial_t\mu^\epsilon + \nabla\cdot(\mu^\epsilon\nabla^\perp\psi^\epsilon) &= O(\epsilon), \\ \Delta\psi^\epsilon &= \mu^\epsilon, \end{aligned}$$

which, when we set $O(\epsilon) = 0$, we recognize as the vorticity formulation of the 2-d Euler incompressible equation.

Let us comment on this scaling: We consider a small solution to SG, i.e., a solution where $\rho - 1$ is small. We then expand this solution by a factor of ϵ^{-1} and study it on a time scale of order ϵ^{-1}.

From a physical point of view, this asymptotic study may be seen as a justification of the consistency of the SG approximation when $d = 2$. Indeed, when $d = 2$, the Euler equations are not affected by the Coriolis force, i.e., the solutions to

$$\partial_t v + v \cdot \nabla v = -\nabla p, \tag{6}$$

$$\mathrm{div} v = 0 \tag{7}$$

and to

$$\partial_t v + v \cdot \nabla v + f v^\perp = -\nabla p, \tag{8}$$

$$\mathrm{div} v = 0 \tag{9}$$

are the same, since the term $v^\perp$ can be considered a pressure term (remember that $v = \nabla^\perp \Phi$). The term f is just a time scale, and the geostrophic regime is the one where $\frac{v}{L} \ll f$, where v denotes here the typical size of v and L the typical space scale of the system. Then, note that if $v(t,x)$ is a solution to (6), so is $v^\epsilon = \epsilon v(\epsilon t, x)$. But the ratio $\frac{v^\epsilon}{L}$ goes to 0 as ϵ goes to 0 (note that the space scales for v and v^ϵ are the same). Hence, in the limit $\epsilon \to 0$, i.e., for small solutions to Euler, the geostrophic approximation should be valid. It is precisely in this regime that we show that the SG system and the incompressible Euler system are asymptotically close to each other, since for SG, a small solution is one where ρ is close to 1. Hence what we show is the following: Let ρ^0 be a "small" initial data for SG. Consider μ^ϵ obtained from ρ as explained above; then μ^ϵ is close to some ω where ω solves the 2-d Euler incompressible equation in vorticity form

$$\partial_t \omega + \nabla \cdot (\omega \nabla^\perp \phi) = 0,$$
$$\Delta \phi = \omega.$$

In other words, when ρ goes to 1, ρ is equivalent to a solution of Euler, on a time that goes to infinity.

The study of this "quasi-neutral" limit is done in two different ways: One uses a modulated energy method similar to the one used in [4] and [5] and is valid for weak solutions. The other uses a more classical expansion of the solution, and regularity estimates, and is similar to the method used in [17]. The second method also yields almost global solutions: Indeed, it will be shown in this paper that smooth (say, with Lipschitz density) solutions exists in short time. The asymptotic study of the convergence to Euler shows that the Lipschitz bound on the solution remains valid on a time that goes to infinity when the solution is chosen with an initial condition that converges toward the uniform density.

2. Measure-valued solutions.

2.1. A new definition of weak solutions. We have first the following classical weak formulation of (3): $\rho \in C(\mathbb{R}+, L^1(\mathbb{R}^3) - w)$ is said to be a weak solution of SG if

$$\forall T > 0, \quad \forall \varphi \in C_c^\infty([0,T] \times \mathbb{R}^2),$$
$$\int \partial_t \varphi \, \rho + \nabla \varphi \cdot (\nabla \Psi[\rho] - x)^\perp \, \rho \, dt dx = \int \varphi(T,x) \rho(T,x) dx - \int \varphi(0,x) \rho(0,x) dx,$$

where, for all t, $\Psi[\rho]$ is as in Theorem 1.1. The problematic part in the case of measure-valued solutions is to give sense to the product $\rho\nabla\Psi[\rho]$ since at the point where ρ is singular $\nabla\Psi[\rho]$ is unlikely to be continuous. Therefore we use Theorem 1.1 to write, for any $\rho \in \mathcal{P}_a^2(\mathbb{R}^3)$,

$$\forall \varphi \in C_c^\infty(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \rho \nabla\Psi[\rho]^\perp \cdot \nabla\varphi = \int_\Omega x^\perp \cdot \nabla\varphi(\nabla\Phi[\rho])$$

(the integrals would be performed over $\mathbb{T}^3$ in the periodic case). The property $\nabla\Phi[\rho]_{\#}\chi_\Omega\mathcal{L}^3 = \rho$ is still valid when ρ is only a measure with finite second moment (see Remark 2 after Theorem 1.2). Therefore, the formulation on the right-hand side extends unambiguously to the case where $\rho \notin L^1(\mathbb{R}^2)$.

Geometric interpretation. This weak formulation has a natural geometric interpretation: At a point where $\Psi[\rho]$ is not differentiable, and thus where $\partial\Psi[\rho]$ is not reduced to a single point, $\nabla\Psi[\rho]$ should be replaced by $\bar\partial\Psi[\rho]$, the center of mass of the (convex) set $\partial\Psi[\rho]$. The function $\bar\partial\Psi[\rho]$ coincides Lebesgue almost everywhere with $\nabla\Psi$, and is defined as follows.

DEFINITION 2.1. *The map $\bar\partial\Psi[\rho]$ is defined at every point x by the center of mass with respect to the Lebesgue measure of the set $\partial\Psi[\rho](x)$. In other words, if $\partial\Psi[\rho](x)$ is a k-dimensional convex set, we have*

$$\bar\partial\Psi[\rho](x) = \int_{\partial\Psi[\rho](x)} y \, d\mathcal{L}^k(y).$$

This provides motivation for the following definition of weak measure solutions.

DEFINITION 2.2. *For all $t \in [0,T]$, let $\rho(t)$ be a probability measure of $\mathbb{R}^3$. It is said to be a* weak measure solution *to SG if*

1. *the time-dependent probability measure ρ belongs to $C([0,T], \mathcal{P} - w*)$;*
2. *there exists $t \to R(t)$ nondecreasing such that for all $t \in [0,T]$, $\rho(t,\cdot)$ is supported in $B(0,R(t))$;*
3. *for all $T > 0$ and for all $\varphi \in C_c^\infty([0,T]\times\mathbb{R}^3)$ we have*

$$\begin{aligned} &\int_{[0,T]\times\mathbb{R}^3} \partial_t\varphi(t,x)\ d\rho(dt,x) + \int_{[0,T]\times\Omega} \nabla\varphi(t,\nabla\Phi[\rho(t)](x))\cdot x^\perp\ dtdx \\ &\quad - \int_{[0,T]\times\mathbb{R}^3} \nabla\varphi(t,x)\cdot x^\perp\ d\rho(dt,x) \\ &\quad = \int \varphi(T,x) d\rho(T,x)\ dx - \int \varphi(0,x) d\rho(0,x)\ dx. \end{aligned} \tag{10}$$

This definition is consistent with the classical definition of weak solutions if, for all t, $\rho(t,\cdot)$ is absolutely continuous with respect to the Lebesgue measure.

2.2. Result. Here we prove the following.

THEOREM 2.3.

1. *Let ρ^0 be a probability measure compactly supported. There exists a global weak measure solution to the SG system in the sense of Definition 2.2.*
2. *For any $T > 0$, if $(\rho_n)_{n\in\mathbb{N}}$ is a sequence of weak measure solutions on $[0,T]$ to SG with initial data $(\rho_n^0)_{n\in\mathbb{N}}$, supported in B_R for some $R > 0$ independent of n, the sequence $(\rho_n)_{n\in\mathbb{N}}$ is precompact in $C([0,T],\mathcal{P} - w*)$ and every converging subsequence converges to a weak measure solution of SG.*

Proof of Theorem 2.3. We first show the weak stability of the formulation of Definition 2.2 and the compactness of weak measure solutions. We then use this result to obtain global existence of solutions to the Cauchy problem with initial data a bounded measure.

Weak stability of solutions. We consider a sequence $(\rho_n)_{n\in\mathbb{N}}$ of solutions of SG in the sense of Definition 2.2. The sequence is uniformly compactly supported at time 0. We first show that there exists a nondecreasing function $R(t)$ such that $\rho_n(t)$ is supported in $B(R(t))$ for all t, n.

LEMMA 2.4. *Let $\rho \in C([0,T], \mathcal{P}(\mathbb{R}^3) - w*)$ satisfy (10), and let $\rho^0 = \rho(t=0)$ be supported in $B(0,R^0)$. Then $\rho(t)$ is supported in $B(0, R^0 + C_\Omega t)$, $C_\Omega = \sup_{y\in\Omega}\{|y|\}$.*

Proof. Consider any function $\xi_\epsilon(t,r) \in C^\infty([0,T]\times\mathbb{R})$ such that

$$\begin{aligned}
&\xi_\epsilon(0,r) \equiv 1 \quad \text{if } -\infty < r \le R^0,\\
&\xi_\epsilon(0,r) \equiv 0 \quad \text{if } r \ge R^0 + \epsilon,\\
&\xi_\epsilon(t,r) = \xi_\epsilon(0, r - C_\Omega t),
\end{aligned}$$

with $\xi(0,\cdot)$ nonincreasing. Then applying (10) to the test function $\xi_\epsilon(t,|x|)$, we find

$$\begin{aligned}
\frac{d}{dt}\int \xi_\epsilon(t,|x|)\, d\rho(t,x) &= -\int \partial_r \xi_\epsilon(t,|x|) C_\Omega \, d\rho(t,x)\\
&\quad + \int_\Omega \partial_r \xi_\epsilon(t, |\nabla\Phi[\rho(t)]|) \frac{\nabla\Phi[\rho(t)]}{|\nabla\Phi[\rho(t)]|}\cdot x^\perp \, dx\\
&\ge \int_\Omega \partial_r \xi_\epsilon(t,|\nabla\Phi[\rho(t)]|)(-C_\Omega + |x|)\, dx\\
&\ge 0
\end{aligned}$$

since, by definition of C_Ω, for $x\in\Omega$, $|x| \le C_\Omega$ and ξ_ϵ is nonincreasing with respect to r. Note also that we have used $\int \nabla_x[\xi(t,|x|)]\cdot x^\perp d\rho(t,x)\, dx \equiv 0$. We know, on the other hand, that

$$\begin{aligned}
&\int_{\mathbb{R}^3} \xi_\epsilon(0,|x|) d\rho(0,x) = 1,\\
&\int_{\mathbb{R}^3} \xi_\epsilon(t,|x|) d\rho(t,x) \le 1.
\end{aligned}$$

Therefore we conclude that $\int_{\mathbb{R}^3} \xi_\epsilon(t,|x|) d\rho(t,x) \equiv 1$, which concludes the lemma by letting ϵ go to 0. ∎

From Lemma 2.4, we have

$$\begin{aligned}
&\left| -\int_{[0,T]\times\mathbb{R}^3} \nabla\varphi(t,x)\cdot x^\perp \, d\rho_n(dt,x) + \int_{[0,T]\times\Omega} \nabla\varphi(t, \nabla\Phi[\rho_n(t)](x))\cdot x^\perp \, dtdx \right|\\
&\quad \le C(T)\|\varphi\|_{L^1([0,T], C^1(B_{R(T)})}.
\end{aligned}$$

Thus from Definition 2.2, equation (10), we know that for any time $t \ge 0$, $\partial_t\rho_n(t,\cdot)$ is bounded in the dual of $L^1([0,T], C^1(\mathbb{R}^3))$ and thus in the dual of $L^1([0,T], W^{2,p}(\mathbb{R}^3))$ for $p > 3$ by Sobolev embeddings. Thus for some $p' > 1$ we have

$$\partial_t\rho_n \in L^\infty([0,T], W^{-2,p'}(\mathbb{R}^2)).$$

With the two above results, and using a classical compactness result (see [16, Chapter 1, Lemma 5.1]), we can obtain the following lemma.

LEMMA 2.5. *Let the sequence $(\rho_n)_{n\in\mathbb{N}}$ be as above. There exists $\rho \in C([0,T], \mathcal{P}-w*)$ and a subsequence $(\rho_{n_k})_{k\in\mathbb{N}}$, such that for all $t \in [0,T]$, $\rho_{n_k}(t)$ converges to $\rho(t)$ in the weak-$*$ topology of measures.*

With this lemma, we need to show that for all $\varphi \in C_c^\infty([0,T]\times\mathbb{R}^3)$ we have $\nabla\varphi(t, \nabla\Phi[\rho_n(t)])$ converging to $\nabla\varphi(t,\nabla\Phi[\rho(t)])$ whenever $\rho_n(t)$ converges weakly-$*$ to $\rho(t)$. This last step will be a consequence of the following stability theorem.

THEOREM 2.6 (Brenier [3]). *Let Ω be as above. Let $(\rho_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on $\mathbb{R}^d$, such that for all n, $\int(1+|x|^2)d\rho_n \le C$, let $\Phi_n = \Phi[\rho_n]$, and let $\Psi_n = \Psi[\rho_n]$ be as in Theorem* 1.1. *If, for any $f \in C^0(\mathbb{R}^d)$ such that $|f(x)| \le C(1+|x|^2)$, $\int f\rho_n \to \int \rho f$, then the sequence Φ_n can be chosen in such a way that $\Phi_n \to \Phi[\rho]$ uniformly on each compact set of Ω and strongly in $W^{1,1}(\Omega;\mathbb{R}^d)$, and $\Psi_n \to \Psi[\rho]$ uniformly on each compact set of $\mathbb{R}^d$ and strongly in $W^{1,1}_{loc}(\mathbb{R}^d)$.*

From this result, we obtain that the sequence $\nabla\Phi[\rho_n]$ converges strongly in $L^1(\Omega)$ and almost everywhere (because of the convexity of $\Phi[\rho]$) to $\nabla\Phi[\rho]$. Thus $\nabla\varphi(t,\nabla\Phi[\rho_n])$ converges to $\nabla\varphi(t,\nabla\Phi[\rho])$ in $L^1(\Omega)$ and one can pass to the limit in the formulation of Definition 2.2. This ends the proof of point 2 of Theorem 2.3.

Existence of solutions. We show briefly the existence of a solution to the Cauchy problem in the sense of Definition 2.2. Indeed given ρ^0 the initial data for the problem that we want to solve, by smoothing ρ^0, we can take a sequence ρ_n^0 of initial data belonging to $L^1(\mathbb{R}^2)$, uniformly compactly supported and converging weakly-$*$ to ρ^0. We know already from [1], [10], [20] that for every ρ_n^0, one can build a global weak solution of (3)–(5) that will be uniformly compactly supported on $[0,T]$ for all $T \ge 0$. This sequence will also be a solution in the sense of Definition 2.2. We then use the stability Theorem 2.6 and conclude that, up to extraction of a subsequence, the sequence ρ_n converges in $C([0,T],\mathcal{P}-w*)$ to a weak measure solution of SG with initial data ρ^0. This achieves the proof of Theorem 2.3. □

Remark. One can prove in fact the more general result, valid for nonlinear functionals.

PROPOSITION 2.7. *Let $F \in C^0(\Omega\times\mathbb{R}^d)$, such that $|F(x,y)| \le C(1+|y|^2)$, and let $(\rho_n)_{n\in\mathbb{N}}$ be a bounded sequence of probability measures, Lebesgue integrable, with finite second moment. Let ρ be a probability measure with finite second moment, such that for all $f \in C^0(\mathbb{R}^d)$ such that $|f(x)| \le C(1+|x|^2)$, $\int f d\rho_n \to \int f d\rho$. Then, as n goes to ∞, we have*

$$\begin{aligned}\int_{\mathbb{R}^d} F(\nabla\Psi[\rho_n](x),x)\ d\rho_n(x) &= \int_\Omega F(y,\nabla\Phi[\rho_n](y))\ dy\\ \to_n \int_\Omega F(y,\nabla\Phi[\rho](y))\ dy &=: \int_{\mathbb{R}^d} F(\bar\partial\Psi[\rho](x),x)\ d\rho(x),\end{aligned}$$

where $\bar\partial\Psi[\rho]$ is given in Definition 2.1.

Remark. One checks easily that this definition of $\int_{\mathbb{R}^d} F(\bar\partial\Psi[\rho](x),x)\ d\rho(x)$ is consistent with the definition of $\int_{\mathbb{R}^d} F(\nabla\Psi[\rho](x),x)\ d\rho(x)$ whenever ρ is absolutely continuous with respect to the Lebesgue measure, or $\nabla\Psi[\rho]$ is continuous. Indeed, note that $\nabla\Psi$ and $\bar\partial\Psi$ always coincide Lebesgue almost everywhere, since as a convex and hence Lipschitz function Ψ is differentiable Lebesgue almost everywhere (Rademacher's theorem), and hence $\partial\Psi$ is single-valued Lebesgue almost everywhere.

3. Continuous solutions. What initial regularity is necessary in order to guarantee that the velocity fields remains Lipschitz, or that the flow remains continuous,

at least for a short time? The celebrated Youdovich's theorem for the Euler incompressible equation shows that when $d = 2$, if the initial vorticity data is bounded in L^∞, the flow is Hölder continuous, with Hölder index decreasing to 0 as time goes to infinity. This proof relies on the following regularity property of the Poisson equation: If $\Delta\phi$ is bounded in L^∞, then $\nabla\phi$ is log-Lipschitz. This continuity is enough to define a Hölder continuous flow for the vector field $\nabla\phi^\perp$. Such a result is not valid for the Monge–Ampère equation. As far as we know, the strongest regularity result for Monge–Ampère equations is the following.

3.1. Regularity of solutions to Monge–Ampère equation with Dini continuous right-hand side.

THEOREM 3.1 (Wang [25]). *Let u be a strictly convex Alexandrov solution of*

$$\det D^2 u = \rho \tag{11}$$

with ρ strictly positive. If $w(r)$, the modulus of continuity of ρ, satisfies

$$\int_0^1 \frac{w(r)}{r} dr < \infty, \tag{12}$$

then u is in C^2_{loc}.

We will work here in the periodic case. In this case, u the solution of (11) will be $\Psi[\rho]$ of Theorem 1.2. The arguments of [7], [8], adapted to the periodic case, show that $\Psi[\rho]$ is indeed a strictly convex Alexandrov solution of (11). Therefore we obtain the following corollary of Theorem 3.1.

COROLLARY 3.2. *Let $\rho \in \mathcal{P}(\mathbb{T}^d)$ be such that*

$$0 < m \le \rho \le M,$$
$$\int_0^1 \frac{w(r)}{r} dr = C < \infty,$$

where m, M, C are positive constants. Let $\Psi[\rho]$ be as in Theorem 1.2. We have, for some constant $\mathcal{H}$ depending only on m, M, C,

$$\|\Psi[\rho]\|_{C^2(\mathbb{T}^d)} \le \mathcal{H}.$$

3.2. Result. We will now prove the following.

THEOREM 3.3. *Let ρ^0 be a probability on $\mathbb{T}^3$, such that ρ is strictly positive and satisfies the continuity condition* (12). *Then there exist $T > 0$ and C_1, C_2 depending on ρ^0, such that on $[0, T]$ there exists a solution $\rho(t, x)$ of SG that satisfies the following for all $t \in [0, T]$:*

$$\int_0^1 \frac{w(t, r)}{r} dr \le C_1, \qquad \|\Psi(t, \cdot)\|_{C^2(\mathbb{T}^3)} \le C_2,$$

where $w(t, r)$ is the modulus of continuity (in space) of $\rho(t, .)$.

Proof of Theorem 3.3. Let us first sketch the proof: If $\Psi \in C^2$, then the flow $t \to X(t, x)$ generated by the velocity field $[\nabla\Psi(x) - x]^\perp$ is Lipschitz in space. Since the flow is incompressible, we have $\rho(t, x) = \rho^0(X^{-1}(t, x))$.

Now we use the following property: If two functions f, g have a modulus of continuity, respectively, w_f, w_g, then $g \circ f$ has modulus $w_g \circ w_f$.

Thus if $X^{-1}(t)$ is Lipschitz, we have $w_{\rho^0 \circ X^{-1}(t)} \leq w_{\rho^0}(L\,\cdot)$ with L the Lipschitz constant of $X^{-1}(t)$, and condition (12) remains satisfied.

Remark 4. Note that Hölder continuous functions satisfy condition (12).

Remark 5. Note also that we do not need any integrability on $\nabla\rho$ and the solution of the Eulerian system (3)–(5) still has to be understood in the distributional sense.

A fixed point argument. Let us introduce the seminorm

$$\|\mu\|_{\mathcal{C}} = \int_0^1 \frac{w_\mu(r)}{r} dr \tag{13}$$

defined on $\mathcal{P}(\mathbb{T}^3)$, where we recall that w_μ is the modulus of continuity of μ. We denote by $\mathcal{P}_C$ the set $\mathcal{P}$ equipped with this seminorm, i.e.,

$$\mathcal{P}_C = \{\mu \in \mathcal{P}(\mathbb{T}^3), \|\mu\|_{\mathcal{C}} < \infty\}.$$

From now on, we fix ρ^0 a probability density in $\mathcal{P}_C$, satisfying $m \leq \rho^0 \leq M$, where m and M are strictly positive constants. Let μ be a time-dependent probability density in $L^\infty([0,T];\mathcal{P}_C)$, such that $m \leq \mu(t) \leq M$ for all t. We consider the solution ρ of the initial value problem:

$$\partial_t \rho + (\nabla\Psi[\mu](x) - x)^\perp \cdot \nabla\rho = 0, \tag{14}$$

$$\rho(t=0) = \rho^0. \tag{15}$$

From Theorem 3.1 and its corollary, the vector field $\mathbf{v}[\mu] = (\nabla\Psi[\mu](x) - x)^\perp$ is C^1 uniformly in time; therefore there exists a unique solution to this equation, by the Cauchy–Lipschitz theorem. This solution can be built by the method of characteristics as follows: Consider the flow $X(t,x)$ of the vector field $\mathbf{v}[\mu]$. Then $\rho(t)$ is ρ^0 pushed forward by $X(t)$, i.e., $\rho(t) = \rho^0 \circ X^{-1}(t)$. From the incompressibility of $\mathbf{v}[\mu]$ the condition $m \leq \rho^0 \leq M$ implies that for all $t \in [0,T]$, $m \leq \rho(t) \leq M$.

The initial data ρ^0 being fixed, the map $\mu \mapsto \rho$ will be denoted by $\mathcal{F}$.

The spatial derivative of X, $D_x X$ satisfies

$$\partial_t D_x X(t) = D_x \mathbf{v}[\mu](X) D_x X(t); \tag{16}$$

therefore we have

$$|D_x X(t)| \leq \exp\left(t \sup_{s\in[0,t]} \|D_x \mathbf{v}[\mu](s)\|_{L^\infty}\right). \tag{17}$$

We have also, from (16),

$$\partial_t \left[D_x X^{-1}(t, X(t,x))\right] = D_x X^{-1}(t, X(t,x));$$

hence

$$|D_x X^{-1}(t)| \leq \exp\left(t \sup_{s\in[0,t]} \|D_x \mathbf{v}[\mu](s)\|_{L^\infty}\right). \tag{18}$$

Since $w_{f\circ g} \leq w_f \circ w_g$, and writing $C_t = \exp(t \sup_{s\in[0,t]} \|D_x \mathbf{v}[\mu]\|_{L^\infty})$, we obtain $w_{\rho(t)}(\cdot) \leq w_{\rho^0}(C_t \cdot)$, and

$$\begin{aligned}\int_0^1 \frac{w_{\rho(t)}(r)}{r} dr &\leq \int_0^{C_t} \frac{w_{\rho^0}(r)}{r} dr \\ &\leq \int_0^1 \frac{w_{\rho^0}(r)}{r} dr + (M-m)(C_t - 1)\end{aligned}$$

(using that, for all $r, w_\rho(r) \leq M - m$). Therefore,

$$\|\rho(t)\|_{\mathcal{C}} \leq \|\rho^0\|_{\mathcal{C}} + (M-m)(C_t - 1).$$

Now from Corollary 3.2, and m, M being fixed, there exists a nondecreasing function $\mathcal{H}$ such that

$$\|\mathbf{v}[\mu]\|_{C^1} \leq \mathcal{H}(\|\mu\|_{\mathcal{C}}),$$

and so $C_t \leq \exp(t\mathcal{H}(\|\mu\|_{L^\infty([0,t];\mathcal{P}_C)}))$. Hence we can choose $Q > 1$, and then T such that

$$\|\rho^0\|_{\mathcal{C}} + (M-m)(\exp(T\ \mathcal{H}(Q\|\rho^0\|_{\mathcal{C}})) - 1) = Q\|\rho^0\|_{\mathcal{C}}.$$

Note that for $Q > 1$, we necessarily have $T > 0$. Then the map $\mathcal{F} : \mu \mapsto \rho$ goes now from

$$\mathcal{A} = \big\{\mu,\ \|\mu\|_{L^\infty([0,T];\mathcal{P}_C)} \leq Q\|\rho^0\|_{\mathcal{C}},\ m \leq \mu \leq M\big\}$$

into

$$\mathcal{B} = \big\{\rho,\ \|\rho(t)\|_{\mathcal{C}} \leq \|\rho^0\|_{\mathcal{C}} + (M-m)(\exp(t\ \mathcal{H}(Q\|\rho^0\|_{\mathcal{C}})) - 1)\ \forall t \in [0,T]\big\},$$

and with our choice of $T = T(Q)$, we have $\mathcal{B} \subset \mathcal{A}$. Moreover from the unconditional bounds

$$\rho \leq M,$$
$$\|\mathbf{v}[\mu]\|_{L^\infty([0,T]\times\mathbb{T}^3)} \leq \sqrt{3}/2$$

(see the remark after Theorem 1.2 for the second bound), and using (14), we have also $\|\partial_t\rho\|_{L^\infty([0,T];W^{-1,\infty})} \leq K(M)$ whenever $\rho = \mathcal{F}(\mu)$.

Call $\tilde{\mathcal{A}}$ (resp., $\tilde{\mathcal{B}}$) the set $\mathcal{A} \cap \{\rho, \|\partial_t\rho\|_{L^\infty([0,T];W^{-1,\infty})} \leq K(M)\}$ (resp., $\mathcal{B} \cap \{\rho, \|\partial_t\rho\|_{L^\infty([0,T];W^{-1,\infty})} \leq K(M)\}$); we claim that

- $\mathcal{F}(\tilde{\mathcal{A}}) \subset \tilde{\mathcal{B}} \subset \tilde{\mathcal{A}}$;
- $\tilde{\mathcal{A}}$ is convex and compact for the $C^0([0,T]\times\mathbb{T}^3)$ topology;
- $\mathcal{F}$ is continuous for this topology,

so that we can apply the Schauder fixed point Theorem. We check only the last point, the second being a classical result of functional analysis. So let us consider a sequence $(\mu_n)_{n\in\mathbb{N}}$ converging to $\mu \in \mathcal{A}$ and the corresponding sequence $(\rho_n = \mathcal{F}(\mu_n))_{n\in\mathbb{N}}$. The sequence ρ_n is precompact in $C^0([0,T]\times\mathbb{T}^3)$, from the previous point, and we see (with the stability Theorem 2.6) that it converges to a solution ρ of

$$\partial_t\rho + \nabla\cdot(\rho\mathbf{v}[\mu]) = 0.$$

But, $\mathbf{v}[\mu]$ being Lipschitz, this solution is unique, and therefore $\mathcal{F}(\mu_n)$ converges to $\mathcal{F}(\mu)$, which proves the continuity of $\mathcal{F}$, and ends the proof of existence by the Schauder fixed point Theorem. □

We state here some consequences of the previous result.

COROLLARY 3.4. *Let $\rho^0 \in \mathcal{P}(\mathbb{T}^3)$, such that $0 < m \leq \rho \leq M$.*

1. *If $\rho^0 \in C^\alpha, \alpha \in]0,1]$, for $T^* > 0$ depending on ρ^0, a solution $\rho(t,x)$ to (3, 4, 5) exists in $L^\infty([0,T^*[, C^\alpha(\mathbb{T}^3))$.*
2. *If $\rho^0 \in W^{1,p}, p > 3$, for $T^* > 0$ depending on ρ^0, a solution $\rho(t,x)$ to (3, 4, 5) exists in $L^\infty([0,T[, W^{1,p}(\mathbb{T}^3))$.*

3. *If* $\rho^0 \in C^{k,\alpha}, \alpha \in]0,1], k \in \mathbb{N}$, *for* $T^* > 0$ *depending on* ρ^0, *a solution* $\rho(t,x)$ *to* (3)–(5) *exists in* $L^\infty([0,T^*[, C^{k,\alpha}(\mathbb{T}^3))$.

Moreover, for these solutions, the velocity field is, respectively, in $C^{1,\alpha}(\mathbb{T}^3)$, $W^{2,p}(\mathbb{T}^3)$, *and* $C^{k+1,\alpha}(\mathbb{T}^3)$ *on* $[0,T^*[$.

Proof. We prove only the first point. We use the representation formula $\rho(t) = \rho^0(X^{-1}(t))$. Since Hölder continuous functions satisfy condition (12), we can construct a solution such that $X^{-1}(t)$ remains Lipschitz with respect to the x variable. Then, composing a Hölder continuous function with a Lipschitz function, we obtain a Hölder continuous function, which yields the result.

4. Uniqueness of solutions to SG with Hölder continuous densities.

4.1. Result. Here we prove the following theorem.

THEOREM 4.1. *Suppose that* $\rho^0 \in \mathcal{P}(\mathbb{T}^3)$ *with* $0 < m \leq \rho^0 \leq M$ *and belongs to* $C^\alpha(\mathbb{T}^3)$ *for some* $\alpha > 0$. *From Theorem* 3.3, *for some* $T > 0$ *there exists a solution* $\bar{\rho}$ *to SG in* $L^\infty([0,T], C^\alpha(\mathbb{T}^3))$. *Then every solution of SG in* $L^\infty([0,T'], C^\beta(\mathbb{T}^3))$ *for* $T' > 0, \beta > 0$ *with the same initial data coincides with* $\bar{\rho}$ *on* $[0, \inf\{T, T'\}]$.

Remark 6. The uniqueness of weak solutions is still an open question.

Remark 7. Our proof of uniqueness is thus valid in a smaller class of solutions than the one found in the previous section, the reason for this being the following: During the course of the proof, we will need to solve a Monge–Ampère equation, whose right-hand side is a function of the second derivatives of the solution of another Monge–Ampère equation. In Theorem 3.1, if u is a solution to (11) with a right-hand side satisfying (12), although $u \in C^2$, it is not clear that the second derivatives of u satisfy (12). Actually, it is even known to be wrong in the case of the Laplacian (for a precise discussion on the subject, the reader may refer to [15]). However, from Theorem 4.3 below, if $\rho \in C^\alpha$, then $u \in C^{2,\alpha}$.

What we actually need is a continuity condition on the right-hand side of (11) such that the second derivative of the solution u satisfies (12). This may be a weaker condition than Hölder continuity; however, the proof would not be affected, and therefore it is enough to give it under the present form.

Proof of Theorem 4.1. Let ρ_1 and ρ_2 be two solutions of (3)–(5) in $L^\infty([0,T], C^\beta(\mathbb{T}^3))$ that coincide at time 0. Let X_1, X_2 be the two corresponding Lagrangian solutions, (i.e., solutions of (1,2)). The velocity field being C^1, for all $t \in [0,T]$, $X_1(t,\cdot)$ and $X_2(t,\cdot)$ are both C^1 diffeomorphisms of $\mathbb{T}^d$.

We call $\mathbf{v}_1$ (resp., $\mathbf{v}_2$) the velocity field associated to X_1 (resp., X_2), $\mathbf{v}_i(t,x) = [\nabla\Psi_i(t,x) - x]^\perp, i = 1,2$. We have

$$\begin{aligned}\partial_t(X_1 - X_2) &= \mathbf{v}_1(X_1) - \mathbf{v}_2(X_2)\\ &= (\mathbf{v}_1(X_1) - \mathbf{v}_1(X_2)) + (\mathbf{v}_1(X_2) - \mathbf{v}_2(X_2)).\end{aligned}$$

We want to obtain a Gronwall-type inequality for $\|X_1 - X_2\|_{L^2}$. Since $\mathbf{v}_1$ is uniformly Lipschitz in space (from Theorem 3.3), the first bracket is estimated in the L^2 norm by $C\|X_1 - X_2\|_{L^2}$.

We now need to estimate the second term. We first have that

$$\int |\mathbf{v}_1(X_2) - \mathbf{v}_2(X_2)|^2 = \int \rho_2 |\nabla\Psi_1 - \nabla\Psi_2|^2,$$

and since ρ_2 is bounded, we need to estimate $\|\nabla\Psi_1 - \nabla\Psi_2\|_{L^2}$. This will be done in the following proposition.

PROPOSITION 4.2. *Let X_1, X_2 be mappings from $\mathbb{T}^d$ into itself, such that the densities $\rho_i = X_{i\#}dx, i = 1, 2$, are in $C^\alpha(\mathbb{T}^d)$ for some $\alpha > 0$ and satisfy $0 < m \leq \rho_i \leq M$. Let $\Psi_i, i = 1, 2$, be convex such that*

$$\det D^2\Psi_i = \rho_i$$

in the sense of Theorem 1.1*, i.e., $\Psi_i = \Psi[\rho_i]$. Then*

$$\|\nabla\Psi_1 - \nabla\Psi_2\|_{L^2} \leq C\|X_1 - X_2\|_{L^2},$$

where C depends only on α (the Hölder index of ρ_i), $\|\rho_i\|_{C^\alpha(\mathbb{T}^d)}$, m, and M.

Before giving a proof of this result, we conclude the proof of Theorem 4.1. Proposition 4.2 implies immediately that

$$\|\partial_t(X_1 - X_2)\|_{L^2} \leq C\|X_1 - X_2\|_{L^2},$$

and we conclude the proof of the theorem by a standard Gronwall lemma. ∎

4.2. Energy estimates along Wasserstein geodesics.

Proof of Proposition 4.2. In the proof of this result we will need the following result on optimal transportation of measures by a gradient of convex functions.

THEOREM 4.3 (Brenier [3], McCann [23], Cordero-Erausquin [9], and Caffarelli [6]). *Let ρ_1, ρ_2 be two probability measures on $\mathbb{T}^d$, such that ρ_1 is absolutely continuous with respect to the Lebesgue measure.*

1. *There exists a convex function ϕ such that $\phi - |\cdot|^2/2$ is $\mathbb{Z}^d$-periodic, satisfying $\nabla\phi_\#\rho_1 = \rho_2$.*
2. *The map $\nabla\phi$ is the $d\rho_1$ a.e. unique solution of the minimization problem*

$$\inf_{T_\#\rho_1=\rho_2} \int_{\mathbb{T}^d} \rho_1(x)|T(x) - x|^2_{\mathbb{T}^d}\, dx, \tag{19}$$

   and for all $x \in \mathbb{R}^d$, $|\nabla\phi(x) - x|_{\mathbb{T}^d} = |\nabla\phi(x) - x|_{\mathbb{R}^d}$.
3. *If ρ_1, ρ_2 are strictly positive and belong to $C^\alpha(\mathbb{T}^d)$ for some $\alpha > 0$, then $\phi \in C^{2,\alpha}(\mathbb{T}^d)$ and satisfies pointwise*

$$\rho_2(\nabla\phi)\det D^2\phi = \rho_1.$$

For complete references on the optimal transportation problem (19) and its applications, the reader can refer to [24].

Remark 8. The expression $|\cdot|_{\mathbb{T}^d}$ denotes the Riemannian distance on the flat torus, whereas $|\cdot|_{\mathbb{R}^d}$ is the Euclidean distance on $\mathbb{R}^d$. The second assertion of point 2 means that, for all $x \in \mathbb{R}^d$, $|\nabla\phi(x) - x| \leq \mathrm{diam}(\mathbb{T}^d) = \sqrt{d}/2$.

Remark 9. Here again, note that since $\phi - |\cdot|^2/2$ is periodic, the map $x \mapsto \nabla\phi(x)$ is compatible with the equivalence classes of $\mathbb{R}^d/\mathbb{Z}^d$, and therefore is defined without ambiguity on $\mathbb{T}^d$.

Wasserstein geodesics between probability measures. In this part we use results from [2], [22]. Using Theorem 4.3, we consider the unique (up to a constant) convex potential ϕ such that

$$\nabla\phi_\#\rho_1 = \rho_2,$$
$$\phi - |\cdot|^2/2 \text{ is } \mathbb{Z}^d\text{-periodic.}$$

We consider, for $\theta \in [1,2]$, ϕ_θ defined by

$$\phi_\theta = (2-\theta)\frac{|x|^2}{2} + (\theta-1)\phi.$$

We also consider, for $\theta \in [1,2]$, ρ_θ defined by

$$\rho_\theta = \nabla\phi_\theta \,{}_\#\rho_1.$$

Then ρ_θ interpolates between ρ_1 and ρ_2. This interpolation was introduced in [2] and [22] as the time continuous formulation of the Monge–Kantorovich mass transfer. In this construction, a velocity field v_θ is defined $d\rho_\theta$ a.e. as follows:

$$\begin{aligned}\forall f \in C^0(\mathbb{T}^d;\mathbb{R}^d),\ \int \rho_\theta v_\theta \cdot f &= \int \rho_1 f(\nabla\phi_\theta)\cdot\partial_\theta\nabla\phi_\theta \\ &= \int \rho_1 f(\nabla\phi_\theta)\cdot(\nabla\phi(x)-x). \end{aligned} \tag{20}$$

It is easily checked that the pair ρ_θ, v_θ satisfies

$$\partial_\theta\rho_\theta + \nabla\cdot(\rho_\theta v_\theta) = 0,$$

and for any $\theta \in [1,2]$, we have (see [2])

$$\frac{1}{2}\int_{\mathbb{T}^d}\rho_\theta|v_\theta|^2 = \frac{1}{2}\int_{\mathbb{T}^d}\rho_1|\nabla\phi(x)-x|^2 = W_2^2(\rho_1,\rho_2),$$

where $W_2(\rho_1,\rho_2)$ is the Wasserstein distance between ρ_1 and ρ_2, defined by

$$W_2^2(\rho_1,\rho_2) = \inf_{T_\#\rho_1=\rho_2}\left\{\int\rho_1(x)|T(x)-x|^2_{\mathbb{T}^d}\right\}.$$

The Wasserstein distance can also be formulated as follows:

$$W_2^2(\rho_1,\rho_2) = \inf_{Y_1,Y_2}\left\{\int_{\mathbb{T}^d}|Y_1-Y_2|^2_{\mathbb{T}^d}\right\},$$

where the infimum is performed over all maps $Y_1, Y_2 : \mathbb{T}^d \mapsto \mathbb{T}^d$ such that $Y_{i\#}dx = \rho_i, i = 1,2$. From this definition we have easily

$$W_2^2(\rho_1,\rho_2) \le \int |X_2(t,a)-X_1(t,a)|^2\,da,$$

and it follows that, for every $\theta \in [1,2]$,

$$\int_{\mathbb{T}^d}\rho_\theta|v_\theta|^2 = W_2^2(\rho_1,\rho_2) \le \|X_2-X_1\|_{L^2}. \tag{21}$$

Regularity of the interpolant measure $\boldsymbol{\rho_\theta}$. From Theorem 4.3, for $\rho_1,\rho_2 \in C^\beta$ and pinched between the positive constants m and M, we know that $\phi \in C^{2,\beta}$ and satisfies

$$\det D^2\phi = \frac{\rho_1}{\rho_2(\nabla\phi)}.$$

We now estimate $\rho_\theta = \rho_1[\det D^2\phi_\theta]^{-1}$. From the concavity of $\log(\det(\cdot))$ on symmetric positive matrices, we have

$$\begin{aligned}\det D^2\phi_\theta &= \det((2-\theta)I + (\theta-1)D^2\phi)\\ &\geq [\det D^2\phi]^{\theta-1}\\ &\geq \frac{m}{M}.\end{aligned}$$

Moreover, since $\phi \in C^2$, $\det D^2\phi_\theta$ is bounded by above. Thus ρ_θ is uniformly bounded away from 0 and infinity, and is uniformly Hölder continuous.

Final energy estimate. If we consider, for every $\theta \in [1,2]$, Ψ_θ a solution of

$$\det D^2\Psi_\theta = \rho_\theta, \tag{22}$$

in the sense of Theorem 1.2, and we impose that

$$\int_{\mathbb{T}^d} \Phi_\theta = 0 \tag{23}$$

(see [19]), then Ψ_θ interpolates between Ψ_1 and Ψ_2, and $\Psi_\theta \in C^{2,\beta}$ uniformly, from the regularity of ρ_θ. We will estimate $\partial_\theta \nabla\Psi_\theta$ by differentiating (22) with respect to θ. The fact that Ψ_θ, Φ_θ is differentiable with respect to θ is a consequence of the results of [19]. We will have, following the a priori estimate of [19, Proposition 5.1, Theorem 2.3],

$$\begin{aligned}\partial_\theta\nabla\Phi_\theta, \partial_\theta\nabla\Psi_\theta \in L^\infty([1,2], L^2(\mathbb{T}^d)),\\ \partial_\theta\Phi_\theta, \partial_\theta\Psi_\theta \in L^\infty([1,2], C^\gamma(\mathbb{T}^d))\end{aligned}$$

for some $\gamma \in]0,1[$. (Note that we need the condition (23).)

Let us obtain a precise quantitative estimate in our present case. First we recall the following fact: For M, N two $d \times d$ matrices, $t \in \mathbb{R}$,

$$\det(M + tN) = \det M + t\ (\text{trace } M^t_{co}N) + o(t),$$

where M_{co} is the comatrix (or matrix of cofactors) of M. Moreover, for any $f \in C^2(\mathbb{R}^d;\mathbb{R})$, if M is the comatrix of $D^2 f$, it is a common fact that

$$\forall j \in \{1,\dots,d\},\ \sum_{i=1}^d \partial_i M_{ij} \equiv 0. \tag{24}$$

Hence, denoting by M_θ the comatrix of $D^2\Psi_\theta$, we obtain that $\partial_\theta\Psi_\theta$ satisfies

$$\begin{aligned}\nabla\cdot(M_\theta\nabla\partial_\theta\Psi_\theta) &= \partial_\theta\rho_\theta(t)\\ &= -\nabla\cdot(\rho_\theta v_\theta),\end{aligned} \tag{25}$$

where v_θ is given by (20). From the $C^{2,\beta}$ regularity of Ψ_θ, $D^2\Psi_\theta$ is a C^β smooth, positive definite matrix, and its comatrix M_θ is as well. Thus (25) is uniformly elliptic. If we multiply by $\partial_\theta\Psi_\theta$ and integrate by parts, we obtain

$$\int \nabla^t\partial_\theta\Psi_\theta\, M_\theta\, \nabla\partial_\theta\Psi_\theta = -\int \nabla\partial_\theta\Psi_\theta\cdot v_\theta\rho_\theta.$$

Using that $M_\theta \geq \lambda I$ for some $\lambda > 0$, and combining with the inequality (21) above, we obtain

$$\begin{aligned}\|\nabla \partial_\theta \Psi_\theta(t)\|_{L^2} &\leq \lambda^{-1} \|\rho_\theta v_\theta\|_{L^2} \\ &\leq \lambda^{-1} \|X_2 - X_1\|_{L^2} \left(\sup_\theta \|\rho_\theta\|_{L^\infty} \right)^{1/2}.\end{aligned}$$

The constant λ^{-1} depends only on $m, M, \beta, \{\|\rho_i\|_{C^\beta}, i = 1, 2\}$ and is thus bounded under our present assumptions. We have already seen that ρ_θ is uniformly bounded, and we finally obtain that

$$\|\nabla \Psi_1 - \nabla \Psi_2\|_{L^2} \leq C \|X_1 - X_2\|_{L^2}. \tag{26}$$

This ends the proof of Proposition 4.2. □

Remark 10. In [19], the author obtains also (weaker) estimates of the type of Proposition 4.2 for discontinuous densities ρ_1, ρ_2.

5. Uniqueness of solutions to the 2-d Euler equations with bounded vorticity: A new proof. As stated in the introduction, the method that we have presented here to show uniqueness of solutions to SG is in fact quite general. It has been shown in [18] to yield a uniqueness result for solutions to the Vlasov–Poisson system, the only condition being that the density in the physical space is bounded. In that paper it was also shown that the method could give a new proof of Youdovich's theorem for solutions in the whole space $\mathbb{R}^2$.

We give here a simplified version of this proof in the periodic case.

We start now from the following system:

$$\partial_t \rho + \nabla \cdot (\rho \nabla \psi^\perp) = 0, \tag{27}$$

$$\rho = \Delta \psi, \tag{28}$$

$$\rho(t = 0) = \rho^0. \tag{29}$$

We restrict ourselves to the periodic case, i.e., $x \in \mathbb{T}^2$, ρ, ψ periodic; this implies that ρ has total mass equal to 0. We reprove the following classical result.

THEOREM 5.1 (Youdovich [26]). *Given an initial data $\rho^0 \in L^\infty(\mathbb{T}^2)$ satisfying $\int_{\mathbb{T}^2} \rho^0 = 0$, there exists a unique solution to* (27)–(29) *such that ρ belongs to $L^\infty_{loc}(\mathbb{R}^+ \times \mathbb{T}^2)$.*

Proof of Theorem 5.1. We consider two solutions ρ_1, ψ_1 and ρ_2, ψ_2, such that $\rho_i, i = 1, 2$, are bounded in $L^\infty([0, T] \times \mathbb{T}^d)$. In this case the velocity fields $\mathbf{v}_i = \nabla \psi_i^\perp$ both satisfy (see [21, Chapter 8])

$$\forall (x, y) \in \mathbb{T}^2, \ |x - y| \leq \frac{1}{2}, \ |\mathbf{v}_i(x) - \mathbf{v}_i(y)| \leq C |x - y| \log \frac{1}{|x - y|}.$$

The flows $(t, x) \mapsto X_i(t, x)$ associated to the velocity fields $\mathbf{v}_i = \nabla \psi_i^\perp$ are then Hölder continuous, and one has, for all $t \in [0, T]$, $\rho_i(t) = X_i(t)_\# \rho^0$.

Applying the same technique as before, we need to estimate $\|\nabla \psi_1 - \nabla \psi_2\|_{L^2(\mathbb{T}^2)}$ in terms of $\|X_1 - X_2\|_{L^2(\mathbb{T}^2)}$. In the present case, the energy estimate of Proposition 4.2 will hold under the weaker assumptions that the two densities are bounded.

PROPOSITION 5.2. *Let X_1, X_2 be continuous injective mappings from $\mathbb{T}^d$ into itself, and let ρ^0 be a bounded measure, with $\int_{\mathbb{T}^d} \rho^0 = 0$. Let $\rho_i = X_{i\#} \rho^0, i = 1, 2$.*

Assume that ρ_1, ρ_2 have densities in L^∞ with respect to the Lebesgue measure. Let $\psi_i, i = 1, 2$, be periodic solutions of $\Delta\psi_i = \rho_i, i = 1, 2$. Then we have

$$\|\nabla\psi_1 - \nabla\psi_2\|_{L^2(\mathbb{T}^d)} \leq (2\max\{\|\rho_1\|_{L^\infty}, \|\rho_2\|_{L^\infty}\}\|\rho_0\|_{L^\infty})^{1/2} \|X_1 - X_2\|_{L^2(\mathbb{T}^d)}.$$

Remark 11. In other words, this proposition shows that for ρ_1, ρ_2 bounded, the H^{-1} norm of $\rho_1 - \rho_2$ is controlled by some "generalized" (since here we have unsigned measures) Wasserstein distance between ρ_1 and ρ_2.

Remark 12. We see that we obtain a result as in Proposition 4.2 under the weaker condition that the densities are bounded in L^∞ (and not in C^α). This is because the Laplacian is uniformly elliptic, independently of the regularity of the solution, while the Monge–Ampère operator is uniformly elliptic only for C^2 solutions.

To conclude the proof of Theorem 5.1, note first that for all $C > 0$, we can take T small enough so that $\|X_2 - X_1\|_{L^\infty([0,T]\times\mathbb{T}^2)} \leq C$. Now we have, for the difference $X_1 - X_2$, as long as $|X_1 - X_2| \leq 1/2$,

$$\begin{aligned}\|\partial_t(X_1 - X_2)\|_{L^2} &\leq \|\nabla\psi_1(X_1) - \nabla\psi_1(X_2)\|_{L^2} + \|\nabla\psi_1(X_2) - \nabla\psi_2(X_2)\|_{L^2} \\ &\leq C_1\||X_1 - X_2|\log(|X_1 - X_2|)\|_{L^2} + C_2\|X_1 - X_2\|_{L^2},\end{aligned}$$

where we have used Proposition 5.2 to evaluate the second term. We just need to evaluate $\||X_1 - X_2|\log(|X_1 - X_2|)\|_{L^2}$. We take T small enough so that $\|X_2 - X_1\|_{L^\infty([0,T]\times\mathbb{T}^2)} \leq 1/e$ and notice that $x \mapsto x\log^2 x$ is concave for $0 \leq x \leq 1/e$. Therefore by Jensen's inequality we have

$$\begin{aligned}\int_{\mathbb{T}^2} |X_2 - X_2|^2 \log^2(|X_1 - X_2|) &= \frac{1}{4}\int_{\mathbb{T}^2} |X_2 - X_2|^2 \log^2(|X_1 - X_2|^2) \\ &\leq \frac{1}{4}\int_{\mathbb{T}^2} |X_2 - X_1|^2 \log^2\left(\int_{\mathbb{T}^2} |X_2 - X_1|^2\right),\end{aligned}$$

and some elementary computations finally yield

$$\partial_t\|X_2 - X_1\|_{L^2} \leq C\|X_2 - X_1\|_{L^2}\log\frac{1}{\|X_2 - X_1\|_{L^2}}.$$

The conclusion $X_1 \equiv X_2$ then follows by standard arguments.

5.1. Energy estimates along Wasserstein geodesics.

Proof of Proposition 5.2. The proof of this proposition is very close to the proof of Proposition 4.2, and we will only sketch it, insisting on the specific points. Here the densities ρ_i cannot be of constant sign, since their mean value is zero; hence we introduce ρ_i^+ (resp., ρ_i^-) the positive (resp., negative) part of ρ_i, i.e., $\rho_i = \rho_i^+ - \rho_i^-$. The mappings X_i are supposed injective, and therefore we have $X_{i\#}\rho_0^\pm = \rho_i^\pm$. Now, $\rho_i^\pm$ are positive measures of total mass equal to, say, M, with $M < \infty$.

Wasserstein geodesic. We interpolate between the positive parts ρ_i^+, and the negative part is handled in the same way. As before, we introduce the density $\rho_\theta^+(t)$ that interpolates between $\rho_1^+(t)$ and $\rho_2^+(t)$. In this interpolation, we consider v_θ^+ such that

$$\partial_\theta\rho_\theta^+ + \nabla\cdot(\rho_\theta^+ v_\theta^+) = 0, \tag{30}$$

and we introduce $\rho_\theta^-, v_\theta^-$ as well. Then $\rho_\theta = \rho_\theta^+ - \rho_\theta^-$ has mean value zero. Let the potential ψ_θ be a solution to

$$\Delta\psi_\theta = \rho_\theta. \tag{31}$$

Note that ρ_θ has mean value zero, and therefore this equation is well posed on $\mathbb{T}^2$; moreover ψ_θ interpolates between ψ_1 and ψ_2.

Bound on the interpolant measure ρ_θ. Instead of interpolating between two smooth densities, we interpolate between bounded densities and use the following result from [22].

PROPOSITION 5.3 (McCann [22]). *Let ρ_θ^+ be the Wasserstein geodesic linking ρ_1^+ to ρ_2^+ defined above. Then, for all $\theta \in [1,2]$,*

$$\|\rho_\theta^+\|_{L^\infty} \le \max\left\{\|\rho_1^+\|_{L^\infty}, \|\rho_2^+\|_{L^\infty}\right\}.$$

The same holds for ρ_i^-, ρ_θ^-.

Remark. This property is often referred to as *displacement convexity*.

Energy estimates. Now we impose that $\int_{\mathbb{T}^d} \phi_\theta = 0$. Since $\rho_\theta^\pm, v_\theta^\pm$ are uniformly bounded in L^∞, we have, using (30), that $\partial_\theta \rho_\theta \in L^\infty([1,2]; W^{-1,\infty}(\mathbb{T}^d))$. We can thus differentiate (31) with respect to θ to obtain

$$\Delta \partial_\theta \psi_\theta = \partial_\theta \rho_\theta = -\nabla \cdot (\rho_\theta^+ v_\theta^+ - \rho_\theta^- v_\theta^-), \tag{32}$$

with $v_\theta^\pm$ the interpolating velocity defined as in (20) and satisfying, for all $\theta \in [1,2]$,

$$\int \rho_\theta^\pm(t) |v_\theta^\pm|^2(t) = W_2^2(\rho_1^\pm(t), \rho_2^\pm(t)).$$

Multiplying (32) by $\partial_\theta \psi_\theta$ and integrating over $\theta \in [1,2]$, we obtain

$$\begin{aligned}\|\nabla\psi_1 - \nabla\psi_2\|_{L^2(\mathbb{T}^d)} &\le \int_{\theta=1}^2 \|\rho_\theta^+ v_\theta^+\|_{L^2} + \|\rho_\theta^- v_\theta^-\|_{L^2} \\ &\le W_2(\rho_1^+, \rho_2^+)\Big(\sup_\theta \|\rho_\theta^+\|_{L^\infty}\Big)^{1/2} \\ &\quad + W_2(\rho_1^-, \rho_2^-)\Big(\sup_\theta \|\rho_\theta^-\|_{L^\infty}\Big)^{1/2}.\end{aligned}$$

Note that the energy estimate is easier here than in the Monge–Ampère case, since the problem is immediately uniformly elliptic.

The mappings X_i are injective and satisfy $X_{i\#}\rho_0 = \rho_i$; therefore we have $X_{i\#}(\rho_0^\pm) = \rho_i^\pm$. Hence,

$$\begin{aligned}W_2^2(\rho_1^\pm, \rho_2^\pm) &\le \int \rho_0^\pm |X_1 - X_2|^2 \\ &\le \|\rho_0\|_{L^\infty} \|X_1 - X_2\|_{L^2}.\end{aligned}$$

Using Proposition 5.3, we conclude:

$$\|\nabla\psi_1 - \nabla\psi_2\|_{L^2(\mathbb{T}^d)} \le 2\|\rho_0\|_{L^\infty}^{1/2} \|X_2 - X_1\|_{L^2} \left(\max\left\{\|\rho_1\|_{L^\infty}, \|\rho_2\|_{L^\infty}\right\}\right)^{1/2}.$$

This ends the proof of Proposition 5.2. Note that in our specific case, X_i are Lebesgue measure-preserving invertible mappings; therefore $\|\rho_i^\pm\|_{L^\infty} = \|\rho_0^\pm\|_{L^\infty}$, and the estimate can be simplified in

$$\|\nabla\psi_1 - \nabla\psi_2\|_{L^2(\mathbb{T}^d)} \le 2\|\rho_0\|_{L^\infty} \|X_2 - X_1\|_{L^2(\mathbb{T}^d)}. \qquad \square$$

6. Convergence to the Euler equation.

6.1. Scaling of the system. Here we present a rescaled version of the 2-d SG system and some formal arguments to motivate the next convergence results. Here $x \in \mathbb{T}^2, t \in \mathbb{R}^+$, and for $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$, $\mathbf{v}^\perp$ now means $(-v_2, v_1)$. Introducing $\psi[\rho] = \Psi[\rho] - |x|^2/2$, where $\Psi[\rho]$ is given by Theorem 1.2, the periodic 2-d SG system now reads

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho \nabla \psi^\perp) &= 0, \\ \det(I + D^2\psi) &= \rho. \end{aligned}$$

If ρ is close to one, then ψ should be small, and therefore one may consider the linearization $\det(I + D^2\psi) = 1 + \Delta\psi + O(|D^2\psi|^2)$, which yields $\Delta\psi \simeq \rho - 1$. Thus for small initial data, i.e., $\rho^0 - 1$ small, one expects $\psi, \mu = \rho - 1$ to stay close to a solution of the Euler incompressible equation EI,

$$\partial_t \bar{\rho} + \nabla \cdot (\bar{\rho} \nabla \bar{\phi}^\perp) = 0, \tag{33}$$

$$\Delta \bar{\phi} = \bar{\rho}. \tag{34}$$

We shall rescale the equation in order to consider quantities of order one. We introduce the new unknown

$$\begin{aligned} \rho^\epsilon(t,x) &= \frac{1}{\epsilon}\left(\rho\left(\frac{t}{\epsilon}, x\right) - 1\right), \\ \psi^\epsilon(t,x) &= \frac{1}{\epsilon}\psi\left(\frac{t}{\epsilon}, x\right). \end{aligned}$$

Then we have

$$\begin{aligned} \rho(t) &= 1 + \epsilon\rho^\epsilon(\epsilon t), \\ \Psi[\rho](t) &= |x|^2/2 + \epsilon\psi^\epsilon(\epsilon t), \end{aligned}$$

and we define ϕ^ϵ by

$$\epsilon\phi^\epsilon = |x|^2/2 - \Phi[\rho],$$

so that

$$\nabla\phi^\epsilon = \nabla\psi^\epsilon(\nabla\Phi[\rho]). \tag{35}$$

Hence, at a point $x \in \mathbb{T}^2$, $\nabla\phi^{\epsilon\perp}$ is the velocity of the associated dual point $\nabla\Phi[\rho](x)$. The evolution of these quantities is then governed by the system SG_ϵ,

$$\partial_t \rho^\epsilon + \nabla \cdot (\rho^\epsilon \nabla \psi^{\epsilon\perp}) = 0, \tag{36}$$

$$\det(I + \epsilon D^2\psi^\epsilon) = 1 + \epsilon\rho^\epsilon. \tag{37}$$

Remark. Note that this system admits global weak solutions with initial data any bounded measure $\rho^{\epsilon\,0}$, as long as

$$\int_{\mathbb{T}^2} \rho^{\epsilon\,0} = 0, \tag{38}$$

$$\rho^{\epsilon\,0} \geq -\frac{1}{\epsilon}. \tag{39}$$

Note also that if the pair $(\bar\rho, \bar\phi)$ is a solution to the EI system (33)–(34), so is the pair $(\frac{1}{\epsilon}\bar\rho(\frac{t}{\epsilon}, x),\ \frac{1}{\epsilon}\bar\phi(\frac{t}{\epsilon}, x))$.

We now present the convergence results. We show that solutions of SG_ϵ converge to solutions of EI in the following sense: If $\rho^{\epsilon\,0}$, the initial data of SG_ϵ, is close (in some sense depending on the type of convergence we wish to show) to a smooth initial data $\bar\rho^0$ for EI, then ρ^ϵ and $\bar\rho$ remain close for some time. This time goes to infinity when ϵ goes to 0.

We present two different versions of this result: The first one is for weak solutions of SG_ϵ, and the second one is for Lipschitz solutions.

6.2. Convergence of weak solutions.

THEOREM 6.1. *Let $(\rho^\epsilon, \psi^\epsilon)$ be a weak solution of the* SG_ϵ *system* (36)–(37). *Let $(\bar\rho, \bar\phi)$ be a smooth $C^3([0,T]\times\mathbb{T}^2)$ solution of the* EI *system* (33)–(34). *Let ϕ^ϵ be obtained from ψ^ϵ as in* (35), *and let $H_\epsilon(t)$ be defined by*

$$H_\epsilon(t) = \frac{1}{2}\int_{\mathbb{T}^2} \left|\nabla\phi^\epsilon - \nabla\bar\phi\right|^2.$$

Then

$$H_\epsilon(t) \le (H_\epsilon(0) + C\epsilon^{2/3}(1+t))\exp Ct,$$

where C depends only on $\sup_{0\le s\le t}\{\|D^3\bar\phi(s), D^2\partial_t\bar\phi(s)\ \|_{L^\infty(\mathbb{T}^2)}\}$.

In particular, if $H_\epsilon(0) \le C_0\epsilon^{2/3}$, we have, for all $T>0, t\in[0,T]$,

$$H_\epsilon(t) \le C_T\epsilon^{2/3},$$

where C_T depends on T, C, C_0 above.

Remark 13. Note that $\nabla\phi^{\epsilon\perp}(t,x)$ is the velocity at point $\nabla\Phi[\rho] = x - \epsilon\nabla\phi^\epsilon$. Thus we compare the SG_ϵ velocity at point $x - \epsilon\nabla\phi^\epsilon$ (the dual point of x) with the EI velocity at point x. Our result also allows comparison of the velocities at the same point by noticing that

$$\begin{aligned} G_\epsilon(t) &= \frac{1}{2}\int_{\mathbb{T}^2}\rho\left|\nabla\psi^\epsilon - \nabla\bar\phi\right|^2 \\ &= \frac{1}{2}\int_{\mathbb{T}^2}\left|\nabla\phi^\epsilon - \nabla\bar\phi(x-\epsilon\nabla\phi^\epsilon)\right|^2 \\ &\le C(H_\epsilon(t) + \epsilon^2), \end{aligned}$$

using the smoothness of $\bar\phi$, and if $\mathbf{v}_{sg_\epsilon}, \mathbf{v}_{ei}$ are the respective velocities of the SG_ϵ and EI systems, $G_\epsilon = \int_{\mathbb{T}^2}\rho^\epsilon|\mathbf{v}_{sg_\epsilon} - \mathbf{v}_{ei}|^2$.

Remark 14. The expansion $\det(I + D^2\psi) = 1 + \Delta\psi + O(|D^2\psi|^2)$, used in the heuristic argument above to justify the convergence, relies a priori on the control of $D^2\psi$ in the sup norm. But in Theorem 6.1, the initial data must satisfy $\nabla\psi^\epsilon$ close in the L^2 norm to $\nabla\bar\phi$; this condition means that $D^2\psi^\epsilon$ is close in the H^{-1} norm to $D^2\bar\phi$, which is smooth. This control does not allow us to justify the expansion $\det(I + D^2\psi) = 1 + \Delta\psi + O(|D^2\psi|^2)$, but we see that the result remains valid.

Proof of Theorem 6.1. In all the proof, we use C to denote any quantity that depends only on $\bar\phi$. We use the conservation of the energy of the SG_ϵ system, given by

$$E(t) = \int_{\mathbb{T}^2}|\nabla\phi^\epsilon|^2. \tag{40}$$

This fact, although formally easily justified, is actually not so straightforward for weak solutions, and has been proved by F. Otto in an unpublished work. The argument is explained in [5]. Therefore $E(t) = E_0$. The energy of the smooth solution of EI is given by

$$\bar{E}(t) = \int_{\mathbb{T}^2} |\nabla \bar{\phi}|^2 \tag{41}$$

and also conserved. For all smooth θ, we will use the notation

$$\langle D^2\theta\rangle(t,x) = \int_{s=0}^{1} (1-s) D^2\theta(t, x - s\epsilon\nabla\phi^\epsilon(t,x)).$$

Thus we have the identity

$$\int_{\mathbb{T}^2} \rho^\epsilon\theta = \int_{\mathbb{T}^2} \theta(x - \epsilon\nabla\phi^\epsilon) \tag{42}$$

$$= \int_{\mathbb{T}^2} \theta - \epsilon \int_{\mathbb{T}^2} \nabla\theta\cdot\nabla\phi^\epsilon + \epsilon^2 \int_{\mathbb{T}^2} \langle D^2\theta\rangle \nabla\phi^\epsilon\nabla\phi^\epsilon. \tag{43}$$

Using the energy bound, the last term is bounded by $\epsilon^2\|D^2\theta\|_{L^\infty(\mathbb{T}^2)}E_0$. Then, using the conservation of the energies E and $\bar{E}$ defined, respectively, in (40) and (41), we have

$$\frac{d}{dt}H_\epsilon(t) = -\frac{d}{dt}\int_{\mathbb{T}^2} \nabla\bar{\phi}\cdot\nabla\phi^\epsilon.$$

Using the identity (43), we have, for all smooth θ,

$$\epsilon\int_{\mathbb{T}^2} \nabla\theta\cdot\nabla\phi^\epsilon = -\int_{\mathbb{T}^2}\rho^\epsilon\theta + \int_{\mathbb{T}^2}\theta + \epsilon^2\int_{\mathbb{T}^2}\langle D^2\theta\rangle\nabla\phi^\epsilon\nabla\phi^\epsilon;$$

hence, replacing θ by $\bar{\phi}$ in this identity, we get

$$\frac{d}{dt}H_\epsilon(t) = \frac{1}{\epsilon}\frac{d}{dt}\int_{\mathbb{T}^2}[\rho^\epsilon\bar{\phi} - \bar{\phi} - \epsilon^2\langle D^2\bar{\phi}\rangle\nabla\phi^\epsilon\nabla\phi^\epsilon].$$

We can suppose without loss of generality that $\int_{\mathbb{T}^2}\bar{\phi}(t,x)\,dx \equiv 0$. Then if we define

$$Q_\epsilon(t) = \int_{\mathbb{T}^2} \epsilon\langle D^2\bar{\phi}\rangle\nabla\phi^\epsilon\nabla\phi^\epsilon$$

(note that $|Q_\epsilon(t)| \le C\epsilon$), we have

$$\frac{d}{dt}(H_\epsilon + Q_\epsilon) = \frac{1}{\epsilon}\frac{d}{dt}\int_{\mathbb{T}^2}\rho^\epsilon\bar{\phi}.$$

Hence we are left to compute

$$\begin{aligned}
\frac{1}{\epsilon}\frac{d}{dt}\int_{\mathbb{T}^2}\rho^\epsilon\bar{\phi} &= \frac{1}{\epsilon}\int_{\mathbb{T}^2}\partial_t\rho^\epsilon\bar{\phi} + \rho^\epsilon\partial_t\bar{\phi}\\
&= \frac{1}{\epsilon}\int_{\mathbb{T}^2}\rho^\epsilon\nabla\psi^{\epsilon\perp}\cdot\nabla\bar{\phi} - \epsilon\nabla\phi^\epsilon\cdot\nabla\partial_t\bar{\phi} + \epsilon^2\langle D^2\partial_t\bar{\phi}\rangle\nabla\bar{\phi}\nabla\bar{\phi}\\
&= \frac{1}{\epsilon}\int_{\mathbb{T}^2}\rho^\epsilon\nabla\psi^{\epsilon\perp}\cdot\nabla\bar{\phi} - \int_{\mathbb{T}^2}\nabla\phi^\epsilon\cdot\nabla\partial_t\bar{\phi} + O(\epsilon)\\
&= T_1 + T_2 + O(\epsilon),
\end{aligned}$$

where in the second line we have used (36) for the first term and (43) with $\theta = \partial_t \bar{\phi}$ for the second and third terms. (Remember also that we assume $\int \partial_t \bar{\phi} \equiv 0$.)

We will now use the other formulation of the Euler equation: $\mathbf{v} = \nabla \bar{\phi}^\perp$ satisfies

$$\partial_t \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} = -\nabla p.$$

After a rotation of $\pi/2$, this equation becomes

$$\partial_t \nabla \bar{\phi} + D^2 \bar{\phi} \nabla \bar{\phi}^\perp = \nabla p^\perp,$$

and thus for T_2 we have

$$\begin{aligned} T_2 &= -\int_{\mathbb{T}^2} \nabla \phi^\epsilon \cdot \nabla \partial_t \bar{\phi} \\ &= \int_{\mathbb{T}^2} \nabla \phi^\epsilon D^2 \bar{\phi} \nabla \bar{\phi}^\perp. \end{aligned}$$

For T_1, using (35) and (43), we have

$$\begin{aligned} \epsilon T_1 &= \int_{\mathbb{T}^2} \rho^\epsilon \nabla \psi^{\epsilon\perp} \cdot \nabla \bar{\phi} \\ &= \int_{\mathbb{T}^2} \nabla \psi^{\epsilon\perp}(x - \epsilon \nabla \phi^\epsilon) \cdot \nabla \bar{\phi}(x - \epsilon \nabla \phi^\epsilon) \\ &= \int_{\mathbb{T}^2} \nabla \phi^{\epsilon\perp} \cdot \nabla \bar{\phi} - \epsilon \nabla \phi^{\epsilon\perp} D^2 \bar{\phi} \nabla \phi^\epsilon + \epsilon \Xi, \end{aligned}$$

where Ξ is defined by

$$\Xi = \int_{\mathbb{T}^2} \nabla \phi^{\epsilon\perp} \left(D^2 \bar{\phi} - \int_{s=0}^1 D^2 \bar{\phi}(x - s\epsilon \nabla \phi^\epsilon)\, ds \right) \nabla \phi^\epsilon. \tag{44}$$

The term $\int_{\mathbb{T}^2} \nabla \phi^{\epsilon\perp} \cdot \nabla \bar{\phi}$ vanishes identically. Concerning Ξ, we claim the following estimate.

LEMMA 6.2. *Let Ξ be defined by* (44). *Then*

$$|\Xi| \leq C(\epsilon^{\frac{2}{3}} + H_\epsilon),$$

where C depends on $\|D^3 \bar{\phi}\|_{L^\infty}$.

We postpone the proof of this lemma until after the proof of Theorem 6.1. We now obtain

$$\frac{d}{dt}(H_\epsilon(t) + Q_\epsilon(t)) \leq \int_{\mathbb{T}^2} (\nabla \bar{\phi}^\perp - \nabla \phi^{\epsilon\perp}) D^2 \bar{\phi} \nabla \phi^\epsilon + C H_\epsilon + C \epsilon^{2/3}.$$

Noticing that for every $\theta : \mathbb{T}^2 \mapsto \mathbb{R}$ we have

$$\int_{\mathbb{T}^2} \nabla \theta^\perp D^2 \bar{\phi} \nabla \bar{\phi} = \int_{\mathbb{T}^2} \nabla \theta^\perp \cdot \nabla \left(\frac{1}{2} |\nabla \bar{\phi}|^2 \right) = 0,$$

we find that

$$\int_{\mathbb{T}^2} (\nabla \bar{\phi}^\perp - \nabla \phi^{\epsilon\perp}) D^2 \bar{\phi} \nabla \phi^\epsilon = \int_{\mathbb{T}^2} (\nabla \phi^\perp - \nabla \bar{\phi}^{\epsilon\perp}) D^2 \bar{\phi} (\nabla \phi^\epsilon - \nabla \bar{\phi}),$$

and hence

$$\frac{d}{dt}(H_\epsilon(t)+Q_\epsilon(t)) \le -\int_{\mathbb{T}^2}(\nabla\phi^{\epsilon\perp}-\nabla\bar\phi^\perp)D^2\bar\phi(\nabla\phi^\epsilon-\nabla\bar\phi)+CH_\epsilon+C\epsilon^{2/3}$$
$$\le C(H_\epsilon(t)+Q_\epsilon(t)+\epsilon^{2/3}),$$

using the fact that $Q_\epsilon(t)\le C\epsilon$. Therefore

$$H_\epsilon(t)+Q_\epsilon(t)\le (H_\epsilon(0)+Q_\epsilon(0)+C\epsilon^{2/3}t)\exp(Ct)$$

and finally

$$H_\epsilon(t)\le (H_\epsilon(0)+C\epsilon^{2/3}(1+t))\exp(Ct),$$

and the result follows. Check that the constant C depends only on $\sup_{0\le s\le t}\{\|D^3\bar\phi, D^2\partial_t\bar\phi\|_{L^\infty(\mathbb{T}^2)}\}$. This ends the proof of Theorem 6.1 □

Proof of Lemma 6.2. First we show that if $\Theta(R)=\int_{\{|\nabla\phi^\epsilon|\ge R\}}|\nabla\phi^\epsilon|^2$, then, for some $C>0$,

$$\Theta(R)\le C\int|\nabla\phi^\epsilon-\nabla\bar\phi|^2+\frac{C}{R^2}. \tag{45}$$

Indeed, $\int|\nabla\phi^\epsilon|^2\le C$ implies that $\text{meas}\{|\nabla\phi^\epsilon|\ge R\}\le C\frac{1}{R^2}$. Since $|\nabla\bar\phi(t,x)|\le C$ for $(t,x)\in[0,T']\times\mathbb{T}^d$, we have

$$\Theta(R)\le 2\int_{\{|\nabla\phi^\epsilon|\ge R\}}|\nabla\bar\phi|^2+2\int_{\{|\nabla\phi^\epsilon|\ge R\}}|\nabla\phi^\epsilon-\nabla\bar\phi|^2$$
$$\le\frac{2C}{R^2}+2\int|\nabla\phi^\epsilon-\nabla\bar\phi|^2.$$

Hence (45) is proved for C replaced by $\max\{2,2C\}$.

Then, letting

$$K(x)=D^2\bar\phi-\int_{s=0}^1 D^2\bar\phi(x-s\epsilon\nabla\phi^\epsilon)\,ds,$$

we have

$$\Xi\le C\Theta(R)+\int_{|\nabla\phi^\epsilon|\le R}|K(x)||\nabla\phi^\epsilon|^2$$

with $|K(x)|\le C\epsilon|\nabla\phi^\epsilon|$, and thus

$$\Xi\le C\epsilon\int_{|\nabla\phi^\epsilon|\le R}|\nabla\phi^\epsilon|^3+C\Theta(R)$$
$$\le C\left(\epsilon R\int|\nabla\phi^\epsilon|^2+\frac{1}{R^2}+\int|\nabla\phi^\epsilon-\nabla\bar\phi|^2\right)$$
$$\le C\left(\epsilon R+\frac{1}{R^2}+\int|\nabla\phi^\epsilon-\nabla\bar\phi|^2\right)$$

for all R, so for $R=\epsilon^{-1/3}$ we obtain

$$\Xi\le C\epsilon^{2/3}+C\int|\nabla\phi^\epsilon-\nabla\bar\phi|^2.$$

This proves Lemma 6.2 □

6.3. Convergence of strong solutions. We present here another proof of convergence, which holds for stronger norms. Let us consider as above the solution $(\bar\rho, \bar\phi)$ to Euler,

$$\begin{aligned}\partial_t\bar\rho + \nabla\cdot(\bar\rho\nabla\bar\phi^\perp) &= 0,\\ \Delta\bar\phi &= \bar\rho,\end{aligned}$$

and we recall the SG_ϵ system

$$\begin{aligned}\partial_t\rho^\epsilon + \nabla\cdot(\rho^\epsilon\nabla\psi^{\epsilon\perp}) &= 0,\\ \det(I+\epsilon D^2\psi^\epsilon) &= 1+\epsilon\rho^\epsilon.\end{aligned}$$

We then have the following theorem.

THEOREM 6.3. *Let $(\bar\rho,\bar\phi)$ be a solution of EI, such that $\bar\rho\in C^2_{loc}(\mathbb{R}^+\times\mathbb{T}^2)$. Let $\rho^{\epsilon\,0}$ be a sequence of initial data for SG_ϵ satisfying* (38)–(39)*, and such that $\frac{\rho^{\epsilon\,0}-\bar\rho^0}{\epsilon}$ is bounded in $W^{1,\infty}(\mathbb{T}^2)$. Then there exists a sequence $(\rho^\epsilon,\psi^\epsilon)$ of solutions to SG_ϵ that satisfies the following: For all $T>0$, there exists $\epsilon_T>0$, such that the sequence*

$$\frac{\rho^\epsilon-\bar\rho}{\epsilon}, \frac{\nabla\psi^\epsilon-\nabla\bar\phi}{\epsilon}$$

for $0<\epsilon<\epsilon_T$ is uniformly bounded in $L^\infty([0,T],W^{1,\infty}(\mathbb{T}^2))$.

Remark. In the previous theorem, we obtained estimates in the L^2 norm; here we obtain estimates in Lipschitz norm. Estimates of higher derivatives follow in the same way.

Proof of Theorem 6.3. We expand the solution of SG_ϵ as the solution of EI plus a small perturbation of order ϵ and show that this perturbation remains bounded in large norms (at least Lipschitz). We first remark that the assumption on $\bar\rho$ implies that for all $T>0$, $\bar\phi\in L^\infty([0,T];C^3(\mathbb{T}^2))$. Let us write

$$\begin{aligned}\rho^\epsilon &= \bar\rho+\epsilon\rho_1,\\ \psi^\epsilon &= \bar\phi+\epsilon\psi_1.\end{aligned}$$

Rewritten in terms of ρ_1,ψ_1, the SG_ϵ system reads

$$\begin{aligned}&\partial_t\rho_1 + (\nabla\bar\phi+\epsilon\nabla\psi_1)^\perp\cdot\nabla\rho_1 = -\nabla\psi_1^\perp\cdot\nabla\bar\rho,\\ &\Delta\psi_1+\epsilon\ \mathrm{trace}\,[D^2\psi_1 D^2\bar\phi]+\epsilon^2\det D^2\psi_1 = \rho_1-\det D^2\bar\phi.\end{aligned}$$

Differentiating the first equation with respect to space, we find the evolution equation for $\nabla\rho_1$:

(46)

$$\partial_t\nabla\rho_1 + ((\nabla\bar\phi+\epsilon\nabla\psi_1)^\perp\cdot\nabla)\nabla\rho_1 = -(D^2\bar\phi+\epsilon D^2\psi_1)\nabla\rho_1^\perp - D^2\psi_1\nabla\bar\rho^\perp - D^2\bar\rho\nabla\psi_1^\perp.$$

We claim that in order to conclude the proof it is enough to have an estimate of the form

$$\|\psi_1(t,\cdot)\|_{C^{1,1}(\mathbb{T}^2)} \le C(1+\|\rho_1(t,\cdot)\|_{C^{0,1}(\mathbb{T}^2)}), \tag{47}$$

where C depends on $\bar\phi$. Let us admit this bound temporarily, and finish the proof of the theorem: Using (47) and (46), we obtain

$$\frac{d}{dt}\|\nabla\rho_1\|_{L^\infty} \le C(t)(1+\|\nabla\rho_1\|_{L^\infty}+\epsilon\|\nabla\rho_1\|^2_{L^\infty}),$$

where the constant $C(t)$ depends on the $C^2(\mathbb{T}^2)$ norm of $(\bar\rho(t,\cdot),\bar\phi(t,\cdot))$. This quantity is bounded on every interval $[0,T]$.

Thus we conclude using Gronwall's lemma that $\|\nabla\rho_1(t,\cdot)\|_{L^\infty(\mathbb{T}^2)}$ remains bounded on $[0,T_\epsilon]$ with T_ϵ going to T as ϵ goes to 0. We then choose T as large as we want, since when $d=2$ the smooth solution to EI is global in time. From estimate (47) the $W^{1,\infty}$ bound on ρ_1 implies a $W^{2,\infty}$ bound on ψ_1. Then, we remember that

$$\rho_1=\frac{\rho^\epsilon-\bar\rho}{\epsilon},\quad \nabla\psi_1=\frac{\nabla\psi^\epsilon-\nabla\bar\phi}{\epsilon}$$

to conclude the proof of Theorem 6.3. □

Proof of the estimate (47). We write the equation followed by ψ_1 as follows:

$$\Delta\psi_1=-\text{trace }[\epsilon D^2\psi_1 D^2\bar\phi]-\epsilon^2\det D^2\psi_1+\rho_1-\det D^2\bar\phi.$$

We recall that

$$\|fg\|_{C^{2,\alpha}}\le\|f\|_{C^{2,\alpha}}\|g\|_{C^{2,\alpha}};$$

hence, using Schauder $C^{2,\alpha}$ estimates for solutions to the Laplace equation (see [14]), we have

$$\|\psi_1\|_{C^{2,\alpha}}\le C_1(1+\epsilon\|\psi_1\|_{C^{2,\alpha}}+\epsilon^2\|\psi_1\|^2_{C^{2,\alpha}}),\tag{48}$$

where C_1 depends on $\|\bar\phi\|_{C^{2,\alpha}},\|\rho_1\|_{C^\alpha}$. Inequality (48) will be satisfied in two cases: Either for $\|\psi_1\|_{C^{2,\alpha}}\le C_2$ or for $\|\psi_1\|_{C^{2,\alpha}}\ge C_3\epsilon^{-2}$, where C_2,C_3 are positive constants that depend on C_1.

Now we show that ψ^ϵ, the solution of (37), is bounded in $C^{2,\alpha}$ for ρ^ϵ bounded in the C^α norm. We consider for $t\in[0,1]$ ψ^ϵ_t the unique up to a constant periodic solution of

$$\det(I+\epsilon D^2\psi^\epsilon_t)=1+t\epsilon\rho^\epsilon.$$

Differentiating this equation with respect to t, we find

$$M_{ij}D_{ij}\partial_t\psi^\epsilon_t=\rho^\epsilon,$$

where M is the comatrix of $I+\epsilon D^2\psi^\epsilon_t$. From the regularity result of Theorem 4.3, M is C^α and strictly elliptic. From Schauder estimates, we then have $\|\partial_t\psi^\epsilon_t\|_{C^{2,\alpha}}\le C\|\rho^\epsilon\|_{C^{2,\alpha}}$, and integrated over $t\in[0,1]$, we get

$$\|\psi^\epsilon\|_{C^{2,\alpha}}\le C\|\rho^\epsilon\|_{C^{2,\alpha}}.$$

Hence, since $\psi^\epsilon=\bar\phi+\epsilon\psi_1$, we have ψ_1 bounded by C/ϵ in $C^{2,\alpha}$. Hence it cannot be bigger than C_3/ϵ^2, and to satisfy (48), we must have

$$\|\psi_1\|_{C^{2,\alpha}}\le C_2,$$

where C_2 as above depends on $\|\bar\phi\|_{C^{2,\alpha}},\|\rho_1\|_{C^\alpha}$. This proves estimate (47). □

Acknowledgments. The author thanks Mike Cullen for his remarks, and also Yann Brenier, since part of this work was done under his direction, during the author's Ph.D. thesis. He also thanks Robert McCann and the Fields Institute of Toronto for their hospitality.

REFERENCES

[1] J.-D. Benamou and Y. Brenier, *Weak existence for the semigeostrophic equations formulated as a coupled Monge–Ampère/transport problem*, SIAM J. Appl. Math., 58 (1998), pp. 1450–1461.

[2] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[3] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[4] Y. Brenier, *Convergence of the Vlasov-Poisson system to the incompressible Euler equations*, Comm. Partial Differential Equations, 25 (2000), pp. 737–754.

[5] Y. Brenier and G. Loeper, *A geometric approximation to the Euler equations: The Vlasov-Monge-Ampère system*, Geom. Funct. Anal., 14 (2004), pp. 1182–1218.

[6] L. A. Caffarelli, *Interior $W^{2,p}$ estimates for solutions of the Monge-Ampère equation*, Ann. of Math. 2, 131 (1990), pp. 135–150.

[7] L. A. Caffarelli, *A localization property of viscosity solutions to the Monge-Ampère equation and their strict convexity*, Ann. of Math. (2), 131 (1990), pp. 129–134.

[8] L. A. Caffarelli, *The regularity of mappings with a convex potential*, J. Amer. Math. Soc., 5 (1992), pp. 99–104.

[9] D. Cordero-Erausquin, *Sur le transport de mesures périodiques*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 199–202.

[10] M. Cullen and W. Gangbo, *A variational approach for the 2-dimensional semi-geostrophic shallow water equations*, Arch. Ration. Mech. Anal., 156 (2001), pp. 241–273.

[11] M. J. P. Cullen and H. Maroofi, *The fully compressible semi-geostrophic system from meteorology*, Arch. Ration. Mech. Anal., 167 (2003), pp. 309–336.

[12] M. J. P. Cullen and R. J. Purser, *Properties of the Lagrangian semi-geostrophic equations*, J. Atmospheric Sci., 46 (1989), pp. 2684–2697.

[13] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.

[14] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, 1983.

[15] J. Kovats, *Dini-Campanato spaces and applications to nonlinear elliptic equations*, Electron. J. Differential Equations, 37 (1999).

[16] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[17] G. Loeper, *Quasi-neutral limit of the Euler-Poisson and Euler-Monge-Ampère systems*, Comm. Partial Differential Equations, 30 (2005), pp. 1141–1167.

[18] G. Loeper, *Uniqueness of solutions to the Vlasov-Poisson system with bounded density*, J. Math. Pures Appl., in press.

[19] G. Loeper, *On the regularity of the polar factorization for time dependent maps*, Calc. Var. Partial Differential Equations, 22 (2005), pp. 343–374.

[20] M. C. Lopes Filho and H. J. Nussenzveig Lopes, *Existence of a weak solution for the semigeostrophic equation with integrable initial data*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 329–339.

[21] A. J. Majda and A. L. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge Texts Appl. Math. 27, Cambridge University Press, Cambridge, UK, 2002.

[22] R. J. McCann, *A convexity principle for interacting gases*, Adv. Math., 128 (1997), pp. 153–179.

[23] R. J. McCann, *Polar factorization of maps on Riemannian manifolds*, Geom. Funct. Anal., 11 (2001), pp. 589–608.

[24] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.

[25] X. J. Wang, *Remarks on the regularity of Monge-Ampère equations*, in Proceedings of the International Conference on Nonlinear PDE (Hangzhou, 1992), F. H. Lin and G. C. Dong, eds., Academic Press, Beijing, 1992, pp. 257–263.

[26] V. Youdovich, *Non-stationary flows of an ideal incompressible fluid*, Zh. Vych. Mat., 3 (1963), pp. 1032–1066.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 824–844

ON THE CONTROL OF AN INTERACTING PARTICLE ESTIMATION OF SCHRÖDINGER GROUND STATES*

MATHIAS ROUSSET†

Abstract. We consider a general Schrödinger operator $L + V$ on a domain $E \subset \mathbb{R}^d$ and its associated positive ground state h solution to the maximal eigenvalue problem $L(h) + Vh = \lambda h$. In this work, an interacting particle model approximating the pair (h, λ) is studied. When $V \leq 0$, a basic version of this particle system consists of N walkers evolving independently according to the Markov generator L, each walker dying at a rate given by the value of the potential $|V|$ at the walker's current location; when a walker dies, any other one splits in two. The long time distribution of the particle system is then an estimator of h. Under some reasonable assumptions (with examples for $E = \mathbb{R}^d$), we get a nonasymptotic control of the $\mathbb{L}^p$ deviations (resp., the bias) of this estimator with the genuine rate of convergence in $1/\sqrt{N}$ (resp., $1/N$). We also compute explicitly the asymptotic standard deviation of the estimation of λ, which remains bounded in usual mild situations.

Key words. Schrödinger ground states, stochastic particle methods, long time behavior, quantum Monte Carlo

AMS subject classifications. 35Q40, 60J35, 65C35, 81-08

DOI. 10.1137/050640667

1. Introduction.

Our motivation can be split into two steps:

1. Control the long time behavior of an interacting particle approximation of Feynman–Kac formulas with genuine rate of convergence.
2. Use the long time distribution of the particle system as a Monte Carlo estimator of the ground state of Schrödinger operators.

The last question is of very high practical interest in quantum physics and chemistry, where one uses such diffusion Monte Carlo methods to compute observables of systems (see [3], [2] and the references therein). In the difficult yet crucial case of Fermi systems, the so-called fixed node approximation is used [3], [2], where one is resorting to the ground state of a general Schrödinger operator on a domain of $\mathbb{R}^d$.

We focus in this work on the interacting particle system (IPS) studied by Del Moral and Miclo in [7]. In its diffusive time-continuous version, it is particularly well suited to this context. Indeed, the fixed number of particles and the selection mechanism make it liable to be stable in the long run and to give rise to finite variance. Note that it has not yet inspired as such practitioners' heuristics. Several keys are given here to design it in practice, and some toy simulations will soon be available on the author's web page and in [12].

For the analysis we have used some semigroup and martingale techniques inherited from [6]. However, this paper is mostly self-contained. The good rate of convergence of the long time distribution of the IPS is a new result, technically demanding, and proved in a very reasonable setting which includes examples in $\mathbb{R}^d$. Intermediate results can be used to make some proofs of [7] precise (see Remark 5.5). For the stability questions, we have used a Foster–Lyapunov drift criterion to prove uniform

*Received by the editors September 19, 2005; accepted for publication (in revised form) March 10, 2006; published electronically August 7, 2006.
http://www.siam.org/journals/sima/38-3/64066.html
†Laboratoire J. A. Dieudonné, Université de Nice Sophia-Antipolis, 06108 Nice, France (mathias.rousset@polytechnique.org).

exponential convergence of Schrödinger semigroups (Proposition 2.2) under a quite general assumption, which seems to be a new point also.

If $K(x, dy)$ is an integral kernel, φ a test function, and μ a probability measure, we will use the notation $\mu K(\varphi) = \int \varphi(y) K(x, dy)\mu(dx)$. $(\quad)^+$ and $(\quad)^-$ denote, respectively, the positive and negative part.

Let us now give the main results of this paper. Suppose we are given an irreducible strong Feller diffusion X_t in an open connected domain $E \subset \mathbb{R}^d$ with generator L, reversible with respect to a probability measure $\mu(dx) = \frac{h_I^2(x)dx}{\int h_I^2(x)dx}$ for some $h_I \geq 0$. V denotes a potential function such that the Feynman–Kac semigroup

$$P_t^V(\varphi)(x) = \mathbb{E}\left(\varphi(X_t)e^{\int_0^t V(X_s)ds} | X_0 = x\right), \quad \varphi \in \mathbb{L}^2(\mu),$$

is strongly continuous in $\mathbb{L}^2(\mu)$ and Fellerian ($x \mapsto P_t^V(\varphi)(x)$ is continuous for φ bounded; see [8]). It gives rise to its associated self-adjoint Schrödinger operator

$$(L + V)(\varphi) = \lim_{t \to 0^+} \frac{P_t^V(\varphi) - \varphi}{t} \in \mathbb{L}^2(\mu)$$

defined on its domain $\mathcal{D}(L + V)$, where the latter limit exists (see [8]).

Our main example (detailed in section 2.1), which arises in many practical situations of interest (again in [3]), is some "importance sampling" transformation of the usual Schrödinger operator (with $h_I > 0$)

$$(L + V)(.) = h_I^{-1}\left(\frac{\Delta}{2} + V_0\right)(h_I .), \tag{1}$$

which leaves the spectrum of $\frac{\Delta}{2} + V_0$ invariant, and multiplies eigenfunctions by h_I^{-1}. X_t is then a Brownian motion with local drift $\nabla \ln h_I$.

We will work under the following usual assumption.

ASSUMPTION 1. *The spectrum of $L + V$ is bounded by a greatest eigenvalue λ and has a spectral gap $\lambda^* > 0$. λ is associated with a unique eigenfunction $h \in \mathbb{L}^2(\mu)$ (the ground state), which is continuous and strictly positive.*

Note that Assumption 1 is very general and idiomatic; see [8, Chapter 3], [4], [9], [11], and the example of section 2.1.

By spectral theory, we get that

$$P_t^{V-\lambda}(\varphi) \xrightarrow[t \to +\infty]{\exp} h\mu(h\varphi) \quad \text{in } \mathbb{L}^2(\mu)$$

with rate $\lambda^\star > 0$. If the initial probability law η_0 of X_0 has a density in $\mathbb{L}^2(\mu)$, the Cauchy–Schwarz inequality gives

$$\eta_0 P_t^{V-\lambda}(\varphi) \xrightarrow[t \to +\infty]{\exp} \eta_0(h)\mu(h\varphi).$$

This is not sufficient to compute h numerically, since of course λ is unknown. That's why we resort to the renormalized version of the semigroup

$$\eta_t = \frac{\eta_0 P_t^V(\varphi)}{\eta_0 P_t^V(1)} = \frac{\eta_0 P_t^{V-\lambda}(\varphi)}{\eta_0 P_t^{V-\lambda}(1)}.$$

This probability flow verifies from the discussion above that

$$\eta_t(\varphi) \xrightarrow[t \to +\infty]{\exp} \frac{\mu(h\varphi)}{\mu(h)} = \eta_\infty(\varphi),$$

the ground state eigenvalue λ can be recovered from η_∞ by the identity

$$\eta_\infty(V) = \frac{\mu(-L(h) + \lambda h)}{\mu(h)} = \lambda,$$

and the Feynman–Kac semigroup can be recovered from η_t by

$$\eta_0 P_t^V(\varphi) = \eta_t(\varphi) \exp\left(\int_0^t \eta_s(V) ds\right).$$

Thus a stochastic particle approximation of η_t enables the computation of P_t^V, of λ, and of h under a renormalized weak form.

Now we consider continuous and bounded potentials $V \in \mathcal{C}_b(E)$ and smooth test functions $\varphi \in \mathcal{C}_b^\infty(E)$, and we remark then that η_t is a weak solution to the "nonlinear" Fokker–Planck equation

$$\begin{aligned} \partial_t \eta_t(\varphi) &= \eta_t\left(L(\varphi) + (V - \eta_t(V))\varphi\right) \\ &= \eta_t(L_{\eta_t}(\varphi)). \end{aligned} \tag{2}$$

The "nonlinear" Markov generator L_η is a jump perturbation of L defined by (other choices are possible, as in the abstract; see subsection 3.1)

$$L_\eta(\varphi)(x) = L(\varphi)(x) + \int_E (\varphi(y) - \varphi(x)) \left((V(x) - \eta(V))^- + (V(y) - \eta(V))^+\right) \eta(dy).$$

To compute η_t, we construct a particle system associated to this mean-field interpretation. The latter is denoted $\xi_t = (\xi_t^1, \dots, \xi_t^N) \in E^N$ with initial law $\eta_0^{\otimes N}$, and its Markov generator is given by

$$\mathcal{L}(\psi)(\xi) = \sum_{i=1}^N L_{m(\xi)}^{(i)}(\psi)(\xi) \quad \text{with} \quad m(\xi) = \frac{1}{N}\sum_{j=1}^N \delta_{\xi^j} \tag{3}$$

for any $\xi = (\xi^1, \dots, \xi^N) \in E^N$. The exponent (i) means that the operator acts on the ith coordinate of the test function $\psi \in \mathcal{C}_b^\infty(E^N)$. The empirical measure of the particle system ξ_t denoted

$$\eta_t^N = m(\xi_t) = \frac{1}{N}\sum_{j=1}^N \delta_{\xi_t^j}$$

is then a stochastic approximation of η_t and converges to the ground state η_∞ in the long run.

ξ_t consists of N walkers evolving independently according to the Markov generator L, but constrained by the following birth and death mechanism:

1. With rate $(V(\xi_t^i) - \eta_t^N(V))^-$, each walker ξ_t^i jumps to the location of a uniformly randomly chosen walker.
2. With rate $(V(\xi_t^i) - \eta_t^N(V))^+$, a uniformly randomly chosen walker jumps to the location of each walker ξ_t^i.

Under some localization assumptions (Assumptions 1, 2, and 3, with examples in $\mathbb{R}^d$), we prove a strong control on the long time behavior of this IPS:

$$\sup_{T\geq 0}\mathbb{E}(|\eta_T^N(\varphi)-\eta_T(\varphi)|^p)^{1/p}\leq \frac{C_p\|\varphi\|_\infty}{\sqrt{N}},$$
$$\sup_{T\geq 0}|\mathbb{E}(\eta_T^N(\varphi))-\eta_T(\varphi)|\leq \frac{C\|\varphi\|_\infty}{N},$$
$$\sup_{T\geq 0}\|\mathrm{Law}(\xi_T^i)-\eta_T\|_{tv}\leq \frac{C}{N}.$$

To get a more quantitative result, we then consider the asymptotic standard deviation of the estimator of λ,

$$\mathrm{Ad}^2(V)=\lim_{N\to+\infty}\overline{\lim}_{T\to+\infty}N\mathbb{E}\left((\eta_T^N(V)-\lambda)^2\right), \tag{4}$$

and we give an explicit upper bound on the latter, which remains finite in the usual mild situations with general unbounded potentials V.

2. Assumptions and examples. We begin with our main example (see also [3]), which motivates the results of the paper.

2.1. Example. We say that a positive function h has exponential fall-off at infinity as soon as $-\ln h$ goes to infinity at least linearly.

Let E be an open domain of $\mathbb{R}^d$ with boundary ∂E. Classically, we consider the Schrödinger operator $\frac{\Delta}{2}+V_0$, with V_0 continuous on $\bar{E}$ and going to $-\infty$ at infinity $\lim_{+\infty}V_0=-\infty$.

$\frac{\Delta}{2}+V_0$ is then self-adjoint for the core $\mathcal{C}_c^\infty(E)$ of smooth test functions with compact support in E (Dirichlet conditions) and has compact resolvent (see [11, Chapter 13]). The operator thus has a discrete spectrum with maximal eigenvalue λ, a spectral gap λ^*, and a ground state $h_0>0$ on E. h_0 is continuous on $\bar{E}$ with $h_0|_{\partial E}=0$ and has exponential fall-off at infinity (see [1]).

Now we consider the importance sampling transformation (1) for $h_I\in\mathcal{C}^\infty\cap\mathbb{L}^2(\bar{E})$, with $h_I>0$ on E, $h_I|_{\partial E}=0$ and exponential fall-off. The resulting operator $L+V$ then reads

$$L=\frac{\Delta}{2}+\nabla\ln h_I\nabla,$$
$$V=V_0+h_I^{-1}\frac{\Delta}{2}h_I.$$

$L+V$ is self-adjoint for the core $\mathcal{C}_c^\infty(E)$ in $\mathbb{L}^2(\mu)$ with $\mu(dx)=\frac{h_I^2(x)dx}{\int h_I^2(x)dx}$; it has the same spectrum as $\frac{\Delta}{2}+V_0$, but with continuous ground state $h=h_0h_I^{-1}>0$. As a consequence, $L+V$ satisfies Assumption 1.

To stick to our probabilistic setting we additionally ask that the following hold:

1. $\nabla h_I\neq 0$ on ∂E.
2. For some constant a and b, $x.\nabla\ln h_I(x)\leq a|x|^2+b$ for all $x\in\bar{E}$.
3. V or $h_I^{-1}\frac{\Delta}{2}h_I$ is bounded above.

By Proposition 7 of [3], L then defines a nonexplosive strong Feller diffusion X_t in E, verifying the stochastic differential equations (for some Brownian motion $t\mapsto W_t$)

$$dX_t=dW_t+\nabla\ln h_I(X_t)dt,$$

and reversible with respect to μ. Here are two examples satisfying Assumptions 2 and 3.

Remark 2.1. Within the context of section 2.1, Assumptions 2 and 3 are satisfied as soon as V_0 is Hölder continuous, and that there is an $\epsilon > 0$ such that, outside some compact set,

$$\epsilon \leq h = h_I h_0^{-1} \leq \epsilon^{-1}, \tag{5}$$

$$-\epsilon^{-1} \leq h_I^{-1} \frac{\Delta}{2} h_I - h_0^{-1} \frac{\Delta}{2} h_0 \leq -\epsilon. \tag{6}$$

Heuristically, this means that h_I and h_0 must have a similar behavior outside compact sets, h_I being chosen slightly more concave than h_0.

First example for bounded domains. Suppose now that E is bounded and V_0 is Hölder. The Schrödinger operator is regularizing and h_0 is smooth. Note this classical fact:[1] $\nabla h_0 \neq 0$ on ∂E. It is now always possible to construct explicitly an h_I satisfying (5) and (6) and thus Assumptions 1, 2, and 3.

Proof. On the boundary ∂E, ∇h_0, and ∇h_I are nondegenerated and directed along the normal vector of ∂E. This ensures (5). Now adjust the concavity of h_I near the boundary so that (6) is satisfied. ∎

Second example for unbounded domains. This case is slightly more intricate, so we only give a particular explicit example:

Suppose that $E = \mathbb{R}^d$, V_0 is Hölder, and that h_0 has the following expression:

$$h_0(x) = \mathrm{e}^{-\frac{|x|^4}{4} + \epsilon_0(x)},$$

where ϵ_0 is smooth and bounded with bounded first derivatives.

Now if we choose h_I such that, outside some compact set,

$$h_I(x) = \mathrm{e}^{-\frac{|x|^4}{4} + \epsilon_0(x) - \frac{C}{|x|^2}}$$

for some $C > 0$, then (5) and (6), and thus Assumptions 1, 2, and 3, are satisfied.

Proof. Inequality (5) is obvious. A straightforward computation shows that at infinity $V(x) - \lambda = -4C + o(|x|^{-2})$, which gives (6). ∎

Third example for general situations. Here is our last example, which less restrictive (neither V nor h^{-1} shall be bounded). Assumptions 2 and 3 are not satisfied, but the expression of the asymptotic standard deviation of the eigenvalue estimation $\mathrm{Ad}(V)$ (defined by (4)) remains finite, which is a very favorable indication of practical efficiency.

Take $E = \mathbb{R}^d$ and suppose V_0 behaves polynomially at infinity. Choose h_I such that

1. $\ln h_I$ and its first two derivatives are of polynomial behavior;
2. $h = h_0 h_I^{-1}$ is bounded with exponential fall-off.

Then the expression of $\mathrm{Ad}(V)$ remains bounded. Note that it is practically easy to choose such an h_I, since the exponential fall-off of h_0 is known from V_0 (see [1]).

Proof. Recall that $V = V_0 + \frac{\Delta}{2} \ln(h_I) + \frac{1}{2}(\nabla \ln(h_I))^2$ is polynomially dominated, and that $\frac{d\eta_\infty}{d\mu}(x) \propto h(x) = h_0(x) h_I^{-1}(x)$, $\frac{d\eta_\infty}{dx}(x) \propto h_0(x) h_I(x)$, and $\frac{d\mu}{dx}(x) \propto h_I^2(x)$ are bounded with exponential fall-off. The result follows then from Proposition 4.5. ∎

This latter case could be generalized to noncontinuous potentials V_0 lying locally in the Kato class (see [5]).

[1] Consider Theorems 4.7 and 4.19 of [5], and an integration by parts between $\frac{\Delta}{2} h_0$ and positive solutions of the Dirichlet boundary value problem.

2.2. Convergence of semigroups in the uniform sense. We define the nonlinear propagator associated to η_t by

$$\Phi_{t,T}(\nu) = \frac{\nu P^V_{T-t}}{\nu P^V_{T-t}(1)} \in \mathcal{P}(E).$$

By the semigroup property, it verifies the propagation equation $\eta_T = \Phi_{t,T}(\eta_t)$.

In this subsection, we give an assumption for the uniform convergence of $P_t^{V-\lambda}$ and its consequence for the stability of $\Phi_{t,T}$. This will be crucial for the stability of the particle approximation. The only assumption we need is the following.

ASSUMPTION 2. *V is bounded above and there is an $\epsilon > 0$ such that the subset $K_\epsilon = \{x \in E \mid V(x) - \lambda \geq -\epsilon\}$ is relatively compact in E.*

This is a natural physical assumption, which ensures that h is a "strict bound" state in the sense that V is a strict potential barrier outside some compact set.

We then have the following.

PROPOSITION 2.2. *Under Assumptions 1 and 2, the Feynman–Kac semigroup is uniformly exponentially converging in the sense that there is some $C \geq 0$ and $0 < \rho < 1$ such that for any test function φ,*

$$\|P_t^{V-\lambda}(\varphi) - h\mu(h\varphi)\|_\infty \leq \|\varphi\|_\infty C\rho^t.$$

Proof. We use the results developed by Tweedie and collaborators, for instance, in [6]. We consider the strong Feller irreducible Markov diffusion semigroup

$$P_t^h(\varphi) = h^{-1}P_t^{V-\lambda}(h\varphi),$$

its associated diffusion process X_t^h, and its extended generator $L^h = h^{-1}(L + V - \lambda)(h.)$, reversible with respect to $h^2(x)\mu(dx)$. We show that h^{-1} is a strict Lyapunov function for L^h outside $\bar{K}_\epsilon$ (in the sense of condition $(\tilde{\mathcal{D}})$ of [7]). Indeed we have that

1. $\bar{K}_\epsilon$ is compact and thus is a petite set for X_t^h (see [13, Theorems 7.1 and 5.1]);
2. $L^h(h^{-1}) + \epsilon h^{-1} = (V - \lambda + \epsilon)h^{-1}$ is bounded on $\bar{K}_\epsilon$ and negative outside.

So by Theorem 5.2 of [7], X_t^h is h^{-1}-uniformly ergodic, which means that

$$\sup_{|g| \leq h^{-1}} |P_t^h(g)(x) - \mu(h^2 g)| \leq h^{-1}(x)C\rho^t,$$

and gives the result for $\varphi = gh$. □

We then harvest the uniform stability of $\Phi_{t,T}$.

COROLLARY 2.3. *Under Assumptions 1 and 2, we have for some $C \geq 0$, $0 < \rho < 1$, and any $\nu \in \mathcal{P}(E)$*

$$|\Phi_{t,T}(\varphi)(\nu) - \eta_\infty(\varphi)| \leq \|\varphi\|_\infty \frac{C}{\nu(h)}\rho^{T-t}.$$

Proof. We take $\|\varphi\|_\infty \leq 1$ and use the Landau symbol "O" uniformly with respect to t, T, ν, and φ. From Proposition 2.2 we get

$$\begin{aligned}\Phi_{t,T}(\varphi)(\nu) = \frac{\nu P^{V-\lambda}_{T-t}(\varphi)}{\nu P^{V-\lambda}_{T-t}(1)} &= \frac{\nu(h)\mu(h\varphi) + O(\rho^{T-t})}{\nu(h)\mu(h) + O(\rho^{T-t})} \\ &= \frac{\mu(h\varphi) + O(\frac{\rho^{T-t}}{\nu(h)})}{\mu(h) + O(\frac{\rho^{T-t}}{\nu(h)})},\end{aligned}$$

which gives the result. □

2.3. A last assumption. To construct the particle system and carry out the long time analysis, we will need some more boundedness and regularity hypotheses.

ASSUMPTION 3.

1. *V is continuous and bounded.*
2. *$\ln h$ is continuous and bounded.*
3. *For $\varphi \in \mathcal{C}_b^\infty(E)$, $(t,x) \mapsto P_t^V(\varphi)(x)$ is $\mathcal{C}_b^{1,2}(E \times \mathbb{R}^+)$.*

Remark 2.4. $\mathcal{C}_b^{1,2}(E \times \mathbb{R}^+)$ denotes bounded continuous functions of E with continuous time derivative and continuous second order space derivatives. The regularity assumption is probably necessary only for intermediate technical purpose.

The second assumption could be replaced by $\sup_{t,N} \mathbb{E}\big(\frac{1}{(\eta_t^N(h))^p}\big) < +\infty$.

The regularity of $P_t^V(\varphi)$ and V gives the backward Fokker–Planck equation in a pointwise sense.

LEMMA 2.5. *For all $\varphi \in \mathcal{C}_b^\infty(E)$, we have*

$$\begin{aligned}\partial_t P_{T-t}^V(\varphi) &= -P_{T-t}^V(L(\varphi) + V\varphi) \\ &= -L(P_{T-t}^V(\varphi)) + V P_{T-t}^V(\varphi).\end{aligned}$$

3. The IPS approximation.

3.1. The generator of the IPS. In this subsection, we design the interacting particle interpretation of the flow $(\eta_t)_{t\geq 0}$, with initial probability $\eta_0 \in \mathcal{P}(E)$. The bounded potential being given, we first consider two continuous bounded applications $(\mathcal{P}(E), \text{weak topology}) \to (\mathcal{C}_b(E), \| \ \|_\infty)$, whose images are nonnegative functions denoted

$$\eta \mapsto V_\eta^b \geq 0, \quad \eta \mapsto V_\eta^d \geq 0$$

and verifying

$$V_\eta^b(x) - V_\eta^d(x) = V(x) + C_\eta,$$

where C_η does not depend on x (as explained in section 3.2, "b" stands for "birth" and "d" for "death").

We define

$$V_\eta^* = V_\eta^b + V_\eta^d.$$

Example 3.1. Here are several possible choices of the above functions:

1. $V^b = 0, \quad V^d = \sup(V) - V$ (as in the abstract).
2. $V^b = V^+, \quad V^d = V^-$.
3. $V_\eta^b = (V - \eta(V))^+, \quad V_\eta^d = (V - \eta(V))^-$.

The last choice is of fundamental importance since it is invariant by the transformation $V \mapsto V + C$, which leaves η_t invariant.

Recall from (2) that η_t satisfies the fundamental nonlinear Markovian evolution equation

$$\partial_t \eta_t(\varphi) = \eta_t(L_{\eta_t}(\varphi)),$$

but here the nonlinear Markov generator is more generally defined by

$$L_\eta(\varphi)(x) = L(\varphi)(x) + \int (\varphi(y) - \varphi(x))(V_\eta^b(y) + V_\eta^d(x))\, \eta(dy).$$

Indeed we have that

$$\begin{aligned}\eta(L_\eta(\varphi)) &= \eta(L(\varphi)) + \eta(V_\eta^b\varphi) - \eta(V_\eta^d\varphi) + \eta(V_\eta^d)\eta(\varphi) - \eta(V_\eta^b)\eta(\varphi)\\ &= \eta(L(\varphi) + V\varphi) - \eta(\varphi)\eta(V).\end{aligned}$$

If A is a linear operator, the associated formal "carré-du-champs" Γ_A is a bilinear operator defined by

$$\Gamma_A(\varphi,\varphi) = A(\varphi^2) - 2\varphi A(\varphi).$$

Recall that when A is the generator of a Markov process X_t, $\Gamma_A(\varphi,\varphi) \geq 0$ and $\int_0^t \Gamma_A(\varphi,\varphi)(X_s)ds$ is the predictable quadratic variation of the martingale part of $\varphi(X_t)$.

We can then define

$$\Gamma_{L_\eta}(\varphi,\varphi)(x) = \Gamma_L(\varphi,\varphi)(x) + \int (\varphi(y)-\varphi(x))^2(V_\eta^b(y) + V_\eta^d(x))\eta(dy)$$

and remark that

$$\begin{aligned}\eta(\Gamma_{L_\eta}(\varphi,\varphi)) &= \eta(\Gamma_L(\varphi,\varphi)) + \eta(\varphi^2 V_\eta^*) + \eta(\varphi^2)\eta(V_\eta^*) - 2\eta(V_\eta^*\varphi)\eta(\varphi)\\ &= \eta(\Gamma_L(\varphi,\varphi)) + \eta\left((\varphi-\eta(\varphi))^2(V_\eta^* + \eta(V_\eta^*))\right).\end{aligned}$$

We now consider the interacting particle model $(\xi_t)_{t\in\mathbb{R}^+}$ associated to the nonlinear operator L_η as defined in the introduction by its initial law $\eta_0^{\otimes N}$ and its Markov generator $\mathcal{L}$ given in (3). The IPS is a Markov process resulting from a bounded jump perturbation of N independent copies of X_t, and thus is well defined.

When we use as a test function the empirical mean $m(.)(\varphi) \in \mathcal{C}_b^\infty(E^N)$ of a $\varphi \in \mathcal{C}_b^\infty(E)$, we have the following simple form of the generator and its associated carré-du-champs.

LEMMA 3.2.

$$\begin{aligned}\mathcal{L}(m(.)(\varphi)) &= m(.)(L_{m(.)}(\varphi)),\\ \Gamma_{\mathcal{L}}(m(.)(\varphi), m(.)(\varphi)) &= \frac{1}{N} m(.)(\Gamma_{L_{m(.)}}(\varphi,\varphi)).\end{aligned}$$

Proof. The first identity is by definition. The second one is a straightforward formal computation. We use the linearity of $A \mapsto \Gamma_A$ to get

$$\Gamma_{\mathcal{L}}(\psi,\psi) = \sum_{i=1}^N \Gamma^{(i)}_{L_{m(.)}}(\psi,\psi),$$

and since $L_\eta(\text{constant}) = 0$, for any $\xi \in E^N$, we have

$$\begin{aligned}\Gamma^{(i)}_{L_{m(\xi)}}(m(.)(\varphi), m(.)(\varphi))(\xi) &= \frac{1}{N^2} 2\sum_{j\neq i} \varphi(\xi^j) L_{m(\xi)}(\varphi)(\xi^i) + \frac{1}{N^2} L_{m(\xi)}(\varphi^2)(\xi^i)\\ &\quad - 2\left(\frac{1}{N}\sum_j \varphi(\xi^j)\right)\frac{1}{N} L_{m(\xi)}(\varphi)(\xi^i)\\ &= \frac{1}{N^2}\Gamma^{(i)}_{m(\xi)}(\varphi,\varphi)(\xi^i),\end{aligned}$$

and the result follows. □

Now we can state our key tool.

PROPOSITION 3.3. *For all $\varphi_. \in \mathcal{C}_b^{1,2}(E \times \mathbb{R}^+)$, the process*

$$\mathcal{M}_t(\varphi_.) = \eta_t^N(\varphi_t) - \eta_0^N(\varphi_0) - \int_0^t \eta_s^N \left(\partial_s \varphi_s + L_{\eta_s^N}(\varphi_s)\right) ds$$

is a local martingale, with predictable quadratic variation given by

$$\left\langle \mathcal{M}(\varphi_.)\right\rangle_0^t = \frac{1}{N} \int_0^t \eta_s^N \left(\Gamma_{L_{\eta_s^N}}(\varphi_s, \varphi_s)\right) ds,$$

and jumps estimated by

$$|\Delta \mathcal{M}_t(\varphi_.)| \leq \frac{2\|\varphi_t\|}{N}.$$

We recall that

$$\eta_s^N(L_{\eta_s^N}(\varphi)) = \eta_s^N \left(L(\varphi) + (V - \eta_s^N(V))\varphi\right)$$

and

$$\eta_s^N(\Gamma_{L_{\eta_s^N}}(\varphi,\varphi)) = \eta_s^N(\Gamma_L(\varphi,\varphi)) + \eta_s^N \left((\varphi - \eta_s^N(\varphi))^2 (V^*_{\eta_s^N} + \eta_s^N(V^*_{\eta_s^N}))\right).$$

Proof. This is a particular case of the usual martingale problem associated to the Markov process ξ_t. The statement can be proved with a standard application of the Itô formula, with Markov property arguments for the jump part.

The estimate on the jumps follows from the fact that each jump concerns only one particle (see the probabilistic construction in subsection 3.2). □

From the above proposition we immediately get the stochastic differential equation

$$d\eta_t^N(\varphi) = \eta_t^N(L_{\eta_t^N}(\varphi))dt + d\mathcal{M}_t(\varphi),$$

which is a perturbation of (2) of the dynamic of η_t by a martingale whose jumps and predictable quadratic variation are of order $\frac{1}{N}$. In this sense, we already see that η_t^N is a natural approximation of the flow η_t. Of course, this point of view is too elementary to enable an asymptotic analysis mainly because of the nonlinearity of (2).

3.2. Probabilistic construction and genetic interpretation. We start with a more explicit expression for the IPS generator.

PROPOSITION 3.4. *We have $\mathcal{L} = \mathcal{L}^{mut} + \mathcal{L}^{sel}$ with the pair mutation/selection generators defined by*

$$\begin{aligned}
\mathcal{L}^{mut}(\psi)(\xi) &= \sum_{i=1}^N L^{(i)}(\psi)(\xi), \\
\mathcal{L}^{sel}(\psi)(\xi) &= \sum_{i=1}^N V^d_{m(\xi)}(\xi^i) \frac{1}{N} \sum_{j=1}^N (\psi(\xi^{i\to j}) - \psi(\xi)) \\
&\quad + \sum_{i=1}^N V^b_{m(\xi)}(\xi^i) \frac{1}{N} \sum_{j=1}^N (\psi(\xi^{j\to i}) - \psi(\xi)),
\end{aligned}$$

where if $\xi' = \xi^{i\to j}$, then $\xi'^k = \xi^k$ except for $k = i$, where $\xi'^i = \xi^j$.

Proof. The jump part of $\mathcal{L}$ is by definition

$$\mathcal{L}^{\mathrm{sel}}(\psi)(\xi) = \sum_{i=1}^{N} V^d_{m(\xi)}(\xi^i)\left(\frac{1}{N}\sum_{j=1}^{N}\psi(\xi^{i\to j}) - \psi(\xi)\right) + \sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N} V^b_{m(\xi)}(\xi^j)(\psi(\xi^{i\to j}) - \psi(\xi)),$$

and the result follows by exchanging the indexes i and j in the second part of the right-hand side of the identity. ∎

Thus the N walkers evolve according to the following birth and death mechanism, for any $i \in [1, N]$ ($\tau^{d/b,i}_{n+1}$ designing independent exponential clocks of mean 1):

1. Between each jump time, the walkers evolve independently according to the mutation generator L.
2. At random times $T^{d,i}_n$ defined by $\int_{T^{d,i}_n}^{T^{d,i}_{n+1}} V^d_{\eta^N_s}(\xi^i_s)ds = \tau^{d,i}_{n+1}$, a walker is uniformly randomly chosen, and the ith walker then jumps to its location.
3. At random times $T^{b,i}_n$ defined by $\int_{T^{b,i}_n}^{T^{b,i}_{n+1}} V^b_{\eta^N_s}(\xi^i_s)ds = \tau^{b,i}_{n+1}$, a walker is uniformly randomly chosen, and then jumps to the location of the ith walker.

This explains how the selection generator tends to "get rid of" walkers with relatively high potential $V^d_{\eta^N_t}$, and tends to "reproduce" walkers with relatively high potential $V^b_{\eta^N_t}$. The effect of selection is then to favor walkers with relatively high potential $V = V^b_{\eta^N_t} - V^d_{\eta^N_t} - C_{\eta^N_t}$. In this sense, the IPS can be seen as a continuous time genetic algorithm with fitness function V and mutations of generator L.

Moreover this structure enables a nice parallelized implementation, where walkers are individually collecting information from V, but yet learns globally the structure of the ground state h.

In practice, one may use some Euler discretization scheme, and may approximate integrals with sums. This requires at least the continuity of the potential V.

4. Long time behavior of the IPS.

4.1. Nonasymptotic control.

We give directly the main theoretical results of this paper, the proof being postponed to section 5.

THEOREM 4.1 (time-uniform $\mathbb{L}^p$ estimate). *We suppose that Assumptions* 1, 2, *and* 3 *are verified.*

There are constants C_p such that, for all test function $\varphi \in \mathcal{C}_b(E)$ with $\|\varphi\|_\infty \le 1$,

$$\sup_{T\ge 0}\mathbb{E}\left(|\eta^N_T(\varphi) - \eta_T(\varphi)|^p\right)^{1/p} \le \frac{C_p}{\sqrt{N}}.$$

THEOREM 4.2 (bias estimate/time-uniform convergence of a particle). *We suppose that Assumptions* 1, 2, *and* 3 *are verified.*

There is a constant C such that, for all $\varphi \in \mathcal{C}_b(E)$ with $\|\varphi\|_\infty \le 1$,

$$\sup_{T\ge 0}\left|\mathbb{E}\left(\eta^N_T(\varphi)\right) - \eta_T(\varphi)\right| \le \frac{C}{N},$$

$$\sup_{T\ge 0}\|Law(\xi^i_T) - \eta_T\|_{tv} \le \frac{C}{N}.$$

Remark 4.3. One can easily show that the particle system ξ_t is recurrent and ergodic. If the invariant measure is finite, it converges in law to a random variable ξ_∞ (this always happens when E is compact). η_∞^N is then the natural estimator of η_∞ and we have the almost sure convergence

$$\eta_\infty^N \xrightarrow[N\to+\infty]{a.s.} \eta_\infty \quad \text{in the weak topology,}$$

which follows the $\mathbb{L}^p$ estimate for $p=4$ with a Borel–Cantelli argument.

This situation is probably true in general for N large enough, although the positivity seems difficult to prove. Anyway, one can take for η_∞^N any adherent limit (which is a positive measure) of η_t^N under the weak topology of evanescent functions.

We want to lay the emphasis on the difficulty of the proof of these results, which comes from the nonlinear propagation of the error made by the particle approximation.

We also propose an asymptotic study of the standard deviation.

4.2. Long time asymptotic standard deviation. The asymptotic standard deviation gives a quantitative information of the IPS approximation. We show in Proposition 4.5 that the latter is likely to remain bounded in many mild situations of interest.

THEOREM 4.4. *Under Assumptions* 1, 2, *and* 3, *we have for any* $\varphi \in \mathcal{C}_b(E)$ ($\bar\varphi = \varphi - \eta_\infty(\varphi)$)

$$\begin{aligned}&\lim_{N\to+\infty} \overline{\lim}_{T\to+\infty} N\mathbb{E}\left((\eta_T^N(\varphi)-\eta_\infty(\varphi))^2\right)\\ &= \mathrm{Ad}^2(\varphi) = \eta_\infty(\bar\varphi^2) + 2\int_0^{+\infty} \eta_\infty\left(P_s^{V-\lambda}(\bar\varphi)^2(V_{\eta_\infty}^b + \eta_\infty(V_{\eta_\infty}^d))\right) ds.\end{aligned}$$

Note that by Proposition 2.2, the local noise introduced by interactions

$$s \mapsto \eta_\infty\left(P_s^{V-\lambda}(\bar\varphi)^2(V_{\eta_\infty}^b + \eta_\infty(V_{\eta_\infty}^d))\right)$$

is exponentially decreasing with s.

We're interested in clarifying this quantity for the meaningful case $\varphi = V$, which corresponds to the eigenvalue estimation. We will take $V_\eta^b = (V-\eta(V))^+$ and $V_\eta^d = (V-\eta(V))^-$.

PROPOSITION 4.5. *Under Assumption* 1 *only, we have*

$$\mathrm{Ad}^2(V) \le \eta_\infty((V-\lambda)^2) + \frac{1}{\lambda^*}\left\|\frac{d\eta_\infty}{d\mu} \times \left((V-\lambda)^+ + \eta_\infty((V-\lambda)^-)\right)\right\|_\infty \mu((V-\lambda)^2).$$

Proof. Since $\bar\varphi = \varphi - \eta_\infty(\varphi)$ is orthogonal to h in $\mathbb{L}^2(\mu)$, we have by spectral theory $\mu(P_t^{V-\lambda}(\bar\varphi)^2) \le e^{-2\lambda^* t}\mu(\bar\varphi^2)$. The result follows from Theorem 4.4 for $\varphi = V$ with $\eta_\infty(V) = \lambda$. □

When $\frac{d\eta_\infty}{d\mu} = \frac{h}{\mu(h)}$ is bounded with exponential fall-off, this upper bound is expected to remain finite in almost any situation of interest (see the third example of section 2.1).

5. Proofs. Throughout this section we will use the following notation:

1. $T > 0$ will be a deterministic horizon time and we will take $t \in [0,T]$.
2. $n \ge 0$ $p \ge 1$ are integers.
3. $\|\ \|$ is the uniform norm.
4. $\varphi \in \mathcal{C}_b^\infty(E)$ is a test function such that $\|\varphi\| \le 1$, and $\bar\varphi = \varphi - \eta_T(\varphi)$.

5. $C > 0$ a constant independent of test functions, of the time parameters t, T and of the number of particles N. In the same spirit, we will use the Landau notation "O" uniformly with these variables. Note: The constant C and the "O" notation may depend on integers n and p.

We recall that $\|V\| \leq C$ and $\sup_\eta \|V_\eta^*\| \leq C$.

The proofs are based on the use of a "linearized" version of the propagator of η_t defined by

$$Q_{t,T}(\varphi) = \frac{P^V_{T-t}(\varphi)}{\eta_t P^V_{T-t}(1)},$$

which verifies the propagation equation

$$\eta_T(\varphi) = \eta_t Q_{t,T}(\varphi).$$

The main idea is to analyze the martingale part and the predictable part of the process $t \mapsto \eta_t^N Q_{t,T}(\bar{\varphi})$ for $\bar{\varphi} = \varphi - \eta_T(\varphi)$. Because we have $\eta_t(Q_{t,T}(\bar{\varphi})) = 0$, it can be interpreted as a stochastic perturbation of the identically null process. Note that $\eta_T^N(Q_{T,T}(\bar{\varphi})) = \eta_T^N(\varphi) - \eta_T(\varphi)$, which is the quantity we wish to control when $T \to +\infty$.

To lighten the computations, the test function φ will be omitted.

Throughout these proofs, we will use the following stability results.

LEMMA 5.1. *The propagator $Q_{t,T}$ verifies the following properties:*
n being given, there is a C such that for any test function φ

$$\|Q_{t,T}(\varphi)\| \leq C,$$
$$\int_t^T \|Q_{s,T}(\varphi)\|^{2^n} ds \leq C(T-t).$$

Moreover there is some $0 < \rho < 1$ such that for any $\bar{\varphi} = \varphi - \eta_T(\varphi)$

$$\|Q_{t,T}(\bar{\varphi})\| \leq C\rho^{T-t},$$
$$\int_t^T \|Q_{s,T}(\bar{\varphi})\|^{2^n} ds \leq C.$$

Proof. First we write

$$Q_{t,T}(\varphi) = \frac{P^{V-\lambda}_{T-t}(\varphi)(x)}{\eta_t P^{V-\lambda}_{T-t}(1)}.$$

We claim that

$$\frac{1}{\eta_t P^{V-\lambda}_{T-t}(1)} \leq C.$$

Indeed by definition and the semigroup property we have $\eta_t P^{V-\lambda}_{T-t}(1) = \frac{\eta_0 P^{V-\lambda}_T(1)}{\eta_0 P^{V-\lambda}_t(1)}$, and $t \mapsto \eta_0 P^{V-\lambda}_t(1)$ is continuous, positive, and goes from 1 to $\eta_0(h)\mu(h) > 0$.

By Proposition 2.2, we then get for any φ $\|Q_{t,T}(\varphi)\| \leq C$.

For $\bar{\varphi} = \varphi - \eta_T(\varphi)$ we use the decomposition

$$|Q_{t,T}(\bar{\varphi})| = \frac{1}{(\eta_t P^{V-\lambda}_{T-t}(1))^2} |\eta_t P^{V-\lambda}_{T-t}(1) P^{V-\lambda}_{T-t}(\varphi) - \eta_t P^{V-\lambda}_{T-t}(\varphi) P^{V-\lambda}_{T-t}(1)|,$$

and again Proposition 2.2 gives a $0 < \rho < 1$ such that for any $\bar{\varphi}$, $\|Q_{t,T}(\bar{\varphi})\| \leq C\rho^{T-t}$. □

Note the control on the initial error.

LEMMA 5.2. *We have for any* φ

$$\mathbb{E}((\eta_0^N(\varphi) - \eta_0(\varphi))^p) \leq \frac{C}{N^{p/2}}.$$

Proof. Since, at time $t = 0$, all particles are sampled independently with law η_0, $\eta_0^N(\varphi)$ is a sum of N zero-mean independently and identically distribute (i.i.d.) variables. The result is then the Burkholder–Davies–Gundy (BDG) inequality for i.i.d. variables. □

5.1. Precise L^p-estimate of the key martingales. We want to apply Proposition 3.3 to the collection $(Q_{t,T}(\varphi)^{2^n})_{n\geq 0} \equiv (Q_{t,T}^{2^n})_{n\geq 0}$. Recall that $\eta_t(P_{T-t}^V(1)) = \frac{\eta_0 P_T^V(1)}{\eta_0 P_t^V(1)}$ and so

$$\begin{aligned}\partial_t \eta_t\left(P_{T-t}^V(1)\right) &= -\frac{\eta_0 P_T^V(1)}{\eta_0 P_t^V(1)^2}\eta_0 P_t^V(V)\\ &= -\eta_t\left(P_{T-t}^V(1)\right)\eta_t(V)\ ;\end{aligned}$$

this yields, using Lemma 2.5,

$$\begin{aligned}\partial_t Q_{t,T} &= -L(Q_{t,T}) - VQ_{t,T} + \frac{1}{\eta_t(P_{T-t}^V(1))^2}\eta_t\left(P_{T-t}^V(1)\right)\eta_t(V)\\ &= -L(Q_{t,T}) - (V - \eta_t(V))Q_{t,T}\end{aligned}$$

and

$$\partial_t Q_{t,T}^{2^n} = -2^n Q_{t,T}^{2^n-1} L(Q_{t,T}) - 2^n Q_{t,T}^{2^n} \times (V - \eta_t(V))\,.$$

From Proposition 3.3, we obtain a collection of difference of martingales between t and T indexed by n:

$$\begin{aligned}\mathcal{M}_t^T(Q_{\cdot,T}^{2^n}) &= \mathcal{M}_T(Q_{\cdot,T}^{2^n}) - \mathcal{M}_t(Q_{\cdot,T}^{2^n})\\ &= \eta_T^N(Q_{T,T}^{2^n}) - \eta_t^N(Q_{t,T}^{2^n}) - \int_t^T \eta_s^N\left(L(Q_{s,T}^{2^n}) - 2^n Q_{s,T}^{2^n-1} L(Q_{s,T})\right) ds\\ &\quad - \int_t^T \eta_s^N\left(Q_{s,T}^{2^n} \times \left(V - \eta_s^N(V) - 2^n(V - \eta_s(V))\right)\right) ds, \qquad (7)\end{aligned}$$

with predictable quadratic variation given by

$$\begin{aligned}&N\left\langle \mathcal{M}(Q_{\cdot,T}^{2^n})\right\rangle_t^T\\ &= \int_t^T \eta_s^N\left(\Gamma_L\left(Q_{s,T}^{2^n}, Q_{s,T}^{2^n}\right)\right) + \eta_s^N\left(\left(Q_{s,T}^{2^n} - \eta_s^N(Q_{s,T}^{2^n})\right)^2\left(V^* + \eta_s^N(V^*)\right)\right) ds. \qquad (8)\end{aligned}$$

We can get rid of the carré-du-champs term and get the following bounds up to a martingale.

LEMMA 5.3. *For all* $n \geq 0$ *and any test function* φ *we have*

$$N\left\langle \mathcal{M}(Q_{\cdot,T}^{2^n}(\varphi))\right\rangle_t^T \leq C(T - t + 1) - \mathcal{M}_t^T(Q_{\cdot,T}^{2^{n+1}}(\varphi)),$$

and for any centered test function $\bar{\varphi} = \varphi - \eta_T(\varphi)$

$$N\big\langle \mathcal{M}(Q^{2^n}_{.,T}(\bar{\varphi}))\big\rangle^T_t \leq C - \mathcal{M}^T_t(Q^{2^{n+1}}_{.,T}(\bar{\varphi})).$$

Proof. The formal carré-du-champs upper bound (Lemma 6.1) gives

$$\int_t^T \eta^N_s\left(\Gamma_L(Q^{2^n}_{s,T}, Q^{2^n}_{s,T})\right) ds$$
$$\overset{(=\text{ for } n=0)}{\leq} \int_t^T \eta^N_s\left(L(Q^{2^{n+1}}_{s,T}) - 2^{n+1}Q^{2^{n+1}-1}_{s,T}L(Q_{s,T})\right) ds.$$

From (8), we use the above inequality and (7) at rank $n+1$ to find the upper bound

$$\begin{aligned} N\big\langle \mathcal{M}(Q^{2^n}_{.,T})\big\rangle^T_t &\overset{(=\text{ for } n=0)}{\leq} -\mathcal{M}^T_t(Q^{2^{n+1}}_{.,T}) + \eta^N_T(Q^{2^{n+1}}_{T,T}) - \eta^N_t(Q^{2^{n+1}}_{t,T}) \\ &\quad - \int_t^T \eta^N_s\left(Q^{2^{n+1}}_{s,T}\big(V - \eta^N_s(V) - 2^{n+1}(V - \eta_s(V))\big)\right) ds \\ &\quad + \int_t^T \eta^N_s\left(\big(Q^{2^n}_{s,T} - \eta^N_s(Q^{2^n}_{s,T})\big)^2\big(V^* + \eta^N_s(V^*)\big)\right) ds. \end{aligned} \tag{9}$$

The result follows then from Lemma 5.1. □

The case $n = 0$ is of crucial importance. From (7) or Proposition 3.3 we can get

$$d\eta^N_t(Q_{t,T}) = d\mathcal{M}_t(Q_{.,T}) + \big(\eta_t(V) - \eta^N_t(V)\big)\,\eta^N_t(Q_{t,T})dt, \tag{10}$$

which gives for centered test functions, by integrating on $[0,T]$,

$$\begin{aligned} \eta^N_T(\varphi) - \eta_T(\varphi) &= \eta^N_0(Q_{0,T}(\bar{\varphi})) + \mathcal{M}_T(Q_{.,T}(\bar{\varphi})) \\ &\quad + \int_0^T \big(\eta_s(V) - \eta^N_s(V)\big)\left(\eta^N_s(Q_{s,T}(\bar{\varphi})) - \eta_s(Q_{s,T}(\bar{\varphi}))\right) ds. \end{aligned} \tag{11}$$

The martingale part and the initial error $\eta^N_0(Q_{0,T}(\bar{\varphi}))$ is expected to be of order $\frac{1}{\sqrt{N}}$, and the predictable part of order $\frac{1}{N}$.

Note that by developing the right-hand side of (9) with the identity $V = V^b_{\eta^N_s} - V^d_{\eta^N_s} + C_{\eta^N_s}$, the predictable quadratic variation of the martingale gives

$$\begin{aligned} N\big\langle \mathcal{M}(Q_{.,T})\big\rangle^T_0 &= -\mathcal{M}_T(Q^2_{.,T}) + \eta^N_T(Q^2_{T,T}) - \eta^N_0(Q^2_{0,T}) \\ &\quad + 2\int_0^T \eta^N_s(Q^2_{s,T}V^b_{\eta^N_s}) + \eta^N_s(Q^2_{s,T})\eta_s(V^d_{\eta^N_s}) \\ &\quad - \eta^N_s(Q_{s,T})\eta^N_s(Q_{s,T}V^*_{\eta^N_s}) + \eta^N_s(Q^2_{s,T})\big(\eta^N_s(V^b_{\eta^N_s}) - \eta_s(V^b_{\eta^N_s})\big)ds. \end{aligned} \tag{12}$$

We can now state the first result of this section, which is the control of all moments of these martingales.

THEOREM 5.4. *For all* $p \geq 1$, *all* $n \geq 0$, *and all test functions* φ,

$$\mathbb{E}\left(\big(\big[\mathcal{M}(Q^{2^n}_{.,T}(\varphi)\big]^T_t\big)^p\right) \leq \frac{C(T-t+1)^p}{N^p},$$

and for centered test functions $\bar{\varphi}$,

$$\mathbb{E}\left(\big(\big[\mathcal{M}(Q^{2^n}_{.,T}(\bar{\varphi})\big]^T_t\big)^p\right) \leq \frac{C}{N^p}.$$

Proof. Note that by localization, we can suppose that we work with bounded martingales.

Thanks to the Jensen inequality, it is sufficient to prove the inequalities for all $p = 2^q$. We are going to use an induction on q to prove that

$$\begin{aligned}&\forall n \geq 0,\\ &\mathbb{E}\left(\left(\left\langle \mathcal{M}(Q^{2^n}_{\cdot,T})\right\rangle^T_t\right)^{2^q}\right) \leq \frac{C(T-t+1)^{2^q}}{N^{2^q}},\\ &\mathbb{E}\left(\left(\left[\mathcal{M}(Q^{2^n}_{\cdot,T})\right]^T_t\right)^{2^q}\right) \leq \frac{C(T-t+1)^{2^q}}{N^{2^q}}.\end{aligned}$$

For $q = 0$, these inequalities are a direct consequence of Lemma 5.3.

Suppose the inequality true at order q and lower. Again from Lemma 5.3, we get

$$\begin{aligned}&\mathbb{E}\left(N^{2^{q+1}}\left(\left\langle \mathcal{M}(Q^{2^n}_{\cdot,T})\right\rangle^T_t\right)^{2^{q+1}}\right)\\ &\quad\leq C(T-t+1)^{2^{q+1}} + C\mathbb{E}\left(\mathcal{M}^T_t(Q^{2^{n+1}}_{\cdot,T})^{2^{q+1}}\right)\\ &\quad\leq C(T-t+1)^{2^{q+1}} + C\mathbb{E}\left(\left(\left[\mathcal{M}(Q^{2^{n+1}}_{\cdot,T})\right]^T_t\right)^{2^q}\right) \quad \text{(by the BDG inequality).}\end{aligned}$$

By induction, this proves the first upper bound at rank $q+1$.

Now we use the alternate BDG inequality stated in Lemma 6.2 to the martingale $\mathcal{M}(Q^{2^n}_{\cdot,T})$, whose jumps, by Proposition 7, verify $a \leq \frac{2\|Q^{2^n}_{t,T}\|}{N} \leq \frac{C}{N}$. This gives

$$\mathbb{E}\left(\left(\left[\mathcal{M}(Q^{2^n}_{\cdot,T})\right]^T_t\right)^{2^{q+1}}\right) \leq C\sum_{k=0}^{q+1}\frac{1}{N^{2^{q+2}-2^{k+1}}}\mathbb{E}\Big(\left(\left\langle \mathcal{M}(Q^{2^n}_{\cdot,T})\right\rangle^T_t\right)^{2^k}\Big).$$

By induction,

$$\begin{aligned}\mathbb{E}\left(\left(\left[\mathcal{M}(Q^{2^n}_{\cdot,T})\right]^T_t\right)^{2^{q+1}}\right) &\leq C\sum_{k=0}^{q+1}\frac{(T-t)^{2^k}}{N^{2^{q+2}-2^{k+1}+2^k}}\\ &\leq \frac{C(T-t)^{2^{q+1}}}{N^{2^{q+1}}},\end{aligned}$$

which proves the second upper bound at rank $q+1$.

The case of centered test functions is identical. □

Remark 5.5. Some results of [6] use $\mathbb{L}^p$ estimates of a similar martingale (Lemma 3.23) whose proof uses the Itô formula but may be incorrect. Resorting to the Itô formula seems to be intractable, and the techniques used for Theorem 5.4 enables us to clarify these results.

5.2. Proof of Theorem 4.1. First of all let's define the following quantity:

$$I_p(N) = \sup_{T,\varphi}\mathbb{E}\left(\left(\eta^N_T(\varphi) - \eta_T(\varphi)\right)^p\right).$$

In this subsection, we will prove the time-uniform estimate $I_p(N) \leq \frac{C}{N^{p/2}}$. We start with our first key lemma.

LEMMA 5.6. *There is an $\epsilon > 0$ independent of p such that*

$$I_p(N) \leq \frac{C}{N^{\epsilon p/2}}.$$

Proof. Let's fix T and use the decomposition

$$\eta_T^N(\varphi) - \eta_T(\varphi) = \underbrace{\eta_T^N(\varphi) - \Phi_{t,T}(\eta_t^N)(\varphi)}_{a(t)} + \underbrace{\Phi_{t,T}(\eta_t^N)(\varphi) - \eta_T(\varphi)}_{b(t)}.$$

$a(t)$ can be controlled by the stochastic errors made by the particle approximation between t and T. $b(t)$ can be controlled by the stability property of the limiting propagator Φ between t and T. $b(0)$ can also be controlled by the error made by the initial condition. We then optimize the whole in t.

Control of $a(t)$. Let us define the continuous finite variation process

$$A_{t_1}^{t_2} = \exp\left(\int_{t_1}^{t_2} (\eta_s^N(V) - \eta_s(V))ds\right).$$

An elementary integration by parts for $s \in [t,T]$ gives

$$\begin{aligned} d\big(A_t^s\, \eta_s^N(Q_{s,T})\big) &= \eta_s^N(Q_{s,T})\, A_t^s\, (\eta_s^N(V) - \eta_s(V))\, ds \\ &\quad + A_t^s\, d\eta_s^N(Q_{s,T}) \\ &= A_t^s\, d\mathcal{M}_s(Q_{.,T}) \quad \text{(by (10))}. \end{aligned}$$

Integrating from t to T and simplifying by $(A_t^T)^{-1}$ gives

$$\eta_T^N(\varphi) - (A_t^T)^{-1}\eta_t^N(Q_{t,T}(\varphi)) = (A_t^T)^{-1}\int_t^T A_t^s\, d\mathcal{M}_s(Q_{.,T}(\varphi)). \tag{13}$$

Now, recalling that $\Phi_{t,T}(\eta_t^N)(\varphi) = \frac{(A_t^T)^{-1}\eta_t^N(Q_{t,T}(\varphi))}{(A_t^T)^{-1}\eta_t^N(Q_{t,T}(1))}$, we write $a(t)$ as

$$a(t) = \eta_T^N(\varphi) - (A_t^T)^{-1}\eta_t^N(Q_{t,T}(\varphi)) - \Phi_{T-t}(\eta_t^N)(\varphi)\big(1 - (A_t^T)^{-1}\eta_t^N(Q_{t,T}(1))\big),$$

which using (13) gives the upper bound

$$|a(t)| \le (A_t^T)^{-1}\left(\left|\int_t^T A_t^s\, d\mathcal{M}_s(Q_{.,T}(\varphi))\right| + \left|\int_t^T A_t^s\, d\mathcal{M}_s(Q_{.,T}(1))\right|\right),$$

and thus

$$\begin{aligned} \mathbb{E}\left(|a(t)|^p\right) &\le Ce^{2\|V\|p(T-t)}\mathbb{E}\left(\left|\int_t^T A_t^s\, d\mathcal{M}_s(Q_{.,T}(\varphi))\right|^p\right) \\ &\le Ce^{2\|V\|p(T-t)}\mathbb{E}\left(\left|\int_t^T (A_t^s)^2\, d\big[\mathcal{M}(Q_{.,T}(\varphi))\big]_s\right|^{p/2}\right) \text{ (by the BDG inequality)} \\ &\le Ce^{4\|V\|p(T-t)}\mathbb{E}\left(\left|\big[\mathcal{M}(Q_{.,T}(\varphi))\big]_t^T\right|^{p/2}\right) \\ &\le Ce^{4\|V\|p(T-t)}\frac{(T-t+1)^{p/2}}{N^{p/2}} \quad \text{(by Theorem 5.4)} \\ &\le C\frac{R^{p(T-t)}}{N^{p/2}} \quad (\text{for } R = \mathrm{e}^{4\|V\|+1} > 1). \end{aligned}$$

Control of $b(t)$. We have for some $0 < \rho < 1$, as a direct consequence of Corollary 2.3,

$$\begin{aligned} \mathbb{E}(|b(t)|^p) &= \mathbb{E}\big(\big|\Phi_{t,T}(\eta_t^N)(\varphi) - \Phi_{t,T}(\eta_t)(\varphi)\big|^p\big) \\ &\le \left(\mathbb{E}\left(\frac{1}{\eta_t^N(h)^p}\right) + 1\right)C\rho^{T-t} \\ &\le C\rho^{T-t} \quad \text{by Assumption 3.} \end{aligned}$$

Control of $b(0)$. We remark that $\eta_0(Q_{0,T}(1)) = 1$ and write $b(0)$ as

$$b(0) = \big(\eta_0^N(Q_{0,T}(\varphi)) - \eta_0(Q_{0,T}(\varphi))\big) + \Phi_{0,T}(\eta_0^N)(\varphi)\big(\eta_0(Q_{0,T}(1)) - \eta_0^N(Q_{0,T}(1))\big),$$

which gives by Lemma 5.2

$$\mathbb{E}(|b(0)|^p) \leq \frac{C}{N^{p/2}}.$$

Global control. We have

$$\mathbb{E}((a(0)+b(0))^p) \leq C\frac{R^{pT}+1}{N^{p/2}},$$
$$\mathbb{E}((a(t)+b(t))^p) \leq C\frac{R^{p(T-t)}}{N^{p/2}} + C\rho^{p(T-t)} \quad (\forall t \in [0,T]).$$

Now we take $\epsilon = \frac{-\ln\rho}{-\ln\rho + \ln R}$ and remark that

$$\frac{R^{\frac{1}{2}\frac{\ln N}{\ln R - \ln\rho}}}{N^{1/2}} = \frac{1}{N^{\epsilon/2}}$$

and

$$\rho^{\frac{1}{2}\frac{\ln N}{\ln R - \ln\rho}} = \frac{1}{N^{\epsilon/2}}.$$

We then get $\mathbb{E}((\eta_T^N(\varphi) - \eta_T(\varphi))^p) \leq \frac{C}{N^{\epsilon p/2}}$ from the first inequality when $T \leq \frac{1}{2}\frac{\ln N}{\ln R - \ln\rho}$, and from the second one otherwise for $T - t = \frac{1}{2}\frac{\ln N}{\ln R - \ln\rho}$. ∎

Now we go back to (11), which readily gives

$$\big(\eta_T^N(\varphi) - \eta_T(\varphi)\big)^p \leq C\big(\eta_0^N(Q_{0,T}(\bar\varphi))\big)^p + C\mathcal{M}_T^p(Q_{\cdot,T}) + C\bigg(\int_0^T |\eta_s^N(V) - \eta_s(V)||\eta_s^N(Q_{s,T}(\bar\varphi))|ds\bigg)^p.$$

From Lemma 5.2,

$$\mathbb{E}(\eta_0^N(Q_{0,T}(\bar\varphi))^p) \leq \frac{C}{N^{p/2}}.$$

From Theorem 5.4,

$$\mathbb{E}\big(\mathcal{M}_T^p(Q_{\cdot,T}(\bar\varphi))\big) \leq \frac{C}{N^{p/2}}.$$

On the other hand, using the Hölder inequality, we also have

$$\bigg(\int_0^T |\eta_s^N(V) - \eta_s(V)||\eta_s^N(Q_{s,T}(\bar\varphi))|ds\bigg)^p$$
$$\leq \int_0^T |\eta_s^N(V) - \eta_s(V)|^p \bigg|\eta_s^N\bigg(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}\bigg)\bigg|^p \|Q_{s,T}(\bar\varphi)\|ds\bigg(\int_0^T \|Q_{s,T}(\bar\varphi)\|ds\bigg)^{p-1}$$
$$\leq C\int_0^T |\eta_s^N\bigg(\frac{V}{\|V\|}\bigg) - \eta_s\bigg(\frac{V}{\|V\|}\bigg)\bigg|^p \bigg|\eta_s^N\bigg(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}\bigg) - \eta_s\bigg(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}\bigg)\bigg|^p \|Q_{s,T}(\bar\varphi)\|ds.$$

Taking expectations, and then using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\mathbb{E}\left(\left(\int_0^T |\eta_s^N(V)-\eta_s(V)||\eta_s^N(Q_{s,T}(\bar\varphi))|ds\right)^p\right) &\le C\int_0^T I_{2p}(N)\|Q_{s,T}(\bar\varphi)\|ds\\ &\le CI_{2p}(N),\end{aligned} \tag{14}$$

which gives on the whole for all $p \le 1$

$$I_p(N) \le \frac{C}{N^{p/2}} + I_{2p}(N).$$

Applying this result to Lemma 5.6 gives

$$I_p(N) \le \frac{C}{N^{\inf(2\epsilon,1)p/2}},$$

and by induction we get

$$I_p(N) \le \frac{C}{N^{p/2}},$$

which ends the proof.

5.3. Proof of Theorem 4.2. We take expectation in (11) to find

$$\mathbb{E}(\eta_T^N(\varphi)) - \eta_T(\varphi) = \int_0^T \mathbb{E}\left(\left(\eta_s(V)-\eta_s^N(V)\right)\eta_s^N\left(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}\right)\right)\|Q_{s,T}(\bar\varphi)\|ds.$$

As above, we use the Cauchy–Schwarz inequality and find that

$$\begin{aligned}|\mathbb{E}(\eta_T^N(\varphi)) - \eta_T(\varphi)| &\le C\int_0^T I_2(N)\|Q_{s,T}(\bar\varphi)\|ds\\ &\le CI_2(N) \le \frac{C}{N},\end{aligned} \tag{15}$$

which gives the estimate on the bias.

The second result is a direct consequence of exchangeability of particles.

5.4. Proof of Theorem 4.4. The study of the asymptotic standard deviation relies on the following asymptotic behavior.

LEMMA 5.7. *For all $T > 0$, and $\bar\varphi = \varphi - \eta_T(\varphi)$,*

$$\begin{aligned}\mathbb{E}\left((\eta_T^N(\varphi)-\eta_T(\varphi))^2\right) &= \frac{1}{N}\mathbb{E}(\eta_T^N(\bar\varphi)^2)\\ &\quad + \frac{2}{N}\mathbb{E}\left(\int_0^T \eta_s^N\left(Q_{s,T}^2(\bar\varphi)V_{\eta_s^N}^b\right) + \eta_s^N\left(Q_{s,T}^2(\bar\varphi)\right)\eta_s(V_{\eta_s^N}^d)ds\right) + O\left(\frac{1}{N^{3/2}}\right).\end{aligned}$$

Proof. Again we start from (11):

$$\begin{aligned}\eta_T^N(\varphi) - \eta_T(\varphi) &= \underbrace{\eta_0^N(Q_{0,T}(\bar\varphi)) + \mathcal{M}_T(Q_{\cdot,T}(\bar\varphi))}_{a}\\ &\quad + \underbrace{\int_0^T \left(\eta_s(V)-\eta_s^N(V)\right)\eta_s^N(Q_{s,T}(\bar\varphi))\,ds}_{b},\end{aligned}$$

which gives

$$\mathbb{E}((\eta_T^N(\varphi)-\eta_T(\varphi))^2)=\mathbb{E}(a^2)+\mathbb{E}(b^2)+2\mathbb{E}(ab).$$

Now we note by the results of the previous section—that is to say, by (14)—that

$$\mathbb{E}(b^2)=\mathbb{E}\bigg(\bigg|\int_0^T(\eta_s(V)-\eta_s^N(V))\eta_s^N\bigg(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}\bigg)\|Q_{s,T}(\bar\varphi)\|ds\bigg|^2\bigg)=O\bigg(\frac{1}{N^2}\bigg),$$

and by Theorem 5.4 and Lemma 5.2

$$\mathbb{E}(a^2)=\mathbb{E}(\mathcal{M}_T(Q_{.,T}(\bar\varphi))^2)+\mathbb{E}(\eta_0^N(Q_{0,T}(\bar\varphi))^2)=O\bigg(\frac{1}{N}\bigg).$$

So we get

$$\mathbb{E}\left((\eta_T^N(\varphi)-\eta_T(\varphi))^2\right)=\mathbb{E}\left(\eta_0^N(Q_{0,T}(\bar\varphi))^2\right)+\mathbb{E}\big(\big\langle\mathcal{M}(Q_{.,T}(\bar\varphi))\big\rangle_0^T\big)+O\bigg(\frac{1}{N^{3/2}}\bigg).$$

Now we shall use the identity (12) to compute the asymptotic (with respect to N) value of $\mathbb{E}(\langle\mathcal{M}(Q_{.,T}(\bar\varphi))\rangle_0^T)$. For this purpose, we note that by Theorem 4.1

$$\begin{aligned}\bigg|\mathbb{E}\bigg(\eta_s^N\bigg(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}\bigg)\eta_s^N\bigg(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}V^*_{\eta_s^N}\bigg)\bigg)\bigg|&\le C\mathbb{E}\bigg(\bigg|\eta_s^N\bigg(\frac{Q_{s,T}(\bar\varphi)}{\|Q_{s,T}(\bar\varphi)\|}\bigg)\bigg|\bigg)\\&=O\bigg(\frac{1}{N^{1/2}}\bigg),\end{aligned}$$

and in the same way

$$\bigg|\mathbb{E}\bigg(\eta_s^N\bigg(\frac{Q^2_{s,T}(\bar\varphi)}{\|Q^2_{s,T}(\varphi)\|}\bigg)\big(\eta_s^N(V^b_{\eta_s^N})-\eta_s(V^b_{\eta_s^N})\big)\bigg)\bigg|=O\bigg(\frac{1}{N^{1/2}}\bigg).$$

Finally we remark that $\eta_0^N(Q_{0,T}(\bar\varphi))$ is a sum of centered i.i.d. random variables, and thus

$$\mathbb{E}\left(\eta_0^N(Q^2_{0,T}(\bar\varphi))\right)=\eta_0(Q^2_{0,T}(\bar\varphi))=\mathbb{E}\left(\eta_0^N(Q_{0,T}(\bar\varphi))^2\right),$$

which ends the proof. □

Now recall that Theorem 4.1 implies by a Borel–Cantelli argument:

$$\eta_s^N\xrightarrow{a.s.}\eta_s\quad\text{(in the weak sense)}$$

We take the limit first when $N\to+\infty$ in Lemma 5.7 uniformly with respect to T, and then when $T\to+\infty$. Of course, the two limits commute. This gives by the Lebesgue convergence theorem

$$\begin{aligned}\lim_{N\to+\infty}N\mathbb{E}\left((\eta_T^N(\varphi)-\eta_T(\varphi))^2\right)&=\eta_T\left((\varphi-\eta_T(\varphi))^2\right)\\&\quad+2\int_0^T\eta_s\big(Q^2_{s,T}(\bar\varphi)V^b_{\eta_s}\big)+\eta_s(Q^2_{s,T}(\bar\varphi))\eta_s(V^d_{\eta_s})ds.\end{aligned}$$

Now we do the change of variables $s\mapsto T-s$ in the above integrand and take the limit $T\to+\infty$. We have

$$\begin{aligned}\eta_{T-s}&\to\eta_\infty,\\\bar\varphi=\varphi-\eta_T(\varphi)&\to\bar\varphi=\varphi-\eta_\infty(\varphi),\\Q_{T-s,T}(\bar\varphi)&\to\frac{P_s^V(\bar\varphi)}{\eta_\infty P_s^V(1)}=P_s^{V-\lambda}(\bar\varphi),\end{aligned}$$

which gives the asymptotic standard deviation.

6. Two general lemmas.

LEMMA 6.1 (an upper bound for the "carré-du-champs" operator). *Let L be a Markov generator and Γ be its associated "carré-du-champs" operator defined by $\Gamma(\varphi,\varphi) = L(\varphi^2) - 2\varphi L(\varphi)$. Then we have the upper bound for all $n \geq 0$:*

$$\Gamma(\varphi^{2^n},\varphi^{2^n}) \leq L(\varphi^{2^{n+1}}) - 2^{n+1}\varphi^{2^{n+1}-1}L(\varphi).$$

Proof. Check out by induction the formal identity

$$\Gamma(\varphi^{2^n},\varphi^{2^n}) = L(\varphi^{2^{n+1}}) - 2^{n+1}\varphi^{2^{n+1}-1}L(\varphi) - \sum_{k=1}^{n} 2^{n+1-k}\varphi^{2^{n+1}-2^k}\Gamma(\varphi^{2^{k-1}},\varphi^{2^{k-1}}),$$

and use the positivity property $\Gamma(\varphi,\varphi) \geq 0$. □

LEMMA 6.2 (BDG inequalities). *Let M be a quasi-left-continuous (i.e., with continuous predictable increasing process) locally square-integrable martingale with $M_0 = 0$ and bounded jumps $\sup_t |\Delta M_t| \leq a < +\infty$. Then there is a constant C (dependant on q) such that*

$$\mathbb{E}\big(\sup_t M_t^{2^{q+1}}\big) \leq C\mathbb{E}\big(\big[M\big]_\infty^{2^q}\big) \leq C\sum_{k=0}^{q} a^{2^{q+1}-2^{k+1}}\mathbb{E}\big(\big\langle M\big\rangle_\infty^{2^k}\big).$$

Proof. The first inequality is the classical BDG inequality (p. 350 of [10]).

For the second, by localization, we can suppose that M is a square-integrable martingale. We are to use the martingale $N = \big[M\big] - \big\langle M\big\rangle$. Because $\big\langle M\big\rangle$ is continuous $\Delta N = \Delta\big[M\big] = (\Delta M)^2$. Moreover, N has finite variation, so

$$\begin{aligned}\big[N\big] &= \sum_{s\leq\cdot}(\Delta N_s)^2 = \sum_{s\leq\cdot}(\Delta M_s)^4 \\ &\leq a^2\sum_{s\leq\cdot}(\Delta M_s)^2 \leq a^2\big[M\big] \\ &\leq a^2(N + \big\langle M\big\rangle). \qquad (16)\end{aligned}$$

We will also us the general fact (C depends on q)

$$\forall x,y,\quad (x+y)^{2^q} \leq C(x^{2^q} + y^{2^q}). \qquad (17)$$

By definition of N and (17) we get

$$\mathbb{E}\big(\big[M\big]_\infty^{2^q}\big) \leq C\mathbb{E}\big(\sup_t N_t^{2^q}\big) + C\mathbb{E}\big(\big\langle M\big\rangle_\infty^{2^q}\big).$$

Now it remains to prove for any $q \geq 1$ that

$$\mathbb{E}\big(\sup_t N_t^{2^q}\big) \leq \sum_{k=0}^{q-1} Ca^{2^{q+1}-2^{k+1}}\mathbb{E}\big(\big\langle M\big\rangle_\infty^{2^k}\big), \qquad (18)$$

which we are going to do by induction on q. For $q = 1$, (18) follows from the BDG inequality applied to N_t, with (16). Suppose (18) is true for a given q. Applying again the BDG inequality to N_t, and using (16) and (17), we get

$$\mathbb{E}\big(\sup_t N_t^{2^{q+1}}\big) \leq Ca^{2^{q+1}}\mathbb{E}\big(\sup_t N_t^{2^q}\big) + Ca^{2^{q+1}}\mathbb{E}\big(\big\langle M\big\rangle_\infty^{2^q}\big).$$

Inequality (18) at rank $q+1$ then follows from the induction hypothesis. □

REFERENCES

[1] S. Agmon, *Lectures on Exponential Decay of Solutions of Second Order Elliptic Equations*, Princeton University Press, Princeton, NJ, 1982.

[2] R. Assaraf, M. Caffarel, and A. Khelif, *Diffusion Monte Carlo methods with a fixed number of walkers*, Phys. Rev. E, 61 (2000), pp. 4566–4575.

[3] E. Cancès, B. Jourdain, and T. Lelièvre, *Quantum Monte Carlo simulations of fermions. A mathematical analysis of the fixed-node approximation*, Math. Models Methods Appl. Sci., to appear.

[4] R. Carmona and J. Lacroix, *Spectral Theory of Random Schrödinger Operators*, Probab. Appl., Birkhäuser Boston, Boston, MA, 1990.

[5] K. L. Chung and Z. Zhao, *From Brownian Motion to Schrödinger's Equation*, 2nd ed., Grundlehren Math. Wiss. 312, Springer, Berlin, 2001.

[6] P. Del Moral and L. Miclo, *Branching and interacting particle systems approximations of Feynman–Kac formulae with applications to nonlinear filtering*, in Séminaire de Probabilités XXXIV, Lecture Notes in Math. 1729, Springer, Berlin, 2000, pp. 1–145.

[7] D. Down, S. P. Meyn, and R. L. Tweedie, *Exponential and uniform ergodicity of Markov processes*, Ann. Probab., 23 (1996), pp. 1671–1691.

[8] M. Fukushima, Y. Oshima, and M. Takeda, *Dirichlet Forms and Symmetric Markov Processes*, de Gruyter Stud. Math. 19, Walter de Gruyter, Berlin, 1994.

[9] P. D. Hislop and I. M. Sigal, *Introduction to Spectral Theory with Application to Schrödinger Operators*, Appl. Math. Sci. 113, Springer, New York, 1996.

[10] P. A. Meyer, *Un cours sur les intégrales stochastiques*, in Séminaire de Probabilités X, Lecture Notes in Math. 511, Springer, Berlin, 1976, pp. 245–400.

[11] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* IV: *Analysis of Operators*, Academic Press, New York, 1978.

[12] M. Rousset, *Continuous Time Population Monte Carlo and Computational Physics*, Ph.D. Thesis, Université Paul Sabatier, Toulouse, France, 2006.

[13] R. L. Tweedie, *Topological conditions enabling use of Harris methods in discrete and continuous time*, Acta Appl. Math., 34 (1994), pp. 175–188.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 845–873

ANALYSIS OF AN INVERSE FIRST PASSAGE PROBLEM FROM RISK MANAGEMENT*

LAN CHENG†, XINFU CHEN†, JOHN CHADAM†, AND DAVID SAUNDERS‡

Abstract. We study the following "inverse first passage time" problem. Given a diffusion process X_t and a probability distribution q on $[0,\infty)$, does there exist a boundary $b(t)$ such that $q(t) = \mathbb{P}[\tau \le t]$, where τ is the first hitting time of X_t to the time-dependent level $b(t)$? A free boundary problem for a parabolic partial differential operator is associated with the inverse first passage time problem. We prove the existence and uniqueness of a viscosity solution to this problem. We also investigate the small time behavior of the boundary $b(t)$, presenting both upper and lower bounds. Finally, we derive some integral equations characterizing the boundary.

Key words. free boundary problem, mathematical finance, first passage problem, boundary crossing, inverse problem

AMS subject classifications. 34A55, 35D05, 60G40, 45G10

DOI. 10.1137/050622651

1. Introduction.

In this paper we study the following free boundary problem: find a boundary $b(t)$ $(t > 0)$ and an unknown function $w = w(x,t)$ $(x \in \mathbb{R},\ t \ge 0)$ such that

$$
\begin{cases}
w_t(x,t) = \frac{1}{2}(\sigma^2 w_x)_x - \mu w_x & \text{for } x > b(t), \quad t > 0, \\
w(x,t) = p(t) & \text{for } x \le b(t), \quad t > 0, \\
w_x(x,t) = 0 & \text{for } x \le b(t), \quad t > 0, \\
0 \le w(x,t) < p(t) & \text{for } x > b(t), \quad t > 0, \\
w(x,0) = \mathbf{1}_{(-\infty,0)}(x) & \text{for } x \in \mathbb{R}, \quad t = 0,
\end{cases}
\tag{1.1}
$$

where $q(t) = 1 - p(t)$ is a given cumulative probability distribution function with the following properties:

$$
1 = p(0) = \lim_{t \searrow 0} p(t), \quad p(t_1) \ge p(t_2) \ge 0 \quad \forall\, t_1 < t_2. \tag{1.2}
$$

This problem arises from the consideration of the first passage times of diffusion processes to curved boundaries. More specifically, let X_t be the solution of the following stochastic differential equation:

$$
dX_t = \mu(X_t,t)dt + \sigma(X_t,t)dB_t, \qquad X_0 = 0, \tag{1.3}
$$

where B_t is a standard Brownian motion on a filtered probability space satisfying the usual conditions, $\mu : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ and $\sigma : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ are smooth bounded functions,

*Received by the editors January 12, 2005; accepted for publication (in revised form) January 26, 2006; published electronically August 22, 2006.
http://www.siam.org/journals/sima/38-3/62265.html

†Department of Mathematics, University of Pittsburgh, 301 Thackeray Hall, Pittsburgh, PA 15260 (cheng0522@hotmail.com, xinfu@pitt.edu, chadam@pitt.edu). The second author was supported by NSF grant DMS-0203991. The third author was supported by NSF grant DMS-9704567.

‡Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada (Dsaunders@math.uwaterloo.ca). This author was supported by NSF grant DMS-0310656.

and $\sigma(x,t) > \varepsilon > 0$ for all $x \in \mathbb{R}, t \geq 0$. For a given function $b : \mathbb{R}_+ \to \mathbb{R}$ we define the first passage time of the diffusion process X_t to the curved boundary $b(t)$ to be

$$\tau = \inf\{t > 0 \mid X_t \leq b(t)\}. \tag{1.4}$$

Two important problems concerning the first passage time of a diffusion process to a curved boundary are the following:

1. *The first passage problem*: Given a barrier function $b(t)$, find the survival probability $p(t)$ that X does not cross b before or at t:

$$p(t) := \mathbb{P}\{\tau > t\}. \tag{1.5}$$

2. *The inverse first passage problem*: Given a survival probability function $p(t)$, find a barrier function $b(t)$, such that (1.5) holds.

The first passage problem is a classical problem in probability and is the subject of a rather large literature. It is also fundamental in many applications of diffusion processes to engineering, physics, biology, and economics. For a survey of techniques for approximating and computing first passage times to curved boundaries and a discussion of their applications in the biological sciences, we refer the reader to [18]. The financial applications motivating the current study are discussed below.

The work of Peskir [16, 17] on the first passage problem is of particular relevance for the inverse problem discussed in this paper. In [17], he derived a sequence of integral equations[1]

$$t^{n/2} H_n\left(\frac{b(t)}{\sqrt{t}}\right) + \int_0^t (t-s)^{n/2} H_n\left(\frac{b(t)-b(s)}{\sqrt{t-s}}\right) \dot{p}(s)ds = 0, \quad n = -1, 0, 1, \ldots, \tag{1.6}$$

where $H_{-1}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $H_n(x) = \int_x^\infty H_{n-1}(z)dz$ for $n \geq 0$. In [16], under the assumption that $b(t)$ is C^1 on $(0,\infty)$, decreasing, and concave, Peskir derived the equality

$$\dot{p}(0+) = -\lim_{t \searrow 0} \frac{1}{2\sqrt{2\pi}} \frac{b(t)}{t^{3/2}} e^{-\frac{b^2(t)}{2t}} = -\lim_{t \searrow 0} \frac{\dot{b}(t)}{\sqrt{2\pi t}} e^{-\frac{b^2(t)}{2t}},$$

provided that the second or third limit exists.

The inverse first passage problem is much harder than the direct problem, and there has been relatively little work on it. The studies on the problem to date are principally concerned with the numerical calculation of the boundary $b(t)$ for a given $p(t)$. There is no publication proving the well-posedness (existence and uniqueness) of the boundary given the survival probability.

Our interest in the inverse first passage problem originates[2] from structural models of credit risk in financial markets. Consider a company whose asset value and debt at time $t \geq 0$ are denoted by A_t and D_t, respectively, with D_t assumed to be deterministic. Assume the following:

1. $D_0 \leq A_0$ and the company is in *default* at a time $t > 0$ if $A_t < D_t$.

[1] In this reference, the derivations are carried out for the case $\sigma \equiv 1$ and $\mu \equiv 0$, i.e., when X_t is a Brownian motion. As mentioned in the reference, the techniques directly extend to other diffusion processes.

[2] We thank A. Kreinin and R. Stamicar for introducing us to this problem and for helpful discussions.

2. $A_t \geq 0$ follows a diffusion process (e.g., a geometric Brownian motion, as in [14]).

It is convenient to use the *default index* X_t and the *barrier function* $b(t)$ defined by

$$X_t := \log \frac{A_t}{A_0}, \qquad b(t) := \log \frac{D_t}{A_0}.$$

Then X_t is a diffusion process satisfying (1.3). In this context, the inverse first passage time problem is the problem of finding the default barrier $b(t)$ given the survival function $p(t)$.

The problem of finding the default barrier $b(t)$ given the survival function $p(t)$ is of critical importance in many problems in modern credit risk management. Hull and White [8, 9] show how the default barrier $b(t)$ can be used, once it is computed, in a model for pricing credit default swaps with counterparty default, and basket credit default swaps, and remark that it could be used to price other, exotic, credit derivatives.[3] They also show how the required input to our problem, the (risk-neutral) survival probability $p(t)$, can be derived from observed market prices. Iscoe, Kreinin, and Rosen [10] show how the inverse first passage problem is a key component in a multistep integrated market and credit risk portfolio model. They state that in their framework, the survival probability $p(t)$ may be derived from the internal model of a financial institution, from an external credit rating agency, or from published transition matrices for credit ratings, assuming a Markov process (see also [12]).

1.1. Mathematical formulation and existing results. Formulating the problems in a PDE setting, we introduce a new function $w(x,t)$ being the probability that the company does not default before or at t and its default index X_t is bigger than x, i.e.,

$$w(x,t) := \mathbb{P}\{X_t > x,\, \tau > t\}. \tag{1.7}$$

Then the density function of X_t when $\tau > t$ can be computed by

$$u(x,t) = \frac{d}{dx}\mathbb{P}\{X(t) \leq x, \tau > t\} = (p(t) - w(x,t))_x. \tag{1.8}$$

From (1.3) and the Kolmogorov forward equation, we see that (assuming sufficient regularity) $w(x,t)$ $(x \in \mathbb{R}, t \geq 0)$ satisfies (1.1). From this we see the following:

- The first passage problem is to solve (1.1) for p, with given b.
- The inverse first passage problem is to solve (1.1) for b, with given p.

The first passage problem can be solved as follows. From the Kolmogorov forward equation, we obtain the following closed system for $u(x,t)$:

$$\begin{cases} u_t(x,t) = \frac{1}{2}(\sigma^2 u)_{xx} - (\mu u)_x & \text{for } x > b(t), \quad t > 0, \\ u(x,t) = 0 & \text{for } x \leq b(t), \quad t > 0, \\ u(x,0) = \delta(x) & \text{for } x > 0, \quad t = 0, \end{cases} \tag{1.9}$$

[3] A standard credit default swap (CDS) is a credit derivative in which a buyer provides a series of payments to a seller in exchange for the seller's commitment to make a payment to the buyer in the event that a third party (the reference entity) defaults on one of its bonds. Pricing a CDS with counterparty default means taking into account the fact that the seller itself may default on its obligations before the reference entity defaults (in which case the buyer would not receive payment). In a basket CDS the seller provides a payment to the buyer when the first of a set of reference entities defaults.

where δ is a Dirac measure concentrated at 0. Given sufficiently regular b, this system has a unique solution. Then p and $\dot{p}$ can be computed from the formulas

$$p(t) = \int_{b(t)}^{\infty} u(x,t)dx \qquad \forall\, t \geq 0, \tag{1.10}$$

$$\dot{p}(t) = -\frac{1}{2}(\sigma^2 u)_x\big|_{x=b(t)} \qquad \forall\, t \geq 0. \tag{1.11}$$

It is possible to compute the solution in a closed form in only a few special cases. However, there is a large literature on numerical and analytic approximations of the solution.

Our system (1.1) is obtained directly from (1.9) and (1.8). Observe that since w is obtained by integrating u, the Dirichlet condition on u at the boundary $b(t)$ (corresponding to an absorbing barrier for the diffusion process) becomes a Neumann condition on w. The additional condition $w(x,t) = p(t), x \leq b(t)$, is obvious from the definition of w (1.7).

Avellaneda and Zhu [1] were the first to use (1.9) and (1.11) to study the inverse first passage problem. They performed a change of variables from X_t to $Y_t = X_t - b(t)$, whose financial meaning is the risk-neutral distance-to-default process for the company. Denote by $f(y,t) = u(y + b(t), t)$ the probability density function of Y_t when $\tau > t$. Equations (1.9) and (1.11) are equivalent to

$$\begin{cases} f_t = \dot{b}(t) f_y - (\mu f)_y + \frac{1}{2}(\sigma^2 f_y)_y & \text{for } y > 0, \quad t > 0, \\ f(0,t) = 0 & \text{for } y = 0, \quad t > 0, \\ f(y,0) = \delta_0(y - b(0)) & \text{for } y > 0, \quad t = 0, \\ \frac{1}{2}\sigma^2 f_y(0,t) + \dot{p}(t) = 0 & \text{for } y = 0, \quad t > 0. \end{cases} \tag{1.12}$$

In [11], Iscoe and Kreinin demonstrated that a Monte Carlo approach can be applied to solve the inverse first passage problem in discrete time, essentially by reducing it to the sequential estimation of conditional distributions. Hull and White [9] also consider a time discretization and compute the boundary by solving a system of nonlinear equations at each time point. Huang and Tian [7] apply a probabilistic method to construct piecewise linear solutions to the problem of constructing the default barriers. In [19], both a Monte Carlo algorithm based on a piecewise linear approximation of the boundary and a method based on the integral equation (1.6) with $n = 1$ are studied.

In this paper, we are particularly interested in the following fundamental questions: (1) Given a survival probability function $p(t)$ satisfying (1.2), is there a barrier function $b(t)$? (2) If there is a barrier function, how many are there? (3) What is the behavior of the barrier function for small times? Namely, we are concerned with the well-posedness (existence and uniqueness) of the free boundary problem (1.1) and the asymptotic properties of $b(t)$ near zero. The motivation for studying the well-posedness of the problem is evident. Aside from its intrinsic interest, knowledge of the behavior of the boundary near zero is essential for effective implementations of numerical solutions to the inverse first passage problem based on equations such as (1.6); see [2].

We point out that solutions to (1.1) are not smooth, so that a weaker concept of solutions has to be used. Instead of using the classical weak solution defined in the distributional sense (see Evans [4]), we use viscosity solutions, introduced by Crandall and Lions in 1981. In this paper, we shall prove the following theorem.

THEOREM 1. *Problem* (1.1) *is a well-posed problem; i.e., for any given* $p(t)$ *satisfying* (1.2), *there is a unique (viscosity) solution.*

The remainder of the paper is organized as follows. In section 2, we provide a definition of the viscosity solution to (1.1) and show there is at most one such solution. In section 3, we establish the existence of a viscosity solution. This is accomplished by first defining and studying a regularization of the problem obtained by penalizing the obstacle (a standard procedure for the obstacle problem; see Friedman [5]). The ε-regularization is carefully designed so that the solution is monotonic in ε, and therefore the existence of a limit as $\varepsilon \to 0$ is automatically guaranteed.[4] We show that the limit is a viscosity solution. In section 4, we study the asymptotic behavior of the boundary as $t \searrow 0$ by providing explicit upper and lower bounds. When $\limsup_{t\searrow 0} -\frac{1-p(t)}{t\dot{p}(t)} < \infty$, we prove that

$$\lim_{t\to 0} \frac{b(t)}{\sqrt{-2t\log(1-p(t))}} = -1.$$

In section 5, we derive integral equations for b when $\sigma \equiv 1$ and $\mu \equiv 0$ under the assumption that p is continuous and nonincreasing.

2. Viscosity solutions and uniqueness. By noticing that $w(x,t) < p(t)$ for all $x > b(t)$ when $\tau > t$, we can state the inverse first passage problem as follows. Find an unknown function $w = w(x,t)$ such that

$$\text{(2.1)} \qquad \begin{cases} \mathcal{L}w = 0 & \text{when } w(\cdot,t) < p(t), \\ 0 \le w(x,t) \le p(t) & \text{for any } (x,t) \in (\mathbb{R}\times(0,\infty)), \\ w(x,0) = \mathbf{1}_{(-\infty,0)}(x) & \text{for } x \in \mathbb{R}, \end{cases}$$

where $\mathcal{L}w := w_t - \frac{1}{2}(\sigma^2 w_x)_x + \mu w_x$. Define the free boundary as

$$b^w(t) := \inf\{x \,|\, w(x,t) < p(t)\}.$$

We can write (2.1) as a variational inequality:

$$\text{(2.2)} \qquad \begin{cases} \max\{\mathcal{L}w, w-p\} = 0 & \text{in } \mathbb{R}\times(0,\infty), \\ w(\cdot,0) = \mathbf{1}_{(-\infty,0)}(\cdot) & \text{on } \ \mathbb{R}\times\{0\}. \end{cases}$$

In this paper, we shall consider viscosity solutions to this variational inequality, as defined below (see also Crandall, Ishii, and Lions [3]). The principle motivation for doing so is to allow the maximum generality in specifying p. Of course, one could consider other notions of weak solution with corresponding assumptions on the problem data. The basic strategy of using penalization to construct approximations and analyze the properties of the solution would still be applicable (see, e.g., Friedman [5]).

For a given p, we define

$$p_*(t) = \liminf_{0\le s\to t} p(s), \quad p^*(t) = \limsup_{0\le s\to t} p(s) \quad \forall t \ge 0.$$

[4]For purely proving existence, such careful design of the sequence is unnecessary. We introduce the complication with a view towards applications in numerical simulations.

Since cumulative probability distribution functions (hence q) are increasing and right continuous, we see that for any b, $p(t) = \mathbb{P}\{\tau > t\} \geq 0$ is decreasing and right continuous, and in particular, $p = p_*$. Furthermore, Blumenthal's zero-one law (see, for example, [13]) implies that we must have either $p(0) = 0$ (in which case the problem is trivial) or $p(0) = 1$. Therefore, in the remainder of the paper, we shall consider only lower semicontinuous p for which $p(0) = 1$.

For a function w defined on $\mathbb{R} \times [0, \infty)$, we define w^* and w_* by

$$w^*(x,t) := \limsup_{y \to x, 0 \leq s \to t} w(y,s) \qquad \forall (x,t) \in \mathbb{R} \times [0,\infty),$$
$$w_*(x,t) := \liminf_{y \to x, 0 \leq s \to t} w(y,s) \qquad \forall (x,t) \in \mathbb{R} \times [0,\infty).$$

A function w is called *upper semicontinuous (USC)* if $w = w^*$ and *lower semicontinuous (LSC)* if $w = w_*$.

In what follows, the parabolic open ball $B_\delta(x,t)$ is defined as

$$B_\delta(x,t) := (x-\delta, x+\delta) \times (t-\delta^2, t) \qquad \forall \delta > 0, (x,t) \in \mathbb{R} \times [0,\infty).$$

For any cylindrical set of the form $D = (s,t) \times \Omega$, where $0 \leq s < t$ and $\Omega \subseteq \mathbb{R}$, the parabolic boundary is defined as

$$\partial_p D = \Big(\partial\Omega \times (s,t)\Big) \cup \Big(\bar{\Omega} \times \{s\}\Big).$$

Definition 1 (viscosity subsolution, supersolution, and solution).

1. *A function w defined on $\mathbb{R} \times (0,\infty)$ is called a* (viscosity) subsolution *if*

$$w = \min\{p, w^*\} \quad \text{in} \quad \mathbb{R} \times (0,\infty),$$

and $\mathcal{L}\varphi(x,t) \leq 0$ whenever φ is smooth and $w^* - \varphi$ attains at (x,t) a local maximum on $\bar{B}_\delta(x,t)$, where $x \in \mathbb{R}$ and $t > \delta^2 > 0$.

2. *A function w defined on $\mathbb{R} \times (0,\infty)$ is called a* (viscosity) supersolution *if*

$$0 \leq w = w_* \quad \text{in} \quad \mathbb{R} \times (0,\infty),$$

and $\max\{w(x,t) - p(t), \mathcal{L}\varphi(x,t)\} \geq 0$ whenever φ is smooth and $w - \varphi$ attains at (x,t) a local minimum on $\bar{B}_\delta(x,t)$, where $x \in \mathbb{R}$ and $t > \delta^2 > 0$.

3. *A function w defined on $\mathbb{R} \times [0,\infty)$ is called a* (viscosity) solution *if w is both a subsolution and a supersolution in $\mathbb{R} \times (0,\infty)$, and for all $x \in \mathbb{R}$,*

$$w(x,0) = \liminf_{y \to x, t \searrow 0} w(y,t) = \mathbf{1}_{(-\infty,0)}, \qquad \limsup_{y \to x, t \searrow 0} w(y,t) = \mathbf{1}_{(-\infty,0]}. \tag{2.3}$$

Remark 2.1. Here we use the default that a viscosity solution is LSC ($w = w_*$). Also, the (probabilistically obvious) condition $w \geq 0$ imposed for supersolutions is to ensure the boundedness of the supersolution, as is usually required. This condition could be relaxed to the assumption that $w \geq -e^{A(1+|x|^2)}$ for some $A > 0$.

To prove the uniqueness of the solution to (2.2), we first establish a few properties of viscosity solutions.

Lemma 2.1. *Let w be a viscosity solution and define*

$$Q := \{(x,t) \in \mathbb{R} \times [0,\infty) \mid w(x,t) < p(t)\}, \qquad \Pi := Q^c = \mathbb{R} \times [0,\infty) \setminus Q.$$

Then

1. *Q is open and w is a smooth solution to $\mathcal{L}w = 0$ in Q;*
2. *$\Pi = \{(x,t) \in \mathbb{R} \times [0,\infty) \mid w(x,t) = p(t)\} = \Pi_0 \cup \Pi_1 \cup \Pi_2$, where*

$$\Pi_0 := \{(x,t) \in \mathbb{R} \times [0,\infty) \mid w^*(x,t) = w_*(x,t) = p(t)\},$$
$$\Pi_1 := \{(x,t) \in \mathbb{R} \times [0,\infty) \mid p^*(t) > w^*(x,t) > w_*(x,t) = p(t)\},$$
$$\Pi_2 := \{(x,t) \in \mathbb{R} \times [0,\infty) \mid p^*(t) = w^*(x,t) > w_*(x,t) = p(t)\}.$$

In particular, if p is continuous, then w is continuous in $\mathbb{R} \times [0,\infty) \setminus \{(0,0)\}$.

Proof. 1. First, we show that Q is open and w is continuous in Q. For any $(x,t) \in Q$, $w(x,t) < p(t)$. As a supersolution, $w(x,t) = w_*(x,t)$, and as a subsolution, $w(x,t) = \min\{p(t), w^*(x,t)\} = w^*(x,t)$. Then $w_* = w = w^*$ at (x,t). Therefore w is continuous at (x,t) and $w < p$ in a neighborhood of (x,t). Consequently, Q is open and w is continuous in Q.

Next, we prove $\mathcal{L}w = 0$ in Q. Let $(x_0,t_0) \in Q$ with $t_0 > 0$. Then $w(x_0,t_0) < p(t_0)$ and w is continuous at (x_0,t_0). There exist positive constants η and δ such that $w < p - \eta$ in $\bar{D} \subset Q$, where $D = (x_0-\delta, x_0+\delta) \times (t_0-\delta^2, t_0+\delta^2)$. Denote by $\tilde{w}$ the solution to $\mathcal{L}\tilde{w} = 0$ in D with continuous initial data $\tilde{w}(\cdot, t_0-\delta^2) = w(\cdot, t_0-\delta^2)$ and boundary data $\tilde{w} = w$ on the parabolic boundary $\partial_p D$. Let $\varepsilon > 0$ and $\varphi^\varepsilon = \tilde{w} - \frac{\varepsilon}{t_0+\delta^2-t}$, $\psi^\varepsilon = \tilde{w} + \frac{\varepsilon}{t_0+\delta^2-t}$. We have that $w - \varphi^\varepsilon > 0$ on $\partial_p D$, and $w - \varphi^\varepsilon \to \infty$ as $t \nearrow t_0 + \delta^2$. Suppose there is a point $(x,t) \in D$ such that $w(x,t) - \varphi^\varepsilon(x,t) \le 0$; then $w - \varphi^\varepsilon$ will have a local minimum in D, say, at (x^*,t^*). Since w is a supersolution and $w(x^*,t^*) < p(t^*)$, we must have $-\frac{\varepsilon}{t_0+\delta^2-t^*} = \mathcal{L}\varphi^\varepsilon \ge 0$, which is a contradiction. Thus $w > \varphi^\varepsilon$. A similar argument gives that $w < \psi^\varepsilon$. Sending $\varepsilon \to 0$ we obtain $w = \tilde{w}$ in D, which implies that w is a smooth solution to $\mathcal{L}w = 0$ in Q.

2. Since $w \le p$, $\Pi := Q^c$. As a subsolution $w = \min\{p, w^*\} \le p$, and as a supersolution, $w = w_*$. It follows that $w_* = w = p \le w^*$ and $w^* \le p^*$ in Π. Hence $w_* = p \le w^* \le p^*$. There are only three possibilities for w^*: (i) $w^* = p$, (ii) $w^* \in (p, p^*)$, and (iii) $w^* = p^* > p$. Thus $\Pi = \Pi_0 \cup \Pi_1 \cup \Pi_2$. ∎

The following lemma characterizes the discontinuities of a solution.

LEMMA 2.2. *Suppose w is a viscosity solution. Then for each $t > 0$, the following hold:*

1. *$w(\cdot,t) = w_*(\cdot,t)$ is continuous in $\mathbb{R}$;*
2. *for each $x \in \mathbb{R}$,*

$$w_*(x,t) = \min\{p(t), w^*(x,t)\} = \lim_{y \to x, s \searrow t} w(y,s), \tag{2.4}$$
$$w^*(x,t) = \lim_{y \to x, s \nearrow t} w(y,s) \le p^*(t); \tag{2.5}$$

3. *if $w^*(x,t) < p^*(t)$, then for some $\delta > 0$, $w = w^*$ in $B_\delta(x,t)$ and w^* is a smooth solution to $\mathcal{L}w^* = 0$ in $\bar{B}_\delta(x,t)$.*

Proof. 1. If $w(x,t) < p(t)$, by Lemma 2.1 w is continuous near (x,t); otherwise $w(x,t) = p(t)$. Then using the subsolution property, $\liminf_{y \to x} w(y,t) \ge w_*(x,t) = w(x,t) = p(t) \ge \limsup_{y \to x} w(y,t)$. Thus $\liminf_{y \to x} w(y,t) = w(x,t) = \limsup_{y \to x} w(y,t)$, and the first assertion follows.

2. Next, we prove (2.4). The first equality is immediate since w is both a subsolution and a supersolution. The second follows by considering separately the cases $w(x,t) < p(t)$ and $w(x,t) = p(t)$ as in the previous step. If $w(x,t) < p(t)$, then w is continuous near (x,t). Therefore $w_*(x,t) = \min\{p(t), w^*(x,t)\} = \lim_{y \to x, s \searrow t} w(y,s)$. If $w(x,t) = p(t)$, then $w_*(x,t) = w(x,t) = p(t) \ge \lim_{y \to x, s \searrow t} w(y,s) \ge w_*(x,t)$. Thus (2.4) holds.

3. Now we prove (2.5) and the third assertion when $w^*(x,t) < p^*(t)$. By the upper semicontinuity of w^*, there exist some positive constants δ and η such that $w(\cdot,\cdot) < p^*(t) - \eta$ in $\bar{B}_\delta(x,t)$. For all $s < t$, $p(s) \geq p^*(t)$. Following the same proof as that for the previous lemma, we conclude that $w^* = w$ in $B_\delta(x,t)$ and w^* is a smooth solution to $\mathcal{L}w^* = 0$ in $\bar{B}_\delta(x,t)$. The third assertion and (2.5) for the case $w^*(x,t) < p^*(t)$ thus follow.

4. Finally, we verify (2.5) for the case $w^*(x,t) = p^*(t)$. For each small $\delta > 0$, we compare w in $B_\delta(x,t)$ with solutions $\bar{w}$ and $\underline{w}$ to

$$\begin{cases} \mathcal{L}\bar{w} = 0 & \text{in } B_\delta, \\ \bar{w} = w^* & \text{on } \partial_p B_\delta \end{cases} \quad \text{and} \quad \begin{cases} \mathcal{L}\underline{w} = 0 & \text{in } B_\delta, \\ \underline{w} = \min\{w^*, p^*(t)\} & \text{on } \partial_p B_\delta, \end{cases}$$

respectively. Note that $w \leq w^* = \bar{w}$ and $\underline{w} = \min\{w^*, p^*(t)\} \leq \min\{w^*, p\} = w$ on $\partial_p B_\delta$ since $\underline{w} \leq p^*(t) \leq p(s)$ for all $s < t$. By the maximum principle, $\underline{w} \leq w \leq \bar{w}$ in B_δ. Observing that $\max_{\bar{B}_\delta}\{\bar{w} - \underline{w}\} = \max_{\partial_p B_\delta}\{\bar{w} - \underline{w}\} \leq p^*(t - \delta^2) - p^*(t)$, we find that

$$\limsup_{y \to x, s \nearrow t} w(y,s) - \liminf_{y \to x, s \nearrow t} w(y,s) \leq \sup_{B_\delta(x,t)} \{\bar{w} - \underline{w}\} \leq p^*(t - \delta^2) - p^*(t).$$

Equation (2.5) then follows by sending $\delta \to 0$. ∎

THEOREM 2 (uniqueness). *There is at most one viscosity solution.*

Proof. Suppose w_1 and w_2 are two solutions. We claim that for any $\eta > 0$,

$$w_1(x,t) \leq w_2(x - \eta, t) \qquad \forall (x,t) \in \mathbb{R} \times [0,\infty). \tag{2.6}$$

Once this is proven, sending $\eta \to 0$ and using Lemma 2.2(1) allows one to conclude that $w_1 \leq w_2$ on $\mathbb{R} \times [0,\infty)$. Exchanging the roles of w_1 and w_2, we also have $w_2 \leq w_1$, so that $w_1 \equiv w_2$.

Suppose that (2.6) is not true; i.e., there exists at least one pair of $(\bar{x}, \bar{t}\,) \in \mathbb{R} \times [0,\infty)$ such that $w_1(\bar{x}, \bar{t}) > w_2(\bar{x} - \eta, \bar{t})$. Then for all sufficiently small positive ε,

$$M_\varepsilon := \sup_{x \in \mathbb{R}, t \geq 0} \{w_1(x,t) - w_2(x - \eta, t) - \varepsilon^4 x^2 - \varepsilon e^t\} > 0.$$

Hence fix such a positive ε such that $\varepsilon \leq 1/(1 + \|\sigma^2\|_\infty + 2\|\sigma\sigma_x - \mu\|_\infty)$. Let $\{(x_n, t_n)\}_{n=1}^\infty$ be a sequence in $\mathbb{R} \times [0,\infty)$ such that the supremum M_ε is attained along the sequence. This sequence is bounded since $0 \leq w_1, w_2 \leq 1$. By taking a subsequence if necessary, there exist the limits $(\hat{x}, \hat{t}\,) := \lim_{n\to\infty}(x_n, t_n)$, $\alpha := \lim_{n\to\infty} w_1(x_n, t_n)$, and $\beta := \lim_{n\to\infty} w_2(x_n - \eta, t_n)$. Note that

$$\begin{gathered} 0 \leq w_{1*}(\hat{x}, \hat{t}\,) \leq \alpha \leq w_1^*(\hat{x}, \hat{t}\,) \leq 1, \\ 0 \leq w_{2*}(\hat{x} - \eta, \hat{t}\,) \leq \beta \leq w_2^*(\hat{x} - \eta, \hat{t}\,) \leq 1, \\ \beta < \alpha. \end{gathered}$$

Consequently,

$$M_\varepsilon = \alpha - \beta - \varepsilon^4 \hat{x}^2 - \varepsilon e^{\hat{t}} > 0. \tag{2.7}$$

Additionally, $|\hat{x}| < \varepsilon^{-2}$. Otherwise $M_\varepsilon \leq \alpha - \beta - 1 - \varepsilon \leq -\varepsilon < 0$, which is a contradiction to (2.7).

Now we show that this is impossible by excluding the following three possibilities:

$$\text{(i)}\ \ \hat{t} = 0; \qquad \text{(ii)}\ \ \hat{t} > 0, \beta < p(\hat{t}); \qquad \text{(iii)}\ \ \hat{t} > 0, \beta \geq p(\hat{t}).$$

Case (i). Suppose $\hat{t} = 0$. If $\hat{x} \geq \eta$, then $0 \leq \beta < \alpha \leq w_1^*(\hat{x}, 0) = 0$; otherwise $1 \geq \alpha > \beta \geq w_{2*}(\hat{x} - \eta, 0) = 1$. Both are impossible.

Case (ii). Suppose $\hat{t} > 0$ and $\beta < p(\hat{t})$. Then $w_{2*}(\hat{x} - \eta, \hat{t}) \leq \beta < p(\hat{t})$. By Lemma 2.1, w_2 is a smooth solution to $\mathcal{L}w_2(\cdot - \eta, \cdot) = 0$ in $\bar{D}$, where $D = (\hat{x} - \delta, \hat{x} + \delta) \times (\hat{t} - \delta^2, \hat{t} + \delta^2)$ for some $\delta > 0$. Let

$$\varphi(x,t) = w_2(x - \eta, t) + \varepsilon^4 x^2 + \varepsilon e^t + (x - \hat{x})^4/\delta^4 + (t - \hat{t})^2/\delta^4.$$

Then φ is smooth in $\bar{D}$ and

$$\max_D \{w_1^* - \varphi\} = \sup_{\bar{D}} \{w_1 - \varphi\} \leq M_\varepsilon \leq w_1^*(\hat{x}, \hat{t}) - \varphi(\hat{x}, \hat{t}).$$

That is, $(\hat{x}, \hat{t})$ is a local maximum of $w_1^* - \varphi$ in D. As w_1 is a subsolution, $\mathcal{L}\varphi(\hat{x}, \hat{t}) \leq 0$. However, $\mathcal{L}\varphi(\hat{x}, \hat{t}) = \varepsilon e^{\hat{t}} - \varepsilon^4\sigma^2 - 2(\sigma\sigma_x - 2\mu)\varepsilon^4\hat{x} \geq \varepsilon - \varepsilon^2\|\sigma^2\|_\infty - 2\varepsilon^2\|\sigma\sigma_x - \mu\|_\infty > 0$ by the smallness of ε. This is a contradiction. Thus case (ii) is impossible.

Case (iii). Suppose $\hat{t} > 0$ and $\beta \geq p(\hat{t})$. Since $p^*(s) \leq p(t)$ for any $s > \hat{t}$, $w_1(x, s) \leq p(s) \leq p^*(t)$ for each x. Then $\sup_x w_1(x, s) \leq p(\hat{t}) \leq \beta < \alpha = \lim_{n\to\infty} w_1(x_n, t_n)$. We claim $t_n < \hat{t}$ for all sufficiently large n; i.e., there exists $N \in \mathbb{N}^+$ such that $t_n < \hat{t}$ for each $n \geq N$. To the contrary, suppose for each $N \in \mathbb{N}^+$ there exists $n > N$ such that $t_n > \hat{t}$. Then $\alpha > \beta \geq \sup_x w_1(x, t_n) > w_1(x_n, t_n)$. Hence there exists $\varepsilon > 0$, independent on n, such that $\alpha > w_1(x_n, t_n) + \varepsilon$. This is a contradiction to $\alpha = \lim_{n\to\infty} w_1(x_n, t_n)$. Consequently, from (2.5), we conclude that

$$\alpha = w_1^*(\hat{x}, \hat{t}) \leq p^*(\hat{t}), \qquad \beta = w_2^*(\hat{x} - \eta, \hat{t}) \leq \alpha < p^*(\hat{t}).$$

By Lemma 2.2(3), for some $\delta > 0$, $w_2^* = w_2$ in $B_\delta(\hat{x} + \eta, \hat{t})$ and w_2^* is a smooth solution to $\mathcal{L}w_2^* = 0$ in $\bar{B}_\delta(\hat{x} + \eta, \hat{t})$. Set

$$\phi(x,t) := w_2^*(x - \eta, t) + \varepsilon^4 x^2 + \varepsilon e^t + (x - \hat{x})^4/\delta^4 + (t - \hat{t})^2/\delta^4.$$

Then, by (2.5) and $w_2^* = w_2$ in $B_\delta(\hat{x} + \eta, \hat{t})$,

$$\max_{\bar{B}_\delta(\hat{x},\hat{t})} \{w_1^* - \phi\} = \sup_{B_\delta(\hat{x},\hat{t})} \{w_1 - \phi\} \leq M_\varepsilon = w_1^*(\hat{x}, \hat{t}) - \phi(\hat{x}, \hat{t}).$$

That is, $w_1^* - \phi$ obtains its local maximum at $(\hat{x}, \hat{t})$. As w_1 is a subsolution, $\mathcal{L}\phi(\hat{x}, \hat{t}) \leq 0$. However, $\mathcal{L}\phi(\hat{x}, \hat{t})\varepsilon e^{\hat{t}} - \varepsilon^4\sigma^2 - 2(\sigma\sigma_x - 2\mu)\varepsilon^4\hat{x} \geq \varepsilon - \varepsilon^2\|\sigma^2\|_\infty - 2\varepsilon^2\|\sigma\sigma_x - \mu\|_\infty > 0$ by the smallness of ε. This is a contradiction. Thus case (iii) is impossible.

The exclusion of cases (i), (ii), and (iii) implies that (2.6) holds for each $\eta > 0$. □

As a product, (2.6) and the uniqueness give the following.

COROLLARY 2.3. *The unique solution w, if it exists, is nonincreasing in x, i.e., $w(x,t) \leq w(x - \eta, t)$ for all $\eta > 0$ and $(x,t) \in \mathbb{R} \times [0, \infty)$.*

3. Existence of a viscosity solution. To establish a solution, we first define and study a regularization of the problem obtained by penalizing the obstacle. This ε-regularization is carefully designed so that the solution is monotonic in ε, and therefore the existence of a limit as $\varepsilon \to 0$ is automatically guaranteed. We prove some regularity properties of the solution to the penalized problem for the purpose of establishing compactness. Finally, we show that the limit is a viscosity solution.

3.1. The regularization. Following the classical penalization technique (see, for example, Friedman [5]) for variational inequalities, we consider a semilinear parabolic equation:

$$\begin{cases} \mathcal{L}w^\varepsilon = -\beta\Big(\varepsilon^{-1}(w^\varepsilon - p^\varepsilon)\Big) & \text{in } \mathbb{R}\times(0,\infty), \\ w^\varepsilon(\cdot,0) = W^\varepsilon(\cdot) & \text{on } \mathbb{R}\times\{0\}, \end{cases} \tag{3.1}$$

where p^ε and W^ε are the smooth approximations of p and $w(\cdot,0) = \mathbf{1}_{(-\infty,0)}$, respectively, and $\beta(\cdot)$ is a smooth function being identically zero in $(-\infty,0]$ and strictly increasing and convex in $[0,\infty)$. For definiteness, we take

$$\beta(s) := \max\{0, s^3\} \qquad \forall\, s\in\mathbb{R}.$$

The particular p^ε and W^ε are chosen so that the solution w^ε is strictly increasing in ε. Define

$$p^\varepsilon(t) := \frac{3}{4}\int_{-1}^{1}(1-z^2)p(t+\varepsilon+\varepsilon z)\,dz - 3\varepsilon^{2/3} \qquad \forall\varepsilon>0, t\ge 0.$$

Then $p^\varepsilon\in C^1([0,\infty))$, and

$$-\frac{1}{\varepsilon} \le \frac{d}{dt}p^\varepsilon(t) \le 0, \quad \frac{d}{d\varepsilon}p^\varepsilon(t) \le -\frac{2}{\varepsilon^{1/3}}, \quad \lim_{\varepsilon\searrow 0} p^\varepsilon(t) = p(t) \quad \forall\varepsilon>0, t\ge 0. \tag{3.2}$$

The first two inequalities follow directly from the definition of p^ε and monotonicity of p. Consequently, for any fixed $t>0$, the limit as $\varepsilon\searrow 0$ of $p^\varepsilon(t)$ exists and is $p(t)$.

When $t=0$, (3.2) yields $\lim_{\varepsilon\searrow 0} p^\varepsilon(0) = p(0) = 1$ and $p^\varepsilon(0)$ is a monotone function of ε. We denote by $\varepsilon^*>0$ the unique constant such that $p^{\varepsilon^*}(0) = 0$ and in what follows assume $\varepsilon\in(0,\varepsilon^*)$.

Fix a smooth function $W(\cdot)$ defined on $\mathbb{R}$ that satisfies

$$W(x) = 0\ \forall\, x\ge 0, \quad W(x) = 1\ \forall\, x\le -1, \quad \dot W \le 0\ \forall\ x\in(-1,0).$$

Set

$$W^\varepsilon(x) := p^\varepsilon(0)\ W(x/\varepsilon) \qquad \forall\, x\in\mathbb{R}.$$

Then W^ε is a smooth function satisfying

$$\frac{d}{d\varepsilon}W^\varepsilon(x)\le 0, \quad \frac{d}{dx}W^\varepsilon(x)\le 0, \quad W^\varepsilon(x) = 0\ \forall\, x\ge 0, \quad W^\varepsilon(x) = p^\varepsilon(0)\ \forall\, x\le -\varepsilon.$$

Before proving the existence of a solution to problem (3.1), we introduce the following functions.

1. Denote by $w_0^\varepsilon(x,t)$ the solution to

$$\begin{cases} \mathcal{L}w_0^\varepsilon = 0 & \text{in } \mathbb{R}\times(0,\infty), \\ w_0^\varepsilon(\cdot,0) = W^\varepsilon(\cdot) & \text{on } \mathbb{R}\times\{0\}. \end{cases} \tag{3.3}$$

Since the problem for w_0^ε is linear, the solution w_0^ε can be expressed as

$$w_0^\varepsilon(x,t) = \int_{\mathbb{R}} K(x,t;y,0)w_0^\varepsilon(y,0)\,dy = p^\varepsilon(0)\int_{-\infty}^{0} K(x,t;y,0)W(y/\varepsilon)\,dy,$$

where $K(x,t;y,s)$ is the fundamental solution associated with the linear operator $\mathcal{L}$. In particular, when $\mathcal{L} = \partial_t - \frac{1}{2}\partial_{xx}$, i.e., $\mu \equiv 0$ and $\sigma \equiv 1$,

$$K(x,t;y,s) = \Gamma(x-y, t-s), \qquad \Gamma(x,t) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}.$$

2. Denote by ρ^ε the solution to

$$\begin{cases} \frac{d}{dt}\rho^\varepsilon(t) = -\beta\left(\frac{\rho^\varepsilon(t) - p^\varepsilon(t)}{\varepsilon}\right) & \text{in } (0,\infty), \\ \rho^\varepsilon(0) = p^\varepsilon(0). \end{cases}$$

Comparing the solution ρ^ε with the functions of the form $p^\varepsilon + C$, where C is constant, one finds that

$$p^\varepsilon(t) \le \rho^\varepsilon(t) \le p^\varepsilon(t) + \varepsilon \|\dot{p}^\varepsilon\|_\infty^{1/3}, \quad \dot{\rho}^\varepsilon(t) \le 0 \quad \forall t \ge 0.$$

Now we are ready to begin studying the properties of (3.1).

THEOREM 3. *For each $\varepsilon > 0$, problem (3.1) admits a unique smooth $(C^{2,1})$ solution in $\mathbb{R} \times [0,\infty)$. The solution is continuously differentiable in ε and satisfies, for all $\varepsilon > 0$ and $(x,t) \in \mathbb{R} \times (0,\infty)$,*

$$w_0^\varepsilon(x,t) + \rho^\varepsilon(t) - \rho^\varepsilon(0) \le w^\varepsilon(x,t) \le \min\{\rho^\varepsilon(t), w_0^\varepsilon(x,t)\}, \tag{3.4}$$

$$w_x^\varepsilon(x,t) < 0, \qquad \frac{d}{d\varepsilon} w^\varepsilon(x,t) < 0. \tag{3.5}$$

Consequently, the following limit exists:

$$w(x,t) := \lim_{\varepsilon \searrow 0} w^\varepsilon(x,t) \qquad \forall\, (x,t) \in \mathbb{R} \times [0,\infty).$$

Proof. 1. Existence and uniqueness of a smooth solution w^ε to (3.1) in $\mathbb{R} \times [0,\infty)$ follows from standard results; see, for example, Friedman [6]. The fact that w^ε satisfies (3.4) follows from the comparison principle. In particular, since

$$\mathcal{L} w_0^\varepsilon + \beta\left(\frac{w_0^\varepsilon - p^\varepsilon}{\varepsilon}\right) = \beta\left(\frac{w_0^\varepsilon - p^\varepsilon}{\varepsilon}\right) \ge 0 = \mathcal{L} w^\varepsilon + \beta\left(\frac{w^\varepsilon - p^\varepsilon}{\varepsilon}\right)$$

and $w_0^\varepsilon(x,0) = w^\varepsilon(x,0)$ we have $w^\varepsilon(x,t) \le w_0^\varepsilon(x,t)$. Similarly, $\mathcal{L}\rho^\varepsilon + \beta(\frac{\rho^\varepsilon - p^\varepsilon}{\varepsilon}) = 0$ and $\rho^\varepsilon(0) = p^\varepsilon(0) \ge \max_{x\in\mathbb{R}} w^\varepsilon(x,0)$, so that $w^\varepsilon(x,t) \le \rho^\varepsilon(x,t)$. Combining the two upper bounds gives the right half of (3.4).

The proof of the lower bound is similar. Set $\underline{w}^\varepsilon := w_0^\varepsilon(x,t) + \rho^\varepsilon(t) - \rho^\varepsilon(0)$. We can compute

$$\begin{aligned} &\mathcal{L}\underline{w}^\varepsilon + \beta\Big(\frac{\underline{w}^\varepsilon - p^\varepsilon}{\varepsilon}\Big) \\ &= \mathcal{L} w_0^\varepsilon(x,t) + \mathcal{L}\rho^\varepsilon(t) + \beta\Big(\frac{\underline{w}^\varepsilon - p^\varepsilon}{\varepsilon}\Big) \\ &= -\left[\beta\Big(\frac{\rho^\varepsilon - p^\varepsilon}{\varepsilon}\Big) - \beta\Big(\frac{\rho^\varepsilon - p^\varepsilon + w_0^\varepsilon - \rho^\varepsilon(0)}{\varepsilon}\Big)\right] \le 0, \end{aligned}$$

since $\beta(\cdot)$ is nondecreasing. We also have

$$w^\varepsilon(x,0) = w_0^\varepsilon(x,0) = \underline{w}^\varepsilon(x,0), \tag{3.6}$$

so the comparison principle implies $w^\varepsilon(x,t) \ge \underline{w}^\varepsilon(x,t)$ for all x,t.

2. The estimates (3.5) can be proved by differentiating the system (3.1). More specifically, to prove the first inequality in (3.5) differentiate (3.1) with respect to x and let $u^\varepsilon := -w^\varepsilon_x$ to obtain

$$\mathcal{A}u^\varepsilon + \frac{1}{\varepsilon}\dot\beta\Big(\frac{w^\varepsilon - p^\varepsilon}{\varepsilon}\Big)u^\varepsilon = 0,$$
$$u^\varepsilon(x,0) = -\frac{d}{dx}w^\varepsilon(x,0) = -\frac{d}{dx}W^\varepsilon(x) \geq 0 \qquad \forall x \in \mathbb{R},$$

where $\mathcal{A}u = \mathcal{L}u - \sigma\sigma_x u_x + (\mu_x - \sigma\sigma_{xx} + (\sigma_x)^2)u$. Since $\frac{1}{\varepsilon}\dot\beta(\frac{w^\varepsilon - p^\varepsilon}{\varepsilon}) > 0$, the maximum principle implies that $-w^\varepsilon_x(x,t) = u^\varepsilon(x,t) > 0$ in $\mathbb{R}\times(0,\infty)$.

For the second half of (3.5), differentiating the system (3.1) with respect to ε we obtain

$$\frac{d}{d\varepsilon}w^\varepsilon(x,0) = \frac{d}{d\varepsilon}W^\varepsilon(x) \leq 0 \qquad \forall x \in \mathbb{R},$$
$$\mathcal{L}\frac{d}{d\varepsilon}w^\varepsilon + \frac{1}{\varepsilon}\dot\beta\Big(\frac{w^\varepsilon - p^\varepsilon}{\varepsilon}\Big)\frac{d}{d\varepsilon}w^\varepsilon = \frac{1}{\varepsilon^2}\dot\beta\Big(\frac{w^\varepsilon - p^\varepsilon}{\varepsilon}\Big)\left\{w^\varepsilon - p^\varepsilon + \varepsilon\frac{d}{d\varepsilon}p^\varepsilon\right\} \leq 0,$$

since $\dot\beta \geq 0$, $w^\varepsilon - p^\varepsilon \leq \rho^\varepsilon - p^\varepsilon \leq \varepsilon\|\dot p^\varepsilon\|_\infty^{1/3} \leq \varepsilon^{2/3}$, and $\frac{d}{d\varepsilon}p^\varepsilon \leq -2\varepsilon^{-1/3}$. Then, by the maximum principle, $\frac{d}{d\varepsilon}w^\varepsilon < 0$ in $\mathbb{R}\times(0,\infty)$. The monotonicity and boundedness of w^ε in ε imply that $w = \lim_{\varepsilon\searrow 0} w^\varepsilon$ exists. ∎

Also note that since w^ε_0 is monotonic in ε and bounded, the limit $w_0 := \lim_{\varepsilon\searrow 0} w^\varepsilon_0$ exists and is the solution to

$$\mathcal{L}w_0 = 0 \ \text{ in } \ \mathbb{R}\times(0,\infty), \qquad w_0(\cdot,0) = \mathbf{1}_{(-\infty,0)}.$$

By Theorem 3, $w^\varepsilon_x < 0$ and $w^\varepsilon(x,t) \geq \rho^\varepsilon(t) - \rho^\varepsilon(0)$ (since $w^\varepsilon_0 \geq 0$). We conclude that $\lim_{x\to\infty} w^\varepsilon$ exists. Similarly, the limit $\lim_{x\to\infty} w^\varepsilon_0$ exists and is nonnegative.

3.2. Continuity estimates and existence. In the previous section, we proved that the limit $w = \lim_{\varepsilon\searrow 0} w^\varepsilon$ of solutions to the system (3.1) exists. In this section, we will prove that this limit gives a viscosity solution to our variational inequality. In order to do so, we need to derive some supplementary estimates on the continuity of w. We first control the behavior of w in the space variable, keeping the time constant.

LEMMA 3.1. *For each $T > 0$, there exists a constant $C_1 = C_1(T)$ that depends only on σ and μ such that for all $\varepsilon \in (0,\varepsilon^*)$, $0 < t \leq T$, and $x, y \in \mathbb{R}$,*

$$-\frac{C_1 p^\varepsilon(0)}{\sqrt{t}} \leq w^\varepsilon_{0x}(x,t) \leq w^\varepsilon_x(x,t) \leq 0. \tag{3.7}$$

Consequently, the limit $w = \lim_{\varepsilon\searrow 0} w^\varepsilon$ satisfies

$$-\frac{C_1}{\sqrt{t}} \leq w_{0x}(x,t) \leq w_x(x,t) \leq 0, \tag{3.8}$$

and we have

$$|w^\varepsilon(x,t) - w^\varepsilon(y,t)| \leq \frac{C_1 p^\varepsilon(0)}{\sqrt{t}}|x-y|, \qquad |w(x,t) - w(y,t)| \leq \frac{C_1}{\sqrt{t}}|x-y|. \tag{3.9}$$

Proof. We will prove (3.7). It is clear that (3.8) then follows by letting $\varepsilon \searrow 0$, and (3.9) can be derived from (3.7) and (3.8) by integration. The upper bounds

in (3.7) can be obtained by differentiating the systems (3.1) and (3.3) with respect to x and applying the maximum principle. Using the notation from the previous theorem, we find that

$$\begin{aligned} \mathcal{A}w_x^\varepsilon = -\varepsilon^{-1}\dot\beta(\varepsilon^{-1}(w^\varepsilon - p^\varepsilon)w_x^\varepsilon \geq 0 = \mathcal{A}w_{0x}^\varepsilon \qquad &\text{in } \ \mathbb{R}\times(0,\infty), \\ w_x^\varepsilon(\cdot,0) = W_x^\varepsilon(\cdot) = w_{0x}^\varepsilon(\cdot,0) \qquad &\text{on } \ \mathbb{R}\times\{0\}, \end{aligned}$$

since $w_x^\varepsilon \leq 0$. Therefore by the maximum principle (the zeroth order term in $\mathcal{A}$ is bounded above) $w_{0x}^\varepsilon \leq w_x^\varepsilon \leq 0$.

Next, we estimate the lower bound of w_{0x}^ε. Differentiating the system (3.3) with respect to x, we obtain

$$\mathcal{A}w_{0x}^\varepsilon = 0 \ \text{ in } \ \mathbb{R}\times(0,\infty), \qquad w_{0x}^\varepsilon(\cdot,0) = W_x^\varepsilon(\cdot) \ \text{ on } \ \mathbb{R}\times\{0\}.$$

This is a linear problem, and the solution can be expressed as

$$w_{0x}^\varepsilon(x,t) = \int_{\mathbb{R}} \tilde K(x,t;y,0)W_y^\varepsilon(y)\,dy,$$

where $\tilde K$ is the fundamental solution associated with the linear operator $\mathcal{A}$, and we therefore have $\tilde K > 0$. Also $W_y^\varepsilon \leq 0$, so for any $0 < t \leq T$,

$$0 \leq -w_{0x}^\varepsilon(x,t) \leq \sup_{x,y\in\mathbb{R}}\{\tilde K(x,t;y,0)\}\int_{\mathbb{R}} -W_y^\varepsilon(y)\,dy \leq \frac{C_1}{\sqrt{t}}\int_{\mathbb{R}} -W_y^\varepsilon(y)\,dy = \frac{C_1 p^\varepsilon(0)}{\sqrt{t}},$$

where

$$C_1 = C_1(T) = \sup_{x,y\in\mathbb{R},0<t<T}\Big\{\sqrt{t}\ \tilde K(x,t;y,0)\Big\},$$

and the above quantity is finite by the standard Gaussian upper bound on the fundamental solution $\tilde K$ (see Friedman [6]). This gives the lower bound in (3.7). ∎

We now proceed to estimate the variation in the time variable, leaving the space variable fixed.

LEMMA 3.2. *For each $T > 0$, there exists a constant $C_2 = C_2(T)$ that depends only on σ and μ such that for all $\varepsilon \in (0,\varepsilon^*)$, $0 < s < t \leq T$, and $x \in \mathbb{R}$,*

$$|w^\varepsilon(x,t) - w^\varepsilon(x,s)| \leq \frac{2C_2 p^\varepsilon(0)\sqrt{t-s}}{\sqrt{s}} + \rho^\varepsilon(s) - \rho^\varepsilon(t), \tag{3.10}$$

$$w^\varepsilon(x,t) - w^\varepsilon(x,s) \leq \frac{2C_2\sqrt{t-s}}{\sqrt{s}}. \tag{3.11}$$

Consequently, the limit $w = \lim_{\varepsilon\searrow 0} w^\varepsilon$ satisfies

$$|w(x,t) - w(x,s)| \leq \frac{2C_2\sqrt{t-s}}{\sqrt{s}} + \rho^\varepsilon(s) - \rho^\varepsilon(t), \tag{3.12}$$

$$w^\varepsilon(x,t) - w^\varepsilon(x,s) \leq \frac{2C_2\sqrt{t-s}}{\sqrt{s}}. \tag{3.13}$$

We remark that when $\sigma \equiv 1$ and $\mu \equiv 0$, $C_1 = C_2 = (2\pi)^{-1/2}$ for all T.

Proof. Once again, we need only worry about the estimates (3.10) and (3.11) for w^ε as the corresponding results for w follow by letting $\varepsilon \searrow 0$. The estimates are derived by first controlling the difference between $w(x,t)$ and the average of w on a small space interval centered at (x,t) and then estimating the difference of the values of such averages with respect to time and space. More specifically, for $\delta > 0$, $s \in (0,T]$, we have, using (3.7),

$$\left| w^\varepsilon(x,s) - \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} w^\varepsilon(y,s)\,dy \right| = \left| \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} \left(w^\varepsilon(x,s) - w^\varepsilon(y,s)\right)\,dy \right|$$
$$\leq \frac{C_1 p^\varepsilon(0)}{2\delta\sqrt{s}}\int_{x-\delta}^{x+\delta} |y-x|\,dy \leq \frac{C_1\delta p^\varepsilon(0)}{\sqrt{s}}.$$

We now consider the variation in time of the integrals of w over intervals of size 2δ. Note that since $\beta(\cdot)$ is increasing and $w^\varepsilon \leq \rho^\varepsilon$,

$$0 \leq \beta(\varepsilon^{-1}(w^\varepsilon - p^\varepsilon)) \leq \beta(\varepsilon^{-1}(\rho^\varepsilon - p^\varepsilon)) = -\dot\rho^\varepsilon(t) \qquad \forall\, t \geq 0, x \in \mathbb{R}.$$

And therefore, for $0 < s < t \leq T$ and $x \in \mathbb{R}$,

$$\left| \int_{x-\delta}^{x+\delta} \{w^\varepsilon(y,t) - w^\varepsilon(y,s)\}\,dy \right| = \left| \int_{x-\delta}^{x+\delta}\int_s^t w^\varepsilon_v(y,v)\,dv\,dy \right|$$
$$= \left| \int_s^t \int_{x-\delta}^{x+\delta} \left(\frac{1}{2}(\sigma^2 w^\varepsilon_y)_y - \mu w^\varepsilon_y - \beta\left(\varepsilon^{-1}(w^\varepsilon - p^\varepsilon)\right)\right) dy\,dv \right|$$
$$\leq \left| \int_s^t \frac{1}{2}\sigma^2 w^\varepsilon_y \Big|_{x-\delta}^{x+\delta} dv \right| + \left| \int_s^t \int_{x-\delta}^{x+\delta} \mu\, w^\varepsilon_y dy\,dv \right| + \left| \int_s^t \int_{x-\delta}^{x+\delta} \beta\left(\varepsilon^{-1}(\rho^\varepsilon - p^\varepsilon)\right) dy\,dv \right|$$
$$\leq p^\varepsilon(0)(t-s)\left(\frac{C_1\|\sigma^2\|_\infty}{\sqrt{s}} + \|\mu\|_\infty \right) + \left| \int_s^t \int_{x-\delta}^{x+\delta} -\frac{d}{dv}\rho^\varepsilon(v) dy\,dv \right|$$
$$\leq p^\varepsilon(0)(t-s)\left(\frac{C_1\|\sigma^2\|_\infty}{\sqrt{s}} + \|\mu\|_\infty \right) + 2\delta(\rho^\varepsilon(s) - \rho^\varepsilon(t)).$$

We can now estimate the continuity in the time variable:

$$|w^\varepsilon(x,t) - w^\varepsilon(x,s)|$$
$$\leq \left| w^\varepsilon(x,t) - \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} w^\varepsilon(y,t)\,dy \right| + \left| \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} \left(w^\varepsilon(y,t) - w^\varepsilon(y,s)\right)\,dy \right|$$
$$+ \left| \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} w^\varepsilon(y,s)\,dy - w^\varepsilon(x,s) \right|$$
$$\leq \frac{C_1 p^\varepsilon(0)}{\sqrt{s}}\left(2\delta + \frac{(t-s)\|\sigma^2\|_\infty}{2\delta} \right) + \frac{(t-s)2\|\mu\|_\infty p^\varepsilon(0)}{2\delta} + \rho^\varepsilon(s) - \rho^\varepsilon(t).$$

By taking $\delta = \frac{1}{2}\sqrt{\|\sigma^2\|_\infty(t-s)}$, we then obtain

$$|w^\varepsilon(x,t) - w^\varepsilon(x,s)| \leq \sqrt{\|\sigma^2\|_\infty(t-s)}\left(2\|w^\varepsilon_x\|^{s,t}_\infty + \frac{\|\mu\|_\infty p^\varepsilon(0)}{\|\sigma^2\|_\infty} \right) + \rho^\varepsilon(s) - \rho^\varepsilon(t)$$
$$\leq \frac{2C_2 p^\varepsilon(0)\sqrt{t-s}}{\sqrt{s}} + \rho^\varepsilon(s) - \rho^\varepsilon(t).$$

Observe that in estimating the upper bound of $w^\varepsilon(x,s) - w^\varepsilon(x,t)$, the term involving the integral of β can be dropped, so we have (3.11). ∎

It is now straightforward to derive the following estimates on the variation of w^ε and w.

LEMMA 3.3. *For each $T > 0$, there exists a constant $C = C(T)$ that depends only on σ and μ such that for all $\varepsilon \in (0, \varepsilon^*)$, $0 < t \le T$, and $x, y \in \mathbb{R}$,*

$$|w^\varepsilon(x,t) - w^\varepsilon(y,s)| \le \frac{Cp^\varepsilon(0)}{\sqrt{s}}\Big\{|x-y| + 2\sqrt{t-s}\Big\} + \rho^\varepsilon(s) - \rho^\varepsilon(t), \tag{3.14}$$

$$w^\varepsilon(x,t) - w^\varepsilon(y,s) \le \frac{Cp^\varepsilon(0)}{\sqrt{s}}\Big\{|x-y| + 2\sqrt{t-s}\Big\}. \tag{3.15}$$

Consequently, the limit $w = \lim_{\varepsilon \searrow 0} w^\varepsilon$ satisfies, for all $0 < s < t \le T$ and $x, y \in \mathbb{R}$,

$$|w(x,t) - w(y,s)| \le \frac{C}{\sqrt{s}}\Big\{|x-y| + 2\sqrt{t-s}\Big\} + p(s) - p(t), \tag{3.16}$$

$$w(x,t) - w(y,s) \le \frac{C}{\sqrt{s}}\Big\{|x-y| + 2\sqrt{t-s}\Big\}. \tag{3.17}$$

Proof. Using the previous two lemmas we have

$$\begin{aligned} &|w^\varepsilon(x,t) - w^\varepsilon(y,s)| \le |w^\varepsilon(x,t) - w^\varepsilon(x,s)| + |w^\varepsilon(x,s) - w^\varepsilon(y,s)| \\ &\le \frac{Cp^\varepsilon(0)}{\sqrt{s}}\left\{|x-y| + 2\sqrt{t-s}\right\} + \rho^\varepsilon(s) - \rho^\varepsilon(t). \end{aligned}$$

This proves (3.14). The proof of (3.15) is similar, and (3.16), (3.17) follow by sending $\varepsilon \searrow 0$. ∎

With the results from Lemma 3.3 we are now in a position to verify that w is indeed a viscosity solution to the variational inequality.

THEOREM 4. *Assume $p(\cdot)$ defined on $[0,\infty)$ is nonnegative, decreasing, and lower semicontinuous, with $p(0) = 1$. Then there is a unique viscosity solution, and it can be obtained as the limit $w := \lim_{\varepsilon \searrow 0} w^\varepsilon$.*

Proof. First, we verify that w satisfies the initial condition (2.3). For any $t > 0$, from (3.4) $|w^\varepsilon(\cdot,t) - w_0^\varepsilon(\cdot,t)| \le \rho^\varepsilon(0) - \rho^\varepsilon(t)$, so that

$$\|w(\cdot,t) - w_0(\cdot,t)\|_\infty \le \lim_{\varepsilon \searrow 0} \|w^\varepsilon(\cdot,t) - w_0^\varepsilon(\cdot,t)\|_\infty \ \le p(0) - p(t),$$

where

$$w_0(x,t) = \lim_{\varepsilon \to 0} w_0^\varepsilon(x,t) = \int_{-\infty}^0 K(x,t;y,0)\,dy,$$

$$\|w(\cdot,t) - w_0(\cdot,t)\|_\infty = \sup_{x \in \mathbb{R}} |w(\cdot,t) - w_0(\cdot,t)|.$$

Sending $t \searrow 0$, we see that w satisfies (2.3).

To verify that w is a viscosity solution in $\mathbb{R} \times (0,\infty)$ we consider two cases for each $(x,t) \in \mathbb{R} \times (0,\infty)$: (i) $p(t) - w(x,t) > 0$ and (ii) $p(t) - w(x,t) \le 0$.

Case (i). Suppose $p(t) - w(x,t) > 0$. Let $D_\delta = (x-\delta, x+\delta) \times (t-\delta^2, t+\delta^2)$ for all $\delta > 0$. Then, for each $(y,s) \in D_\delta$,

$$|\rho^\varepsilon(s) - \rho^\varepsilon(t)| \le |\rho^\varepsilon(s) - p^\varepsilon(s)| + |p^\varepsilon(s) - p^\varepsilon(t)| + |p^\varepsilon(t) - \rho^\varepsilon(t)| \le |p^\varepsilon(s) - p^\varepsilon(t)| + 2\varepsilon^{2/3},$$

since $0 \le \rho^\varepsilon - p^\varepsilon \le \varepsilon^{2/3}$. As $p^\varepsilon(\cdot)$ is decreasing, when $s > t$,

$$p^\varepsilon(t) - p^\varepsilon(s) + |p^\varepsilon(t) - p^\varepsilon(s)| = 2\left(p^\varepsilon(t) - p^\varepsilon(s)\right) \le 2\left(p^\varepsilon(t) - p^\varepsilon(t+\delta^2)\right);$$

and when $s \le t$,

$$p^\varepsilon(t) - p^\varepsilon(s) + |p^\varepsilon(t) - p^\varepsilon(s)| = 0 \le 2\left(p^\varepsilon(t) - p^\varepsilon(t+\delta^2)\right).$$

Using (3.14), we can compute

$$\begin{aligned} w^\varepsilon(y,s) - p^\varepsilon(s) &\le w^\varepsilon(x,t) + \tfrac{(2+2\sqrt{2})C\delta}{\sqrt{t-\delta^2}} + |\rho^\varepsilon(s) - \rho^\varepsilon(t)| - p^\varepsilon(s) \\ &\le \tfrac{(2+2\sqrt{2})C\delta}{\sqrt{t-\delta^2}} + w^\varepsilon(x,t) - p^\varepsilon(t) + p^\varepsilon(t) - p^\varepsilon(s) + |p^\varepsilon(t) - p^\varepsilon(s)| + 2\varepsilon^{2/3} \\ &\le \tfrac{(2+2\sqrt{2})C\delta}{\sqrt{t-\delta^2}} + w^\varepsilon(x,t) - p^\varepsilon(t) + 2(p^\varepsilon(t) - p^\varepsilon(t+\delta^2)) + 2\varepsilon^{2/3}. \end{aligned}$$

Then, if we take δ small enough,

$$\begin{aligned} \limsup_{\varepsilon\to 0} \max_{\bar{D}_\delta} \left\{w^\varepsilon - p^\varepsilon\right\} &\le \limsup_{\varepsilon\to 0} \tfrac{(2+2\sqrt{2})C\delta}{\sqrt{t-\delta^2}} \\ &\quad + w^\varepsilon(x,t) - p^\varepsilon(t) + 2(p^\varepsilon(t) - p^\varepsilon(t+\delta^2)) + 2\varepsilon^{2/3} \\ &\le \tfrac{(2+2\sqrt{2})C\delta}{\sqrt{t-\delta^2}} + w(x,t) - p(t) + 2(p(t) - p(t+\delta^2)) < 0. \end{aligned} \tag{3.18}$$

Thus, for all sufficiently small positive ε, $w^\varepsilon - p^\varepsilon < 0$ in $\bar{D}_\delta$. Consequently, $\mathcal{L}w^\varepsilon = -\beta(\frac{w^\varepsilon - p^\varepsilon}{\varepsilon}) = 0$ in $\bar{D}_\delta$. The limit w is then a smooth solution to $\mathcal{L}w = 0$ in D_δ.

Case (ii). Suppose $w(x,t) - p(t) \ge 0$. However, $w - p \le 0$ in $\mathbb{R}\times[0,\infty)$ since $w^\varepsilon \le \rho^\varepsilon$ and $\lim_{\varepsilon\searrow 0}\rho^\varepsilon(t) = p(t)$ in $\mathbb{R}\times[0,\infty)$. Hence, we must have $w(x,t) = p(t) = \min\{p(t), w^*(x,t)\}$. From (3.17)

$$w(x,t) - w_*(x,t) = \limsup_{y\to x, s\to t}(w(x,t) - w(y,s)) \le \limsup_{y\to x, s\to t}\left(\frac{C}{\sqrt{s}}\left\{|x-y| + 2\sqrt{t-s}\right\}\right) = 0.$$

Therefore $w_*(x,t) = w(x,t) = p(t)$. Thus the semicontinuity requirements for a viscosity solution hold.

In this case, we clearly have $\max\{w(x,t) - p(t), \mathcal{L}\varphi(x,t)\} \ge 0$ for *any* smooth φ. Therefore w is a supersolution. It remains to verify the differential inequality for subsolutions. To this end, let φ be a smooth function on $\bar{B}_\delta$, where $B_\delta = B_\delta(x,t)$ such that $w^*(y,s) - \varphi(y,s)$ attains at (x,t) a local maximum on $\bar{B}_\delta$. Set

$$\psi(y,s) = \varphi(y,s) + (y-x)^4/\delta^4 + (s-t)^2/\delta^4.$$

For each small positive ε, $w^\varepsilon - \psi$ attains a global maximum on $\bar{B}_\delta$. Denote any such point of maximum by $(y_\varepsilon, s_\varepsilon)$. Then $(w^\varepsilon - \psi)_s \ge 0$, $(w^\varepsilon - \psi)_{yy} \le 0$, $(w^\varepsilon - \psi)_y = 0$ at $(y_\varepsilon, s_\varepsilon)$. Thus, $\mathcal{L}\psi(y_\varepsilon, s_\varepsilon) \le \mathcal{L}w^\varepsilon(y_\varepsilon, s_\varepsilon) = -\beta(\varepsilon^{-1}(w^\varepsilon - p^\varepsilon)) \le 0$. If $(\bar{x},\bar{t})$ is a limit point of $\{(y_\varepsilon, s_\varepsilon)\}$ as $\varepsilon \to 0$, then $\mathcal{L}\psi(\bar{x},\bar{t}) \le 0$. Thus, it suffices to show that $(\bar{x},\bar{t}) = (x,t)$.

Since $w^\varepsilon \le w \le w^*$,

$$\begin{aligned} \limsup_{\varepsilon\searrow 0} \max_{\bar{B}_\delta}\left\{w^\varepsilon - \psi\right\} &\le \limsup_{\varepsilon\searrow 0}\left\{w^*(y_\varepsilon, s_\varepsilon) - \psi(y_\varepsilon, s_\varepsilon)\right\} \\ &\le w^*(\bar{x},\bar{t}) - \psi(\bar{x},\bar{t}) \le \max_{\bar{B}_\delta}(w^* - \varphi) - |\bar{x}-x|^4/\delta^4 - |\bar{t}-t|^2/\delta^4. \end{aligned}$$

On the other hand, from (3.16) and (3.17), we see that $w^*(\hat{x},t) = \lim_{s\nearrow t} w(\hat{x},s)$, so that

$$\begin{aligned}\limsup_{\varepsilon\searrow 0} \max_{\bar{B}_\delta}\Big\{w^\varepsilon - \psi\Big\} &\geq \lim_{s\nearrow t}\limsup_{\varepsilon\searrow 0}\{w^\varepsilon(x,s)-\psi(x,s)\}\\ &= \lim_{s\nearrow t}\{w(x,s)-\psi(x,s)\} = w^*(x,t)-\psi(x,t) = \max_{\bar{B}_\delta}(w^*-\varphi).\end{aligned}$$

Thus, we must have $(\bar{x},\bar{t}) = (x,t)$. This completes the proof. □

3.3. The differential equation and the free boundary problem. Since

$$0 \leq \beta(\varepsilon^{-1}(w^\varepsilon - p^\varepsilon)) \leq -\dot{\rho}^\varepsilon,$$

and $\rho^\varepsilon(\cdot)$ is decreasing, by weak compactness of measures, as $\varepsilon \to 0$,

$$\begin{aligned}\beta(\varepsilon^{-1}(w^\varepsilon - p^\varepsilon)) &\longrightarrow \gamma \qquad \text{as a measure in } \mathbb{R}\times[0,\infty),\\ \mathcal{L}w &= \gamma \qquad \text{on } \mathbb{R}\times(0,\infty),\end{aligned}$$

where γ is a Radon measure satisfying

$$0 \leq \gamma\, dx\, dt \leq -dx\, dp(t).$$

In addition, from Lemma 2.1(1), γ is supported on the set $w = p$.

Now suppose that p is continuous. Then $\gamma = \dot{p}$ on the contact set Π (noticing that Π_2 is empty). Hence, w is the solution to

$$\mathcal{L}w = \dot{p}(t)\mathbf{1}_{\{w=p\}} \text{ in } \mathbb{R}\times(0,\infty), \qquad w(\cdot,0) = \mathbf{1}_{(-\infty,0)} \text{ on } \mathbb{R}\times\{0\}. \tag{3.19}$$

Using a free boundary approach, this can be written as the solution to the free boundary problem for (b,w):

$$\begin{cases} \mathcal{L}w = \dot{p}(t)\mathbf{1}_{x<b(t)} & \text{in } \mathbb{R}\times(0,\infty),\\ b(t) := \inf\{x \mid w(x,t) < p(t)\} & \forall\ t \geq 0,\\ w(\cdot,0) = \mathbf{1}_{(-\infty,0)} & \text{on } \mathbb{R}\times\{0\}.\end{cases} \tag{3.20}$$

We emphasize that this formulation works only when p is continuous, since if p is not continuous at s, then

$$\mathcal{L}w = \min\{p(s) - w^*(x,\tau), 0\}\cdot\delta(t-s) \quad \text{on } \mathbb{R}\times\{s\},$$

where δ is the Dirac measure.

Remark 3.1. Suppose $\|\dot{p}\|_\infty := \sup_{t\geq 0}|\dot{p}(t)|$ is finite. Then $\|\dot{p}_\varepsilon\|_\infty \leq \|\dot{p}\|_\infty$ and $\rho^\varepsilon - p^\varepsilon \leq \varepsilon\|\dot{p}\|^{1/3}$. Consequently, $\dot{\rho}^\varepsilon = -\beta(\|\dot{p}\|^{1/3}) = -\|\dot{p}\|_\infty$. Hence

$$0 \leq \gamma^\varepsilon(x,t) \leq \|\dot{p}\|_\infty \qquad \forall (x,t)\in\mathbb{R}\times[0,\infty).$$

It is then easy to show that $w^\varepsilon(x,t) - w_0^\varepsilon \to w - w_0$ in $W_r^{2,1}([-R,R]\times[0,R^2])$ for any $r > 1$ and any $R > 0$.

4. Estimation of the free boundary. In this section, we provide both upper and lower bounds for the free boundary

$$b(t) := \inf\{x \in \mathbb{R} \mid w(x,t) < p(t)\} \in [-\infty, \infty] \qquad \forall\, t > 0$$

in the case of Brownian motion, i.e., when $\sigma \equiv 1$ and $\mu \equiv 0$.

Recall the notation $q(t) = 1 - p(t)$. Note that for any $s > 0$, $0 = q(0) = q^*(0) \leq q(s)$, and since p is LSC, q is USC. We define

$$\dot{q}(s) := \liminf_{t \nearrow s} \frac{q(s) - q(t)}{s - t} \in [0, \infty].$$

The following lemma is obvious from the probabilistic interpretation of our problem since it states that $\mathbb{P}[X_t \leq b(t)] \leq \mathbb{P}[\tau \leq t]$. Its analytic derivation is equally simple.

LEMMA 4.1. *For every $t > 0$,*

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{b(t)/\sqrt{2t}} e^{-z^2}\, dz \leq q(t).$$

Proof. We need only consider the case $b(t) > -\infty$. Since $w(x,t) \leq w_0(x,t)$,

$$\begin{aligned} 1 - q(t) = p(t) = w(b(t),t) \leq w_0(b(t),t) &= \int_{-\infty}^{0} \Gamma(b(t) - y, t)\, dy \\ &= 1 - \frac{1}{\sqrt{2\pi t}} \int_0^{\infty} e^{-\frac{(b(t)-y)^2}{2t}}\, dy = 1 - \frac{1}{\sqrt{\pi}} \int_{-\infty}^{b(t)/\sqrt{2t}} e^{-z^2} dz. \qquad \square \end{aligned}$$

We observe that the above lemma immediately yields an upper bound on $b(t)$:

$$b(t) \leq \sqrt{t} \cdot \Phi^{-1}(q(t)),$$

where $\Phi(z)$ is the cumulative distribution function of a standard normal random variable, $\Phi(z) = \int_{-\infty}^{z} e^{-w^2/2} \frac{dw}{\sqrt{2\pi}}$. We now proceed to the more difficult task of deriving useful lower bounds on $b(t)$.

LEMMA 4.2 (method for lower bounds). *Assume that $\underline{w}$ defined on $\mathbb{R} \times [0,t]$ satisfies*

$$\text{(4.1)} \qquad \begin{cases} \mathcal{L}\underline{w} = 0 & \text{in } \ \mathbb{R} \times (0,t], \\ \underline{w}(\cdot,0) \leq w(\cdot,0) & \text{on } \ \mathbb{R} \times \{0\}, \\ \underline{w} \leq p & \text{on } \ \mathbb{R} \times (0,t), \\ \underline{w}(s,t) \geq p(t) & \text{at } \ (s,t). \end{cases}$$

Then

$$\underline{w} \leq w \ \ \text{in } \ \mathbb{R} \times [0,t), \qquad s \leq b(t).$$

Proof. First, consider the case where p is continuous at t, so $p(t) = p^*(t)$. For each $\varepsilon > 0$, let $\phi^\varepsilon = \underline{w} - \varepsilon e^r - \varepsilon x^2$. We claim that $\phi^\varepsilon \leq w$ on $\mathbb{R} \times [0,t]$. Suppose not; then $w - \phi^\varepsilon$ can attain a global negative minimum, say, at $(\hat{x}, \hat{r})$. Since $w(x,0) - \phi^\varepsilon(x,0) = w(x,0) - \underline{w}(x,0) + \varepsilon + \varepsilon x^2 \geq \varepsilon$, $\hat{r} > 0$. Therefore $\mathcal{L}\phi^\varepsilon(\hat{x},\hat{r}) = -\varepsilon e^{\hat{r}} + \varepsilon < 0$. As a

supersolution, $\max\{w(\hat{x},\hat{r})-p(\hat{r}), \mathcal{L}\phi^\varepsilon(\hat{x},\hat{r})\} \geq 0$; hence we must have $w(\hat{x},\hat{r})-p(\hat{r}) \geq 0$. The condition $\underline{w} \leq p$ on $\mathbb{R}\times(0,t)$, and $p^*(t) = p(t)$ implies that $\underline{w} \leq p$ on $\mathbb{R}\times(0,t]$. Then $w(\hat{x},\hat{r}) < \phi^\varepsilon(\hat{x},\hat{r}) < \underline{w}(\hat{x},\hat{r}) \leq p(\hat{r})$. This is a contradiction. Thus $\phi^\varepsilon \leq w$ in $\mathbb{R}\times(0,t]$ for each $\varepsilon > 0$. Sending $\varepsilon \searrow 0$, we conclude that $\underline{w} \leq w$ in $\mathbb{R}\times(0,t]$.

In general, let $\{t_n\}$ be a sequence of positive numbers such that $t_n \nearrow t$ as $n \to \infty$, and $p(\cdot)$ is continuous at t_n. Then $\underline{w} \leq w$ in $\mathbb{R}\times[0,t_n]$. Sending $n \to \infty$ we obtain $\underline{w} \leq w$ in $\mathbb{R}\times[0,t)$.

As a subsolution, $w = \min\{p, w^*\}$. From the above argument, $w^*(s,t) \geq \underline{w}(s,t) \geq p(t)$. Therefore $w(s,t) = w_*(s,t) = p(t)$. By the definition of $b(\cdot)$, we conclude that $b(t) \geq s$. ∎

LEMMA 4.3 (a criterion for lower bounds). *For each $s < 0 < t$, let*

$$Q(s,r) := q(r) - \frac{2}{\sqrt{\pi}} \int_{-\infty}^{s/\sqrt{2r}} e^{-z^2}\, dz.$$

Suppose (s,t) is such that

$$s < 0 < t, \quad Q(s,r) \leq Q(s,t) \quad \forall\, r \in (0,t). \tag{4.2}$$

Then $b(t) \geq s$.

Proof. Let $\underline{w}$ be the solution to

$$\begin{cases} \mathcal{L}\underline{w} = 0 & \text{in } \mathbb{R}\times(0,t], \\ \underline{w}(\cdot,0) = \theta\, \mathbf{1}_{(2s,0)} & \text{on } \mathbb{R}\times\{0\}. \end{cases}$$

where

$$\theta = p(t) \left(\frac{2}{\sqrt{\pi}} \int_{s/\sqrt{2t}}^{0} e^{-z^2} dz \right)^{-1}.$$

We claim $\underline{w}$ satisfies (4.1).

1. Since the problem for $\underline{w}$ is linear, it can be expressed as

$$\underline{w}(x,r) = \theta \int_{2s}^{0} \Gamma(x-y,r)dy = \frac{\theta}{\sqrt{\pi}} \int_{\frac{x}{\sqrt{2r}}}^{\frac{x-2s}{\sqrt{2r}}} e^{-z^2}\, dz \qquad \forall\, x \in \mathbb{R}, r > 0. \tag{4.3}$$

In particular, when $x = s$,

$$\underline{w}(s,t) = \frac{\theta}{\sqrt{\pi}} \int_{s/\sqrt{2t}}^{-s/\sqrt{2t}} e^{-z^2} dz = \frac{2\theta}{\sqrt{\pi}} \int_{s/\sqrt{2t}}^{0} e^{-z^2} dz = p(t).$$

2. By (4.3), we find that

$$\max_{x\in\mathbb{R}} \underline{w}(x,r) = \underline{w}(s,r) = \frac{p(t) \int_{s/\sqrt{2r}}^{0} e^{-z^2} dz}{\int_{s/\sqrt{2t}}^{0} e^{-z^2} dz} \qquad \forall r > 0. \tag{4.4}$$

For any $s < 0 < r$, we can compute

$$\frac{2}{\sqrt{\pi}} \int_{s/\sqrt{2r}}^{0} e^{-z^2}\, dz = 1 - \frac{2}{\sqrt{\pi}} \int_{-\infty}^{s/\sqrt{2r}} e^{-z^2} dz = 1 + Q(s,r) - q(r) = Q(s,r) + p(r). \tag{4.5}$$

From (4.4) and (4.5), for all $r \in (0,t)$,

$$\begin{aligned}\max_{x\in\mathbb{R}} \underline{w}(x,r) - p(r) &= \frac{p(t)[Q(s,r)+p(r)]}{Q(s,t)+p(t)} - p(r)\\ &= \frac{p(t)Q(s,r) - p(r)Q(s,t)}{Q(s,t)+p(t)}\\ &\le \frac{p(t)[Q(s,r)-Q(s,t)]}{Q(s,t)+p(t)} \le 0,\end{aligned}$$

i.e., $\underline{w} \le p$ on $\mathbb{R}\times(0,t)$.

3. Since $Q(s,\cdot)$ is increasing for any $s<0$, $Q(s,t) \ge \lim_{r\searrow 0} Q(s,r) = 0$. In particular, when $r=t$, (4.5) reads as

$$\frac{2}{\sqrt{\pi}} \int_{s/\sqrt{2t}}^{0} e^{-z^2}\,dz = Q(s,t) + p(t) \ge p(t).$$

Therefore $\theta \le 1$, and thus $\underline{w}(\cdot,0) \le w(\cdot,0)$ on $\mathbb{R}\times\{0\}$. Lemma 4.2 now gives $b(t) \ge s$. □

The above lemma has an interesting probabilistic interpretation. Recalling that we are assuming that X_t is a Brownian motion, define $\underline{X}_t = \min_{v\le t} X_v$. Then

$$\begin{aligned}Q(s,r) &= q(r) - \frac{2}{\sqrt{\pi}} \int_{-\infty}^{s/\sqrt{2r}} e^{-z^2}\,dz\\ &= \mathbb{P}[\tau \le r] - 2\mathbb{P}[X_r \le s]\\ &= \mathbb{P}[\tau \le r] - \mathbb{P}[\underline{X}_r \le s],\end{aligned}$$

where the last line follows from the reflection principle (see, for example, Karatzas and Shreve [13]). Rearranging (4.2) then yields that if $s<0<t$ are such that

$$\mathbb{P}[\underline{X}_r > s \text{ and } \underline{X}_t \le s] \le \mathbb{P}[r < \tau \le t]$$

for all $r \in (0,t)$, then $b(t) \ge s$. This result is clear in the case that $b(t)$ is monotone but requires a more careful argument, as above, in the general case.

Before we continue, we provide an interesting application of Lemmas 4.3 and 4.1.

COROLLARY 4.4. *For each $t>0$, let $\zeta(t) \in (-\infty,0)$ and $\nu(t) \in \mathbb{R}$ be defined by*

$$q(t) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\zeta(t)/\sqrt{2t}} e^{-z^2}\,dz = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\nu(t)/\sqrt{2t}} e^{-z^2}\,dz.$$

1. *Suppose ζ is a constant function. Then the exact solution to* (1.1) *is given by*

$$\text{(4.6)} \qquad \begin{cases} w(x,t) = \frac{1}{\sqrt{\pi}} \int_{x/\sqrt{2t}}^{(x-2\zeta)/\sqrt{2t}} e^{-z^2}\,dz & \forall\, x \ge \zeta, t>0,\\ b(t) = \zeta & \forall\, t>0.\end{cases}$$

2. *Suppose $\zeta(r) \le \zeta(t)$ for all $r \in (0,t)$. Then $\zeta(t) \le b(t) \le \nu(t)$.*
3. *Suppose $\dot\zeta(t) \ge 0$ for all $t \in (0,T]$. Then*

$$\zeta(t) \le b(t) \le \nu(t) \quad \forall t \in (0,T], \quad \lim_{t\searrow 0} \frac{b(t)}{\zeta(t)} = 1.$$

Proof. 1. The first assertion may be verified by a direct computation. We note that it agrees with the formula for the first hitting time of Brownian motion to the level ζ (see, e.g., Karatzas and Shreve [13, pages 94–96]).

2. Suppose $\zeta(r) \le \zeta(t)$ for all $r \in (0,t)$. Set $s = \zeta(t)$. Then

$$Q(s,r) = q(r) - \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\zeta(t)/\sqrt{2r}} e^{-z^2}\, dz = \frac{2}{\sqrt{\pi}} \int_{\zeta(t)/\sqrt{2r}}^{\zeta(r)/\sqrt{2r}} e^{-z^2} dz \le 0 = Q(s,t)$$

for each $r \in (0,t)$. Thus, by Lemma 4.3, $b(t) \ge s = \zeta(t)$. This is the lower bound for $b(t)$.

For $t > 0$, Lemma 4.1 reads as

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{b(t)/\sqrt{2t}} e^{-z^2} dz \le q^*(t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\nu^*(t)/\sqrt{2t}} e^{-z^2} dz,$$

which implies that $b(t) \le \nu^*(t)$.

3. Suppose $\dot{\zeta}(t) \ge 0$ in $(0,T]$. Then by Corollary 4.4(2) $\zeta(t) \le b(t) \le \nu(t)$ for all $t \in (0,T]$. To complete the proof, it remains to estimate the difference between $\alpha(t) := \nu^*(t)/\sqrt{2t}$ and $\gamma(t) := \zeta(t)/\sqrt{2t}$. Let $\delta(t) = \ln 2/(-2\gamma(t) - 1)$. Since $\lim_{t \searrow 0} \gamma(t) = -\infty$, we conclude that for all small positive $t > 0$, $\delta(t) \in (0,1)$. Note that

$$\begin{aligned}
\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\alpha} e^{-z^2}\, dz = q(t) &= \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\gamma} e^{-z^2}\, dz = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\gamma+\delta} e^{-z^2+2\delta z-\delta^2} dz \\
&\le \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\gamma+\delta} e^{-z^2+2\delta(\gamma+\delta)-\delta^2} dz = \frac{2e^{\delta(2\gamma+\delta)}}{\sqrt{\pi}} \int_{-\infty}^{\gamma+\delta} e^{-z^2} dz \\
&= \frac{2e^{-\ln 2+\delta^2-\delta}}{\sqrt{\pi}} \int_{-\infty}^{\gamma+\delta} e^{-z^2}\, dz \le \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\gamma+\delta} e^{-z^2}\, dz.
\end{aligned}$$

Thus, $\alpha(t) \le \gamma(t) + \delta(t)$. Then

$$0 \le 1 - \frac{b(t)}{\zeta(t)} \le 1 - \frac{\nu(t)}{\zeta(t)} = 1 - \frac{\alpha(t)}{\gamma(t)} \le \frac{\delta(t)}{-\gamma(t)} = \frac{\ln 2}{2\gamma^2(t) + \gamma(t)}.$$

The third assertion of the lemma thus follows by sending $t \to 0$. □

Next, we present a sufficient condition for $Q(s,\cdot)$ to attain its maximum in $(0,t]$ at t. Note that for $0 < r < t$, $Q(s,r) \le Q(s,t)$ is equivalent to

$$\frac{q(t) - q(r)}{t - r} \ge \frac{2}{\sqrt{\pi}(t-r)} \left\{ \int_{-\infty}^{s/\sqrt{2t}} e^{-z^2} dz - \int_{-\infty}^{s/\sqrt{2r}} e^{-z^2} dz \right\} = \frac{-s\, e^{-s^2/(2\theta)}}{\theta^{3/2}\sqrt{2\pi}},$$

where $r < \theta \le t$. The second equality follows from the mean value theorem. To achieve the maximum in $(0,t]$ for $Q(s,\cdot)$, it is sufficient to have

$$\inf_{0<r<t} \frac{q(t) - q(r)}{t - r} \ge \sup_{r<\theta\le t} \frac{|s| e^{-s^2/(2\theta)}}{\theta^{3/2}\sqrt{2\pi}} = \frac{|s| e^{-s^2/(2t)}}{t^{3/2}\sqrt{2\pi}}, \tag{4.7}$$

provided that $s \le -\sqrt{3t}$. Taking the best possible $s \le -\sqrt{3t}$ for the inequality (4.7) to hold, we then obtain the following.

LEMMA 4.5. *Assume that $t > 0$ and*

$$k(t) := \inf_{0\le r\le t} \frac{q(t) - q(r)}{t - r} > 0.$$

Then

$$b(t) \geq \max\left\{ s \;\middle|\; s \leq -\sqrt{3t}\,;\, \frac{|s|}{\sqrt{2\pi}t^{3/2}} e^{-s^2/(2t)} \leq k(t) \right\}.$$

As an immediate consequence of the lemma, we have the following.

COROLLARY 4.6. *If $\dot{q}(t) > 0$, then $b(t) > -\infty$.*

We end this section with the following estimate on the asymptotic behavior of the boundary near 0. As mentioned above, this estimate is important when solving the inverse first passage problem numerically using integral equations such as (1.6); see [2].

THEOREM 5. *Assume that*

$$\limsup_{t \searrow 0} \frac{q(t)}{t\dot{q}(t)} < \infty. \tag{4.8}$$

Then

$$\lim_{t \searrow 0} \frac{b(t)}{\sqrt{-2t \log q(t)}} = -1. \tag{4.9}$$

Consequently, in special cases the following hold:

1. *when $q(t) = A\,t^m$, where A and m are positive constants,*

$$b(t) = -\sqrt{-2mt \log t}\;\; [1 + o(1)], \qquad \lim_{t \searrow 0} o(1) = 0;$$

2. *when $q(t) = A\,e^{-\gamma^2/(2t^m)}$, where A, m, γ are positive constants,*

$$b(t) = -\gamma\, t^{(1-m)/2}\;\; [1 + o(1)], \qquad \lim_{t \searrow 0} o(1) = 0.$$

In particular,

$$\lim_{t \searrow 0} b(t) = \begin{cases} -\infty & \text{if } m > 1, \\ \gamma & \text{if } m = 1, \\ 0 & \text{if } 0 < m < 1. \end{cases}$$

Proof. The idea is to estimate $k(r)$ via $q(r)/r$. Under the assumption (4.8), there exist positive constants C and T such that

$$0 < q(r) \leq C\, r\, \dot{q}(r) \qquad \forall\, r \in (0, T].$$

For any $0 < r < t \leq T$, we can compute

$$\begin{aligned} C\,(q(t) - q(r)) &\geq \int_r^t C\,\dot{q}(\theta)\,d\theta \geq \int_r^t \frac{q(\theta)}{\theta}\,d(\theta - r) \\ &= \frac{q(\theta)(\theta - r)}{\theta}\Big|_{\theta=r}^{\theta=t} - \int_r^t (\theta - r)\frac{\theta\dot{q}(\theta) - q(\theta)}{\theta^2}\,d\theta \\ &= \frac{(t-r)q(t)}{t} - \int_r^t \dot{q}(\theta)\,d\theta + \int_r^t \frac{r\theta\dot{q}(\theta) + (\theta - r)q(\theta)}{\theta^2}\,d\theta \\ &\geq (t-r)\frac{q(t)}{t} - [q(t) - q(r)]. \end{aligned}$$

That is, $(C+1)(q(t)-q(r)) \geq (t-r)\, q(t)/t$. It follows that

$$k(t) = \inf_{0<r<t} \frac{q(t)-q(r)}{t-r} \geq \frac{1}{C+1}\frac{q(t)}{t}.$$

Now fix $t \in (0,T]$, when $s \leq -\sqrt{3t}$, to be

$$\frac{|s|}{t^{3/2}\sqrt{2\pi}} \int_{-\infty}^{s/\sqrt{2t}} e^{-z^2} dz \leq \frac{|s|}{t^{3/2}\sqrt{2\pi}} e^{s^2/(2t)}.$$

Let $s < -\sqrt{3t}$ be the solution to

$$\frac{|s|}{\sqrt{2\pi}t^{3/2}} \int_{-\infty}^{s/\sqrt{2t}} e^{-z^2} dz = \frac{1}{C+1}\frac{q(t)}{t}. \tag{4.10}$$

For small t, $q(t)$ is small, so that $s/\sqrt{t} \ll -1$ and we can use the expansion, for $a < 0$,

$$\int_{-\infty}^{a} e^{-z^2} dz = \int_{a^2}^{\infty} \frac{e^{-x}dx}{2\sqrt{x}} = \frac{e^{-a^2}}{2|a|} - \int_{a^2}^{\infty} \frac{e^{-x}dx}{4x^{3/2}} = \frac{e^{-a^2}}{2|a|}\left\{1 - \frac{\theta}{2a^2}\right\},$$

where $\theta = \theta(a) \in (0,1)$. Hence, the equation for s reads as

$$e^{-s^2/(2t)}\left\{1 - \frac{\theta}{s^2/(2t)}\right\} = \frac{\sqrt{2\pi}}{C+1} q(t).$$

It then follows that

$$\begin{aligned} |s| &= \sqrt{2t}\left(-\log q(t) + \log(1-\theta t/s^2) + \log[(C+1)/\sqrt{2\pi}]\right)^{1/2} \\ &\leq \sqrt{-2t\log q(t)}\Big\{1 + (\log[(C+1)/\sqrt{4\pi}])/(-\log q(t))\Big\}^{1/2}. \end{aligned}$$

By Lemma 4.5, we then have

$$\begin{aligned} b(t) \geq s &\geq -\sqrt{-2t\log q(t)}\left\{1 + \frac{\log[(C+1)/\sqrt{2\pi}]}{|\log q(t)|}\right\}^{1/2} \\ &= -\sqrt{-2t\log q(t)}\{1+o(1)\}. \end{aligned}$$

This gives the lower bound for $b(t)$; now we estimate the upper bound. From Lemma 4.1,

$$q(t) \geq \frac{1}{\sqrt{\pi}} \int_{-\infty}^{b(t)/\sqrt{2t}} e^{-z^2}\, dz = \frac{\sqrt{t}}{\sqrt{2\pi}|b(t)|}\left\{1 - \frac{\theta\, t}{b^2(t)}\right\} e^{-b(t)^2/(2t)}.$$

This implies that

$$\begin{aligned} b(t) &\leq -\sqrt{-2t\log q(t)}\left\{1 + \frac{\log[1-\theta t/b^2(t)] - \log[\sqrt{2\pi}|b(t)|/\sqrt{t}]}{-\log q(t)}\right\}^{1/2} \\ &\leq -\sqrt{-2t\log q(t)}\left\{1 - \frac{O(1)\log|\log q(t)|}{|\log q(t) + o(1)|}\right\}^{1/2}. \end{aligned}$$

The assertion (4.9) thus follows. The remainder of the theorem is a direct application of (4.9). □

5. Integral equations. As in section 4, we assume $\sigma \equiv 1$ and $\mu \equiv 0$. Also we assume that p (and therefore q) is continuous. Then the solution to (3.19) can be expressed as

$$\begin{aligned} w(x,t) &= \int_{-\infty}^{0} \Gamma(x-y,t)dy + \int_0^t dp(s) \int_{-\infty}^{b(s)} \Gamma(x-y,t-s)\,dy \\ &= 1 - \int_0^{\infty} \Gamma(x-y,t)dy + \int_0^t dp(s) \left(1 - \int_{b(s)}^{\infty} \Gamma(x-y,t-s)\,dy\right) \\ &= p(t) - \int_{-\infty}^{x} \Gamma(z,t)\,dz + \int_0^t dq(s) \int_{-\infty}^{x-b(s)} \Gamma(z,t-s)\,dz, \end{aligned} \tag{5.1}$$

where the second equation is obtained by using $\int_{\mathbb{R}} \Gamma(x-y,s)dy = 1$ for all $s > 0$.

Now assume that b is smooth. Differentiating w, we can derive

$$u(x,t) = -w_x(x,t) = \Gamma(x,t) - \int_0^t \Gamma(x-b(s),t-s)\,dq(s). \tag{5.2}$$

Also, for $x \neq b(t)$, we can further differentiate to obtain

$$u_x(x,t) = \Gamma_x(x,t) - \int_0^t \Gamma_x(x-b(s),t-s)dq(s), \tag{5.3}$$

$$\begin{aligned} u_t(x,t) &= \Gamma_t(x,t) - \int_0^t \Gamma_t(x-b(s),t-s)dq(s) \\ &= \Gamma_t(x,t) + \int_0^t \left(\frac{d}{ds}\Gamma(x-b(s),t-s) + \dot{b}(s)\Gamma_x(x-b(s),t-s)\right) dq(s) \\ &= \Gamma_t(x,t) + \int_0^t \dot{b}(s)\Gamma_x(x-b(s),t-s)dq(s) + \int_0^t \dot{q}(s)d\Gamma(x-b(s),t-s) \\ &= \Gamma_t(x,t) + \int_0^t \dot{b}(s)\Gamma_x(x-b(s),t-s)dq(s) - \dot{q}(0)\Gamma(x,t) \\ &\quad - \int_0^t \Gamma(x-b(s),t-s)d\dot{q}(s), \end{aligned} \tag{5.4}$$

where the second equation is obtained by the equality

$$\Gamma_t(x-b(s),t-s) = -\frac{d}{ds}\Gamma(x-b(s),t-s) - \dot{b}(s)\Gamma_x(x-b(s),t-s)$$

and the third equation by using integration by parts to $\int_0^t \dot{q}(s)d\Gamma(x-b(s),t-s)$.

From potential theory, for any b and f with the certain regularity, we have

$$\lim_{x \to b(t) \pm 0} \int_0^t f(s)\Gamma_x(x-b(s),t-s)ds = \int_0^t f(s)\Gamma_x(b(t)-b(s),t-s)\,ds \mp f(t). \tag{5.5}$$

This can be derived as follows:

$$\begin{aligned}
&\lim_{x\to b(t)\pm 0}\int_0^t f(s)\Gamma_x(x-b(s),t-s)ds\\
&\quad=-\lim_{x\to b(t)\pm 0}\int_0^t f(s)\left[\frac{(b(t)-b(s))}{\sqrt{2\pi}(t-s)^{3/2}}e^{-\frac{(x-b(s))^2}{2(t-s)}}+\frac{(x-b(t))}{\sqrt{2\pi}(t-s)^{3/2}}e^{-\frac{(x-b(s))^2}{2(t-s)}}\right]ds\\
&\quad=\int_0^t f(s)\Gamma_x(b(t)-b(s),t-s)\,ds-\lim_{x\to b(t)\pm 0}\int_0^t f(s)\frac{x-b(t)}{\sqrt{2\pi}(t-s)^{3/2}}e^{-\frac{(x-b(s))^2}{2(t-s)}}ds\\
&\quad=\int_0^t f(s)\Gamma_x(b(t)-b(s),t-s)\,ds\\
&\qquad-\frac{2}{\sqrt{\pi}}\lim_{x\to b(t)\pm 0}\int_{\frac{x-b(t)}{\sqrt{2t}}}^{\pm\infty} f\left(t-\frac{(x-b(t))^2}{2\eta^2}\right)e^{-(\eta+\frac{b(t)-b(s)}{\sqrt{2(t-s)}})^2}d\eta\\
&\quad=\int_0^t f(s)\Gamma_x(b(t)-b(s),t-s)\,ds-\frac{2}{\sqrt{\pi}}\int_0^{\pm\infty} f(t)e^{-\eta^2}d\eta\\
&\quad=\int_0^t f(s)\Gamma_x(b(t)-b(s),t-s)\,ds\mp f(t).
\end{aligned}$$

Note that $w(x,t)=p$ and $0=u(x,t)=u_x(x,t)=u_t(x,t)$ for $x<b(t)$. Sending x to $b(t)$ from below in (5.1) and (5.2) we then obtain

$$\int_{-\infty}^{b(t)}\Gamma(z,t)dz=\int_0^t dq(s)\int_{-\infty}^{b(t)-b(s)}\Gamma(z,t-s)dz, \tag{5.6}$$

$$\Gamma(b(t),t)=\int_0^t\Gamma(b(t)-b(s),t-s)\,dq(s), \tag{5.7}$$

which reflect the free boundary condition $w(b(t),t)=p(t)$ and the condition $u(b(t),t)=0$, respectively. Sending x to $b(t)$ from below in (5.3) and (5.4) and using (5.5), we have

$$\dot q(t)=\Gamma_x(b(t),t)-\int_0^t\Gamma_x(b(t)-b(s),t-s)\,dq(s), \tag{5.8}$$

$$-\dot b(t)\dot q(t)=\Gamma_t(b(t),t)+\int_0^t\dot b(s)\Gamma_x(b(t)-b(s),t-s)dq(s),$$

$$-\dot q(0)\Gamma(b(t),t)-\int_0^t\Gamma(b(t)-b(s),t-s)d\dot q(s). \tag{5.9}$$

Equation (5.8) reflect the free boundary condition $u_x(b(t)^-,t)=0$ and $u_x(b(t)^+,t)=\dot q(t)$. Similarly, (5.9) reflects the free boundary condition that $u_t(b(t)^-,t)=0$ and $u_t(b(t)^+,t)=-\dot b(t)u_x(b(t)^+,t)=-\dot b(t)\dot q(t)$.

Clearly, these identities can provide numerical schemes much more flexible and economical than integrating the corresponding PDEs. For this purpose, it is necessary to study solutions to each of these identities.

One observes that if $b(\cdot)$ is a solution to (5.7), then $b_1(t):=-b(t)$ is also the solution. Hence, we need to be careful when considering solutions to the integral equation.

THEOREM 6. *Let $q:[0,\infty)\to[0,1)$ be continuous, increasing, and $q(0)=0$. Assume that $b:(0,T]\to\mathbb{R}$ is a continuous function. Then $x=b(t)$, $t\in(0,T]$, is the solution to the free boundary problem, provided that one of the following holds.*

1. *b satisfies* (5.6) *for all* $t \in (0,T]$*;*
2. *b satisfies* (5.7) *for all* $t \in (0,T]$*,* $b(t) < 0$ *for all sufficiently small positive* t*, and the function*

$$t \to q_{1/2}(t) := \int_0^t \frac{\dot{q}(t)}{\sqrt{2\pi(t-s)}} ds$$

is continuous in $(0,T]$ *with* $q_{1/2}(0+) = 0$*;*
3. *b satisfies* (5.8), $\lim_{t\searrow 0} \frac{b(t)}{\sqrt{t}} = -\infty$*,* $\dot{q}$ *is continuous in* $[0,T]$*, and the function*

$$t \to q^b_{3/2} := \int_0^t \frac{|b(t) - b(s)|}{\sqrt{2\pi}(t-s)^{3/2}} dq(s)$$

is continuous on $(0,T]$ *and is uniformly bounded.*

The analogous condition for (5.9) is too technical, and hence we omit it here.

Proof. With the given continuous function b, we define $w(x,t)$ as in (5.1). Note that

$$\begin{aligned} 0 \geq w(x,t) - w_0(x,t) &= \int_0^t dp(s) \int_{-\infty}^{b(s)} \Gamma(x-y, t-s) dy \\ &= \int_0^t dp(s) \left(1 - \int_{b(s)}^{\infty} \Gamma(x-y,t-s) dy \right) \\ &= -q(t) + \int_0^t dq(t) \int_{-\infty}^{x-b(s)} \Gamma(z, t-s) dz \geq -q(t). \end{aligned}$$

This implies that

$$|w(x,t) - w_0(x,t)| \leq q(t).$$

When $t = 0$, it reads as $|w(x,0) - w_0(x,0)| \leq q(0) = 0$, so that $w(x,0) = w_0(x,0) = \mathbf{1}_{(-\infty,0)}$.

In addition, upon differentiation, $\mathcal{L}w = \dot{p}\mathbf{1}_{\{x<b(t)\}} \leq 0$ as a measure in $\mathbb{R} \times (0,T]$. This can verified as follows. Direct calculation gives

$$\begin{aligned} w_{xx} &= -\Gamma_x(x,t) + \int_0^t \Gamma_x(x - b(s), t-\tau) dq(s), \\ w_t &= \dot{p}(t) - \int_{-\infty}^x \Gamma_t(z,t) dz + \int_0^t \int_{-\infty}^{x-b(s)} \Gamma_t(z, t-s) dz dq(s). \end{aligned}$$

Using the fact that $\Gamma_t = \frac{1}{2}\Gamma_{xx}$, we obtain

$$\int_{-\infty}^x \Gamma_t(z,t) dz = \int_{-\infty}^x \frac{1}{2}\Gamma_{xx}(z,t) dz = \frac{1}{2}\Gamma_x(z,t)\Big|_{z=-\infty}^{z=x} = \frac{1}{2}\Gamma_x(x,t).$$

When $x < b(s)$,

$$\begin{aligned} \int_0^t \int_{-\infty}^{x-b(s)} \Gamma_t(z, t-s) dz dq(s) &= \frac{1}{2} \int_0^t \Gamma_x(z, t-s)\Big|_{z=-\infty}^{z=x-b(s)} dq(s) \\ &= \frac{1}{2} \int_0^t \Gamma_x(x - b(s), t-\tau) dq(s). \end{aligned}$$

When $x > b(s)$,

$$\begin{aligned}
&\int_0^t \int_{-\infty}^{x-b(s)} \Gamma_t(z, t-s)dzdq(s) \\
&= \frac{1}{2}\int_0^t \Gamma_x(z, t-s)\Big|_{z=-\infty}^{z=(b(t)-b(s))^-} dq(s) + \frac{1}{2}\int_0^t \Gamma_x(z, t-s)\Big|_{z=(b(t)-b(s))^+}^{z=x-b(s)} dq(s) \\
&= \frac{1}{2}\int_0^t \Gamma_x(x-b(s), t-s)dq(s) + \lim_{z\nearrow(b(t)-b(s))} \frac{1}{2}\int_0^t \Gamma_x(z, t-s)dq(s) \\
&\quad - \lim_{z\searrow(b(t)-b(s))} \frac{1}{2}\int_0^t \Gamma_x(z, t-s)dq(s) \\
&= \dot{q}(t) + \frac{1}{2}\int_0^t \Gamma_x(x-b(s), t-s)dq(s) \\
&= -\dot{p}(t) + \frac{1}{2}\int_0^t \Gamma_x(x-b(s), t-s)dq(s).
\end{aligned}$$

Thus $\mathcal{L}w = \dot{p}\mathbf{1}_{\{x<b(t)\}} \le 0$ in $\mathbb{R}\times(0,T]$ holds.

It remains to show that $w(x,t) = p(t)$ for $x \le b(t)$ and $w < p(t)$ for $x > b(t)$.

(1) Assume the condition of the first assertion. We define

$$v(x,t) := w(x,t) - p(t) = -\int_{-\infty}^{x} \Gamma(z,t)\,dz + \int_0^t dq(s)\int_{-\infty}^{x-b(s)} \Gamma(z, t-s)\,dz.$$

Upon differentiation, $\mathcal{L}v = 0$ in $\{x < b(t)\}$. Note that v is bounded, continuous, and, by (5.6), $v(b(t),t) = 0$. It follows that $v(x,t) \equiv 0$ for all $x \le b(t)$, i.e., $w = p(t)$ for any $x \le b(t)$. Also by differentiation, we see that $\mathcal{L}v = -\dot{p} \le 0$ in $\{x > b(t)\}$. The strong maximum principle gives $v < 0$ in $\{x > b(t)\}$. That is, $w < p(t)$ in $\{x > b(t)\}$. Thus w is a variational solution.

(2) Assume the condition of the second assertion. We see that $u := -w_x$ given by (5.2) is continuous in $\mathbb{R}\times(0,\infty)$. For every small $\varepsilon > 0$, the function u satisfies $\mathcal{L}u = 0$ in

$$\Omega_\varepsilon := \{(x,t) \mid x < b(t), t \in (\varepsilon, T]\}.$$

Also, from (5.7), $u(b(t),t) = 0$ for all $t \in (0,T]$. Since $b(t) < 0$ for small positive t, we can assume that $b(\varepsilon) < 0$. It then follows from (5.2) that for all $x \le b(\varepsilon)$,

$$|u(x,\varepsilon)| \le \max\{\Gamma(x,\varepsilon), q_{1/2}(\varepsilon)\} \le \max\{\Gamma(b(\varepsilon),\varepsilon), q_{1/2}(\varepsilon)\} \le q_{1/2}(\varepsilon),$$

since for any $t \in (0,T]$

$$\int_0^t \Gamma(b(t)-b(s), t-s)dq(s) \le \int_0^t \frac{\dot{q}(s)}{\sqrt{2\pi(t-s)}}\,ds = q_{1/2}(t),$$

which holds for ε as well. It then follows from the maximum principle that

$$\max_{\Omega_\varepsilon} |u| \le q_{1/2}(\varepsilon).$$

Sending ε to 0 from above, we obtain $w_x = u \equiv 0$ in $\{(x,t) \mid x \le b(t), t > 0\}$. w is constant in $\{x \le b(t)\}$, and $w(-\infty,t) = p(t)$ imply that $w \equiv p(t)$ in $\{x \le b(t)\}$. From the first assertion, the second assertion of the theorem thus holds.

(3) Assume the conditions in the third assertion. Let $u := -w_x$ be given by (5.2) and $u_x := -w_{xx}$ by (5.3) when $x \neq b(t)$. Since b, $\dot{q}$, and $q^b_{3/2}$ are continuous, and (5.5) holds for $f = \dot{q}$, sending x to $b(t)$ from below in the equation for u_x and using (5.8) we derive that $u_x(b(t)^-, t) = 0$.

Next, we show that $u_x \equiv 0$ in $\{x < b(t)\}$. To do this, we first show that u_x given in (5.3) is uniformly bounded in $\{x < b(t)\}$. First, the boundedness of $q^b_{3/2}$ and (5.8) imply that $\Gamma_x(b(t), t)$ is uniformly bounded in $(0, T]$. Next, as $b(t) < -\sqrt{3t}$ for small positive t, we see that $0 < \Gamma_x(x, t) < \Gamma(b(t), t)$ for all $x < b(t)$. Thus $\Gamma_x(x, t)$ is bounded for all $x < b(t)$.

For $x < b(t)$, let $A_1 = \{s \in (0, t] \mid b(t) - x > 2|b(t) - b(s)|\}$ and $A_2 = [0, t] \setminus A_1$. Then

$$\int_0^t \Gamma_x(x - b(t), t - s)dq(s) = I_1 + I_2, \qquad I_i = \int_{A_i} \Gamma_x(x - b(t), t - s)dq(s).$$

Note that

$$|I_2| \leq \int_0^t \frac{|b(t) - b(s)|\dot{q}(s)}{2\sqrt{\pi}|t - s|^{3/2}} ds \leq 2q^b_{3/2}(t)$$

is uniformly bounded. To estimate I_i, notice that when $x - b(t) > 2|b(t) - b(s)|$, $(x - b(s))^2 = (x - b(t) - (b(t) - b(s)))^2 \geq \frac{1}{4}(x - b(t))^2$. Thus,

$$|I_1| \leq \int_0^t \frac{|x - b(t)|\dot{q}(s)e^{-|x-b(t)|^2/[16(t-s)]}ds}{\sqrt{2\pi}(t - s)^{3/2}} \leq \|\dot{q}\|_\infty,$$

and therefore u_x is uniformly bounded in $\{x < b(t)\}$.

Since $\mathcal{L}u_x = 0$ in $\{x < b(t), t > 0\}$, $u_x((b(t) - 0), t) = 0$, and $u_x(x, 0) = 0$ for all $x < 0$, a special maximum principle then implies that $u_x \equiv 0$ in $\{x < b(t)\}$. Using $u(-\infty, t) = 0$ we then conclude that $u \equiv 0$. Following (2), the third assertion of the theorem follows. □

Acknowledgments. The authors would like to thank Alex Kreinin, Soiliou Daw Namoro, Dan Rosen, Rob Stamicar, Ivan Yotov, the participants in the PhiMAC seminar at McMaster University, and two anonymous referees for helpful comments and suggestions.

REFERENCES

[1] M. Avellaneda and J. Zhu, *Modeling the distance-to-default process of a firm*, Risk, 14 (2001), pp. 125–129.
[2] L. Cheng, *Analysis and Numerical Solution of an Inverse First Passage Problem from Risk Management*, University of Pittsburgh, Pittsburgh, PA, 2005.
[3] M. G. Crandall, H. Ishii, and P. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[4] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.
[5] A. Friedman, *Variational Principles and Free-Boundary Problems*, John Wiley and Sons, New York, 1982.
[6] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.
[7] H. Huang and W. Tian, *Constructing default boundaries*, Banque & Marches, Jan.-Feb. (2006), pp. 21–28.
[8] J. Hull and A. White, *Valuing credit default swaps* I: *No counterparty default risk*, J. Derivatives, 8 (2000), pp. 29–40.

[9] J. Hull and A. White, *Valuing credit default swaps* II: *Modeling default correlations*, J. Derivatives, 8 (2001), pp. 12–21.
[10] I. Iscoe, A. Kreinin, and D. Rosen, *An integrated market and credit risk portfolio model*, Algo. Research Quarterly, 2 (1999), pp. 21–37.
[11] I. Iscoe and A. Kreinin, *Default Boundary Problem*, Internal Paper, Algorithmics Inc., Toronto, ON, Canada, 2002.
[12] R. B. Israel, J. S. Rosenthal, and J. Z. Wei, *Finding generators for Markov chains via empirical transition matrices with applications to credit ratings*, Math. Finance, 11 (2001), pp. 245–265.
[13] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer, New York, 1996.
[14] R. C. Merton, *On the pricing of corporate debt: The risk structure of interest rates*, J. Finance, 29 (1974), pp. 449–470.
[15] B. Øksendal, *Stochastic Differential Equations*, 6th ed., Springer, Berlin, 2002.
[16] G. Peskir, *Limit at zero of the Brownian first-passage density*, Probab. Theory Related Fields, 124 (2002), pp. 100–111.
[17] G. Peskir, *On integral equations arising in the first-passage problem for Brownian motion*, J. Integral Equations Appl., 14 (2002), pp. 397–423.
[18] L. M. Ricciardi, A. Di Crescenzo, V. Giorno, and A. G. Nobile, *An outline of theoretical and algorithmic approaches to first passage time problems with applications to biological modeling*, Math. Japonica, 50 (1999), pp. 247–322.
[19] C. Zucca, L. Sacerdote, and G. Peskir, *On the inverse first-passage problem for a Wiener process*, submitted.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 874–893

VORTEX LATTICES IN ROTATING BOSE–EINSTEIN CONDENSATES*

AMANDINE AFTALION† AND XAVIER BLANC†

Abstract. The structure of the vortex lattice for a fast rotating condensate in a harmonic trap has been studied experimentally and numerically: it is an almost regular hexagonal lattice, with a distortion on the edges. In this paper, we provide rigorous proofs of results announced in [A. Aftalion, X. Blanc, and J. Dalibard, *Phys. Rev. A*, 71 (2005), p. 023611]. We analyze the vortex pattern in the framework of the Gross–Pitaevskii energy using wave functions in the lowest Landau level. We compute the energy of a regular triangular lattice and of a class of distorted lattices and find the optimal distortion which provides a decay of the wave function similar to an inverted parabola.

Key words. Bose–Einstein condensates, lattice, lowest Landau level, averaging process, Gross–Pitaevskii, vortices

AMS subject classifications. 35Jxx, 35Q40, 81V45

DOI. 10.1137/050632889

1. Introduction.

One of the special features of Bose–Einstein condensates (BEC), related to superfluidity, is the existence of quantized vortices. These vortices can be observed in different types of experiments, one of them being the equivalent of what is known for helium as the rotating bucket experiment [17]. When a normal fluid is rotated, the velocity field inside the fluid is governed by solid body rotation. In contradistinction, a quantum fluid such as a BEC, described by a macroscopic wave function, nucleates vortices. This has been observed experimentally recently, in particular in the ENS group [13, 27] but also in [1, 28]. When the rotational velocity is small, there are only a few vortices in the system [12]. Their three dimensional shape is of interest, as has been described in [6, 7] using tools developed by [10] for Ginzburg–Landau vortices. At high velocity, the size of the condensate and the number of vortices increase: vortices arrange themselves in a lattice [1, 14, 18, 31, 33], referred to as an Abrikosov lattice due to the analogy with superconductors. Such a system can be related to homogeneous media since there are two scales emerging: the size of vortices (of order 1) and the size of the condensate (much larger). In this regime, vortices have approximately the same size as their mutual distance, which is very different from the lower rotation regime. Hence different mathematical tools need to be introduced relying on averaging effects (see [8]).

The description of the vortex lattice has been the focus of very recent papers in the condensed matter physics community, starting with the seminal paper of Ho [21] and very recently of Fischer and Baym [20], Baym and Pethick [9], Cooper, Komineas, and Read [15], Watanabe, Baym, and Pethick [37], and Sheehy and Radzihovsky [34]. Our aim is to provide mathematical insight to the lattice pattern. In particular, we show why a regular lattice has a higher energy than a distorted one and describe

*Received by the editors June 1, 2005; accepted for publication (in revised form) February 7, 2006; published electronically August 22, 2006. This work was supported by the fund of the French Ministry for Research, ACI "Nouvelles interfaces des mathématiques," CNRS, Collège de France, Région Ile de France, and DRED.

http://www.siam.org/journals/sima/38-3/63288.html

†Université Pierre et Marie Curie–Paris 6, CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, Paris, F-75005 France (aftalion@ann.jussieu.fr, blanc@ann.jussieu.fr).

the appropriate distortion and how the distortion towards the edges modifies the decay of the wave function. This is performed using the minimization of the Gross–Pitaevskii energy. This framework, known as the mean field Quantum Hall regime, is acceptable only if the number of vortices is much smaller than the number of atoms in the condensate, which is the case of the present experiments. Otherwise, one has to consider other models, as in [16, 35]. The reduction of the quantum many-body Hamiltonian to the Gross Pitaevskii energy is an open question for this fast rotating regime. It has been derived in the case of no rotation by Lieb, Seiringer, and Yngvason [25], and in the case of fixed rotation by Lieb and Seiringer [24].

The energy that we want to minimize is the following:

$$\int_{\mathbb{R}^2} \left(\frac{1}{2}|\nabla\phi(x)|^2 + \frac{1}{2}|x|^2|\phi(x)|^2 - \Omega\,(L\phi)\,(x)\overline{\phi}(x) + \frac{1}{2}G|\phi(x)|^4 \right) dx \tag{1.1}$$

under $\int_{\mathbb{R}^2} |\phi|^2 = 1$, where $x = (x_1, x_2) \in \mathbb{R}^2$, $L = i(x_2\partial_{x_1} - x_1\partial_{x_2})$ is the angular momentum, Ω is the rotational velocity. The contributions to the energy are the kinetic term $(1/2)|\nabla\phi|^2$, the term due to the potential trapping the atoms $(1/2)|x|^2|\phi|^2$ and the rotating term $-\Omega L$. The last term is due to atomic interactions, G being a positive constant equal to Na, where N is the number of atoms in the condensate and a the scattering length. In fact, we should have taken a three-dimensional version of this energy, but as we are interested in a rotational velocity Ω close to 1, the minima of the 2D and 3D energies are asymptotically the same (see [3]). Indeed, the effective perpendicular trapping frequency becomes $\sqrt{1-\Omega^2}$, and is thus much smaller than the trapping frequency in the x_3 direction, which is fixed: the condensate is strongly confined along the x_3 axis and it is expected that, as Ω tends to 1, the 3D wave function is well approximated by the product of a Gaussian in the x_3 direction times a 2D wave function minimizing (1.1). The kinetic and rotational terms in the energy (1.1) are the beginning of the expansion of a complete square. Adding and subtracting the missing term modifies the trapping potential, creating what is called the effective trapping potential, and the energy (1.1) can be rewritten as follows:

$$E(\phi) = \int_{\mathbb{R}^2} \frac{1}{2}|\nabla\phi - i\Omega x^{\perp}\phi|^2 + \frac{1}{2}(1-\Omega^2)|x|^2|\phi|^2 + \frac{1}{2}G|\phi|^4, \tag{1.2}$$

where $x^{\perp} = (-x_2, x_1)$. In order for the energy to be bounded below, we need to have $\Omega < 1$, which means that the trapping potential remains stronger than the rotating force. As $\Omega < 1$ approaches one, the extension of the condensate increases. The first term in the energy is identical to the energy of a particle placed in a uniform magnetic field 2Ω. It is also reminiscent of type II superconductors near the second critical field H_{c2}. The minimizers for

$$\int_{\mathbb{R}^2} \frac{1}{2}|\nabla\phi - i\Omega x^{\perp}\phi|^2 \text{ under } \int_{\mathbb{R}^2} |\phi|^2 = 1 \tag{1.3}$$

are well known [23, 26] through the study of the eigenvalues of the operator $-(\nabla - i\Omega x^{\perp})^2$. The minimum is Ω and it is achieved in a space of infinite dimension called the lowest Landau level (LLL). This space is spanned by

$$\phi(x_1, x_2) = P(z)e^{-\Omega|z|^2/2} \text{ with } z = x_1 + ix_2, \tag{1.4}$$

where P varies in the space of polynomials. The other eigenvalues are $(2k+1)\Omega$, $k \in \mathbb{N}$.

We will see that as Ω approaches 1, the second and third term in the energy (1.2) produce a contribution of order $\sqrt{1-\Omega}$, which is much smaller than the gap between two eigenvalues of $-(\nabla - i\Omega x^\perp)^2$, namely 2Ω. Thus, it is natural, as a first step, to restrict to the minimizers of (1.3) and minimize the energy (1.2) in this reduced infinite dimensional space. Since we want to keep the same space as Ω varies, we will use the rescaled wave function

$$\psi(x) = \frac{1}{\sqrt{\Omega}}\phi\left(\frac{x}{\sqrt{\Omega}}\right), \tag{1.5}$$

which satisfies the condition $\int |\psi|^2 = 1$. Therefore, the energy (1.2) provides $E(\phi) = \tilde{E}(\psi)$ with

$$\tilde{E}(\psi) = \int_{\mathbb{R}^2} \frac{\Omega}{2}|\nabla\psi - ix^\perp\psi|^2 + \frac{1-\Omega^2}{2\Omega}|x|^2|\psi|^2 + \frac{1}{2}G\Omega|\psi|^4, \tag{1.6}$$

and the condition (1.4) becomes

$$\psi(x) = P(z)e^{-|z|^2/2} \text{ with } P(z) = A\prod_{i=1}^{n}(z-z_i) \text{ and } z = x_1 + ix_2. \tag{1.7}$$

For such a ψ, the first term of the energy is equal to Ω. Hence, we find that the energy $\tilde{E}(\psi)$ is equal to

$$E_{LLL}(\psi) := \Omega + \int_{\mathbb{R}^2} \frac{1-\Omega^2}{2\Omega}|x|^2|\psi|^2 + \frac{G\Omega}{2}|\psi|^4. \tag{1.8}$$

The issue is to minimize the energy (1.8) on the space (1.7), and in particular understand the optimal location of the zeros or vortices z_i and the averaged behavior of ψ on large spheres.

The ansatz (1.7) is motivated by very recent physics papers: in a seminal paper, Ho [21] computed the energy (1.6) of a configuration of the type (1.7), where the z_i are located on a triangular lattice and found that the wave function averaged over vortex cells has a Gaussian decay. This was confirmed by [9]. Only recently did Cooper, Komineas, and Read [15] observe numerically the distortion of the lattice on the edges of the condensate and the decay of the wave function, which is closer to an inverted parabola than a Gaussian. The inverted parabola profile can be explained in two ways: either by taking contributions to the wave function in excited Landau levels or by distorting the lattice in the LLL—that is, moving the z_i in (1.7) away from a regular lattice. Watanabe, Baym, and Pethick [37] claimed (without rigorously proving it) that the minimizer is achieved by a distorted lattice inside the LLL. In this paper, we want to understand rigorously the distortion of the lattice and are going to provide a rigorous proof of the results announced in [2]: in [2], we have performed numerical computations, fixing an upper bound on the number of zeros and using a conjugate gradient on the z_i to find a minimizer of the energy. This provides the pattern for vortices illustrated in Figure 1. On the left, we have plotted the z_i and on the right $|\psi|$, where ψ is related to the z_i through (1.7): in a central region, vortices

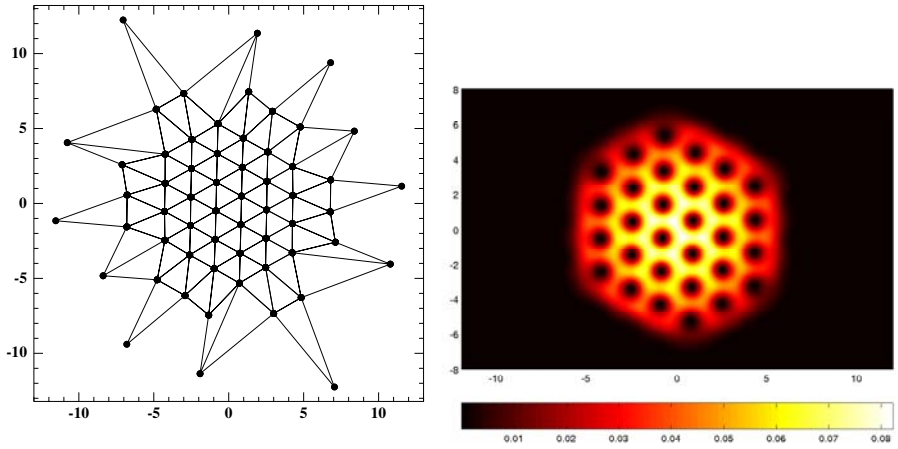


FIG. 1. *An example of (left) a configuration of z_i minimizing the energy for $\Omega = 0.999$, $G = 3$, and $n = 58$; (right) density plot of $|\psi|$.*

are located on a regular triangular lattice, while the lattice is distorted towards the edges. The density plot of $|\psi|$ shows that the only visible vortices are the central ones in the regular lattice part, the outer ones being in a region of very low density.

If we minimize the energy (1.8) without any ansatz, we find that the minimizer is

$$|\psi_{\min}(r)|^2 = \frac{2}{\pi R_0^2}\left(1 - \frac{r^2}{R_0^2}\right)_+, \quad R_0 = \left(\frac{4G\Omega^2}{\pi(1-\Omega^2)}\right)^{1/4}. \tag{1.9}$$

This is a first indication of the main scale of the problem: the extension of the condensate is of size R_0, which is large when Ω approaches 1. The value of the energy yields a lower bound for E_{LLL}:

$$\epsilon_{\min} := E_{LLL}(\psi_{\min}) - \Omega \underset{\Omega\to 1}{\sim} \frac{2\sqrt{2}}{3}\sqrt{\frac{G(1-\Omega)}{\pi}}\,. \tag{1.10}$$

Since $\psi_{\min}$ has compact support, it is not in the closure of the space spanned by (1.7). Thus, computing the energy of $\psi_{\min}$ requires going back to (1.6). This yields a contribution due to the kinetic energy of order R^2 and hence much larger than (1.10). Nevertheless, with an appropriate location of the zeros z_i, the aim is to build a test function whose energy can be of the same order as (1.10) and whose decay weakly approaches that of (1.9) as Ω tends to 1. Our main results, announced in [2], are the following: if the vortices are located on a regular lattice, the wave function decays like a Gaussian and we provide a rigorous proof of the energy estimate obtained by Ho [21]. On the other hand, we are able to build distortions of the lattice which improve the energy and hence modify the decay of the wave function. This decay is similar to that of an inverted parabola, as already observed by [15, 33, 37].

In what follows, ℓ will denote a regular hexagonal lattice, whose unit cell Q, centered at the origin has volume V. Moreover, we will identify complex numbers and vectors in $\mathbb{R}^2$ and in particular dz will denote the two dimensional Lebesgue measure $dx = dx_1\, dx_2$ when $z = x_1 + ix_2$ and $x = (x_1, x_2)$. The symbol $\fint$ is the average of an ℓ-periodic function: $\fint f = \frac{1}{V}\int_Q f$.

THEOREM 1.1. *Let ℓ be a regular hexagonal lattice and Q its unit cell. Let*

$$\psi_R(z) = A_R \prod_{j\in\ell\cap B_R} (z-j)e^{-|z|^2/2} \tag{1.11}$$

with A_R chosen such that $\|\psi_R\|_{L^2(\mathbb{R}^2)} = 1$. If $V = |Q| > \pi$, we define

$$\frac{1}{\sigma^2} = 1 - \frac{\pi}{V}, \tag{1.12}$$

then we have, as R tends to ∞,

$$|\psi_R(z)| \longrightarrow \psi(z) = \frac{1}{\sqrt{\pi}\sigma}\eta(z)e^{-|z|^2/(2\sigma^2)} \text{ in } L^p\left(\mathbb{R}^2, (1+|z|^2)dz\right) \forall p \geq 1, \tag{1.13}$$

where η is an ℓ-periodic function which vanishes at each point of ℓ. Moreover, η is the unique (up to multiplication by a constant) solution of

$$-\Delta(\ln\eta) = 2\pi\delta_0 - \frac{2\pi}{V} \quad \text{in } Q,$$

with periodic boundary conditions. In addition, $\lim_{R\to+\infty} E_{LLL}(\psi_R) = E_{LLL}(\psi)$. As σ tends to infinity,

$$E_{LLL}(\psi) - \Omega \sim \frac{1-\Omega^2}{2\Omega}\sigma^2 + \frac{1}{4}\frac{G\Omega b}{\pi\sigma^2} \quad \text{where } b = \frac{\fint |\eta|^4}{(\fint |\eta|^2)^2}, \tag{1.14}$$

uniformly with respect to Ω.

The main feature of the periodic lattice is to modify the decay of the Gaussian from $e^{-|z|^2/2}$ to $e^{-|z|^2/2\sigma^2}$, where σ depends on the volume through (1.12). The choice of the normalization by the L^2 norm implies convergence of ψ_R only if $V > \pi$.

We need to choose the optimal σ in (1.14), which yields

$$\sigma^4 = \frac{1}{2}\frac{Gb\Omega^2}{\pi(1-\Omega^2)}. \tag{1.15}$$

This value of σ indeed satisfies $\sigma \to +\infty$ as Ω tends to 1. The volume condition (1.12) matched with the value of σ (1.15) implies

$$V = \pi\left(1 + \sqrt{(1-\Omega)\frac{4\pi}{Gb}}\right) + O(1-\Omega). \tag{1.16}$$

This is close to the value predicted by solid body rotation arguments, π/Ω (see [21]), but different. The estimate of the energy is thus

$$E_{LLL}(\psi) - \Omega \underset{\Omega\to 1}{\sim} \sqrt{\frac{Gb}{\pi}(1-\Omega)}. \tag{1.17}$$

This is to be compared to (1.10), which is better by a factor $\sqrt{8/9b}$, but is of the same magnitude, as $1-\Omega$ is small. Let us emphasize the presence of the coefficient b: it takes into account the averaged vortex contribution in each cell. As in the case of superconductors near H_{c_2}, for the Abrikosov lattice, the optimal lattice minimizing the ratio b is the hexagonal one [22]. An approximate value of b is 1.16. Note that our proof holds with other lattices than the hexagonal one.

The aim of the paper is to improve the numerical factor in front of the square root in (1.17). The main observation is that modifying the location of the vortices from a regular lattice can change the decay of the wave function and hence improve the energy estimate.

THEOREM 1.2. *There exists a sequence of functions ψ_Ω of the form* (1.7), *such that as Ω tends to* 1,

$$E_{LLL}(\psi_\Omega) - \Omega \sim \frac{2\sqrt{2}}{3}\sqrt{\frac{Gb}{\pi}(1-\Omega)}. \tag{1.18}$$

This is closer to the lower bound (1.10) than the regular lattice: the numerical factor is the same as in (1.10), except for the coefficient b, coming from the averaged vortex contribution.

Let us now explain the main ideas of the proof, announced in [2]. For the regular lattice, we split $\ln|\psi_R(z)|$ into $v_R(z) + w_R(z)$ with

$$v_R(z) = \sum_{j\in\ell\cap B_R}\left(\ln|z-j| - \frac{1}{V}\int_Q \ln|z-y-j|\,dy\right), \tag{1.19}$$

$$w_R(z) = \ln A_R - \frac{|z|^2}{2} + \frac{1}{V}\sum_{j\in\ell\cap B_R}\int_Q \ln|z-y-j|\,dy. \tag{1.20}$$

At this stage, we have just added and subtracted the sum of the integrals. As R tends to ∞, we prove that v_R converges to a periodic series v and e^{w_R} to a Gaussian with modified decay $1/\sigma^2$. The computation of the energy uses the fact that for a periodic function η, $\eta(\sigma z)$ converges weakly to the average of η (see Lemma 2.5 and [8, 29]). This allows one to separate the integrals of the Gaussian and of v, which vary on two different scales; the v integral provides the term b.

Let us be more precise about Theorem 1.2. We define the radius R by

$$R = \left(\frac{2Gb}{\pi(1-\Omega)}\right)^{1/4}, \tag{1.21}$$

and we perform a general transformation of the lattice in the following way: for j in ℓ, a regular triangular lattice of unit cell with volume $V = \pi$, we define the transformed lattice ℓ'_R by

$$k \in \ell'_R \text{ if } k = \nu_R(|j|)\,j \text{ for } j \in \ell\cap B_R. \tag{1.22}$$

We assume that ν_R is close to 1 as Ω tends to 1, in the sense that

$$\nu_R^2(r) = 1 + \frac{f(r^2/R^2)}{R^2} + O\left(\frac{1}{R^4}\right), \tag{1.23}$$

where $f(t)$ is a continuous function, such that for some γ, $f(\gamma) = \infty$ and $\int_0^\gamma f(s)\,ds = \infty$. Actually, we will see in (1.28) that the one providing an energy minimizer is $f(s) = 1/(1-s)$.

We would like to apply the same proof as for the regular lattice, using v_R and w_R for this distorted lattice. Contrary to the proof for the regular lattice, we cannot study the two limits $R\to\infty$ in (1.13) and $\sigma\to\infty$ in (1.14) separately, since R is

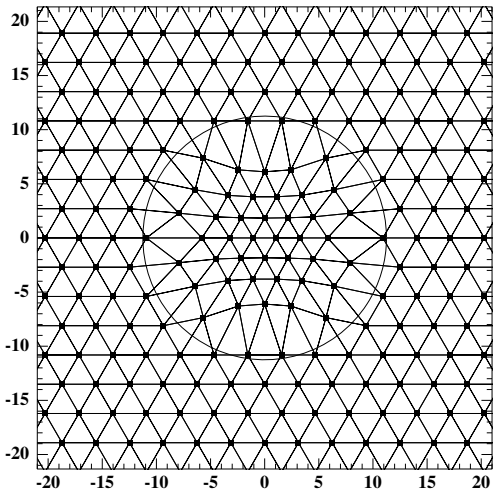


FIG. 2. *A plot of the distorted lattice defined by* (1.24).

now related to Ω through (1.21). Hence, the lattice has a finite extent at each R and we have to pass to the limit in the averaging process at the same time as the scale of the lattice. For technical reasons, we are unable to match w_R inside and outside the lattice and use the dominated convergence theorem.

In order to circumvent this problem, we introduce an outer regular lattice, whose characteristic size tends to infinity in a last step. Let $\alpha \in (0, \gamma)$, R be related to Ω by (1.21), and

$$\lambda_R(r) = \begin{cases} \nu_R(r) & \text{if} \quad r \leq \alpha R, \\ \nu_{\alpha,R} = \nu_R(\alpha R) & \text{if} \quad r > \alpha R, \end{cases} \tag{1.24}$$

and $\ell'_R = \{\lambda_R(|j|)j, \quad j \in \ell\}$. We have plotted such a transformation in Figure 2. For fixed α, we let R tend to ∞, and study the limit of the wave functions vanishing at each point of ℓ'_R:

$$\psi_R(z) = A_R \prod_{j \in \ell} (z - \lambda_R(|j|)j)\, e^{-|z|^2/2}. \tag{1.25}$$

Since α is fixed, $\nu_R(\alpha R)$ tends to 1. We use similar ideas as in the regular lattice case and identify two parts in the limit of $|\psi_R(Rz)|^2$ as R tends to ∞: a periodic part, which is rapidly oscillating and averaged in the limit, multiplied by a profile depending on the transformation f, which is equal to

$$|\psi(z)|^2 = e^{-F(|z|^2)} \mathbf{1}_{B_\alpha}(z) + e^{\alpha^2 f(\alpha^2) - F(\alpha^2) - f(\alpha^2)|z|^2} \mathbf{1}_{B_\alpha^c}(z), \tag{1.26}$$

where $F : \mathbb{R}^+ \to \mathbb{R}$ satisfies $F' = f$. The proof uses as a main tool that λ_R is close to one. As a final step only, once we have passed to the limit $\Omega \to 1$, we let α tend to γ, so that the exterior regular lattice has a unit volume cell which tends to infinity and the outer contribution disappears. We find an estimate for the energy:

$$E_{LLL}(\psi_\Omega) - \Omega \underset{\Omega \to 1}{\sim} \sqrt{\frac{2Gb(1-\Omega)}{\pi}} \int_0^\gamma \left(s e^{-F(s)} + \frac{1}{4} e^{-2F(s)} \right) ds, \tag{1.27}$$

where F is such that $F' = f$ and $\int_0^\gamma e^{-F(s)}\, ds = 1$.

We want to find which type of distortion f provides the optimal energy. The minimizer of (1.27) under $\int_0^\gamma e^{-F(s)}\,ds = 1$ is reached when

$$\gamma = 1 \text{ and } e^{-F(r^2)} = 2(1-r^2)_+. \tag{1.28}$$

Thus, the decay of the wave function is asymptotically an inverted parabola. The corresponding value of f is $f(s) = 1/(1-s)$. The limiting value of the energy is (1.18).

Let us point out that the proof uses two lattices: an initial regular lattice and an image lattice obtained by (1.22) and (1.23). The meaning of $\gamma = 1$ in (1.28) is that the initial lattice is truncated in the ball B_R: the points outside B_R are sent to infinity. There are two regions in the initial lattice: the points sufficiently far away from the circle of radius R, for which ν_R is almost one, and the points close to the circle, at distance less than $\sqrt{R}$ for instance. For the first category of points, the image lattice is an almost regular lattice and the image points are inside the disk B_R. These are the visible points on the density profile. On the contrary, the points close to the circle are strongly modified by (1.23) and sent far away. This allows one to better understand the distorted shape of Figure 1. It turns out that R is both the radius of the "horizon" for the initial lattice, but also the radius of the limiting inverted parabola (1.28). The points which are not visible in the density profile, and are in the distorted region, have nevertheless a contribution in creating the inverted parabola profile, since they provide the decay of the wave function: if they are removed, the decay is that of a Gaussian.

For each Ω, this analysis gives an estimate of the number N_v of points in the distorted lattice: it is given by the number of points in a regular lattice of unit volume π, included in a ball of radius R given by (1.21), namely

$$N_v \sim \sqrt{\frac{2Gb}{\pi(1-\Omega)}}.$$

This is probably a good estimate of the number of significant points to produce an energy of order (1.18). Nevertheless, adding more points in the low density region may lower the energy at the next order term in $(1-\Omega)$. In [2], we have studied numerically the dependence of E_n, the minimum of the energy over polynomials of degree less than or equal to n, which corresponds to at most n vortices. With the parameters $\Omega = 0.999$ and $G = 3$, we find that E_n becomes almost constant for $n \geq 52$ (with a precision of order 10^{-10}) which is comparable to the present estimate $N_v = 47$. Nevertheless, in [4, 5], we have now proved, using Bargmann spaces, that the minimizer has an infinite number of zereos. It means that additional zeroes appear in the low density region and contribute to decrease higher order terms in the energy. Consequently, the present upper bound only provides information for vortices entering in the first order expansion of the energy.

One of our technical tools in the proof is to use an outer regular lattice whose spacing tends to infinity in a last step. If one wanted to get rid of this trick, one would need to count the number of points in the lattice closest to the limiting circle of radius R and estimate the convergence of v_R and w_R due to the fact that these limiting points do not lie on a circle but on the edges of hexagons. We are not able to prove that the finite extension of the lattice (which becomes infinite with an outer regular lattice) does not create a boundary contribution in the energy. These boundary effects seem to be more important than we expected and are related to known problems about counting the number of points of a lattice in an annulus (see [11] and the references therein).

Our results deal with an upper bound for the energy. A natural question would be to also get the lower bound and prove Γ convergence–type results. The present gap between the lower bound (1.10) and the upper bound (1.18) lies in the coefficient b. We believe that an optimal lower bound should match the upper bound and that the limiting inverted parabola should have a radius given by (1.21) instead of (1.9); that is, the optimal inverted parabola should have the coefficient b. Let us give a few more explanations.

A first question is to understand the lower bound for the energy restricted to the LLL and prove that it matches our upper bound (1.18). For the moment, we are unable to prove it, though we expect that reproducing an inverted parabola profile in the space (1.7) requires a lot of vortices and thus creates a contribution in the energy through b. Our numerical computations agree with this. But proving that a minimizing sequence has zeros which are located on an almost regular lattice seems very difficult and is probably related to similar difficulties in crystalization and sphere packing problems.

Another issue is to check that the lower bound of the energy restricted to the LLL should provide the lower bound for the full energy (1.6). More precisely, if ψ is a minimizer of (1.6), we can project it on the LLL and its orthogonal through $\psi = \psi_{LLL} + \psi_\perp$. The upper bound and the properties of the operator (1.3) imply that $\|\psi_{LLL}\|_{L^2}$ is close to 1 and $\|\psi_\perp\|_{L^2}$ is small like $\sqrt{1-\Omega}$. If we had for E_{LLL} a lower bound in the LLL matching the upper bound (1.18), this would imply $\|\psi_\perp\|_{L^2} = o\left(\sqrt{1-\Omega}\right)$. Using the Poincaré inequality of Lu and Pan [26], one can improve the norm and get an L^∞ estimate for $\psi_\perp$. This should allow one to prove that the leading order term in the energy of ψ is $E_{LLL}(\psi_{LLL})$.

The authors of [37] provide a formal argument to exclude contributions of excited Landau levels. Let us give a gist on how to justify it. Let us call ρ the rescaled limiting density distribution given by (1.26), which can either be a Gaussian if the lattice is regular or approach an inverted parabola if the lattice is distorted. We have seen that the energy of this test function is computed through

$$\sqrt{\frac{2Gb(1-\Omega)}{\pi}} \int \left(r^2\rho^2 + \rho^4\right) \text{ under } \int \rho^2 = 1. \tag{1.29}$$

This is done by using a wave function ψ_Ω such that $|\psi_\Omega|^p \to (⨍ |\eta|^p)\rho^p$. In [37], the following ansatz is introduced: the authors replace ψ_Ω by $h_\Omega\psi_\Omega$, where ψ_Ω is the above wave function and h_Ω is a modulation due to the contribution of higher Landau levels. It is in addition assumed that h_Ω varies on a much larger scale than the vortex lattice. This assumption of slow variation allows them to compute the limit of $h_\Omega\psi_\Omega$ using averaging effects and find as a limiting energy of (1.6), when Ω tends to 1,

$$\sqrt{\frac{2Gb(1-\Omega)}{\pi}} \int \frac{1}{2}\rho^2|\nabla h|^2 + r^2\rho^2h^2 + \rho^4h^4 \text{ under } \int \rho^2h^2 = 1. \tag{1.30}$$

Note that $h = 1$ reproduces the previous computations. Minimizing (1.30) and using a convexity argument, we find that $\int \left(h^2\rho^2 - p^2\right)^2$ tend to 0, where p is the rescaled inverted parabola. If ρ itself is close to the inverted parabola, this implies that h is close to 1, and thus the contribution from higher Landau levels is negligible.

Let us point out that other trapping potentials than $|x|^2$ can be dealt with these techniques. In [2], we have addressed the case of $|x|^2 + k|x|^4$ with k small, following recent experiments [13, 36]. For certain values of Ω, a giant vortex can be obtained.

This paper is organized as follows: In section 2, we study the regular lattice case and prove Theorem 1.1. Then, in section 3, we prove Theorem 1.2. Finally, section 4 is devoted to remarks concerning other trapping potentials.

2. Regular lattice. In this section, we prove Theorem 1.1. We first need two technical lemmas:

LEMMA 2.1. *Let ℓ be a lattice, and denote by Q its unit cell centered at 0. Let $Q_R = \bigcup_{k\in\ell\cap B_R}(Q+k)$ and for x in $\mathbb{R}^2$, let*

$$h_R(x) = \int_{Q_R} (\ln|x-x'| - \ln|x'|)\,dx'.$$

Then there exists $C>0$ and $R_0>0$ such that

$$\forall R \geq R_0, \quad h_R(x) \leq \left(\frac{\pi}{2} + \frac{C}{R}\right)|x|^2.$$

Proof. If Q_R was a ball, then the integral could be computed explicitly. Thus, we use a ball close to Q_R and estimate the difference. We separate the integral defining h_R into two parts:

$$h_R(x) = \int_{B_{R-a}} (\ln|x-x'| - \ln|x'|)\,dx' + \int_{Q_R\setminus B_{R-a}} (\ln|x-x'| - \ln|x'|)\,dx',$$

where $a>0$ is independent of R and such that $B_{R-a}\subset Q_R$. The first term is the radial solution of $\Delta u = \mathbf{1}_{B_{R-a}}$ such that $u(0)=0$. One easily computes this solution:

$$u(x) = \frac{\pi}{2}|x|^2\mathbf{1}_{B_{R-a}} + \pi(R-a)^2\left(\frac{1}{2} + \ln\left(\frac{|x|}{R-a}\right)\right)\mathbf{1}_{B^c_{R-a}}.$$

Next, we consider the second term defining h_R and use the inequality $\ln(t)\leq\frac{1}{2}(t^2-1)$, valid for any $t>0$:

$$\begin{aligned}\int_{Q_R\setminus B_{R-a}} (\ln|x-x'| - \ln|x'|)\,dx' &\leq \int_{Q_R\setminus B_{R-a}} \frac{1}{2}\left(\frac{|x-x'|^2}{|x'|^2} - 1\right)dx' \\ &= |x|^2\int_{Q_R\setminus B_{R-a}} \frac{dx'}{2|x'|^2} \leq C\frac{|x|^2}{R},\end{aligned}$$

the constant C being independent of R and x. Collecting both results, we infer

$$\begin{aligned}h_R(x) &\leq \frac{\pi}{2}|x|^2\mathbf{1}_{B_{R-a}}(x) + \pi(R-a)^2\left(\frac{1}{2} + \ln\left(\frac{|x|}{R-a}\right)\right)\mathbf{1}_{B^c_{R-a}}(x) + \frac{C}{R}|x|^2 \\ &\leq \left(\frac{\pi}{2} + \frac{C}{R}\right)|x|^2,\end{aligned}$$

here using again $\ln(t)\leq\frac{1}{2}(t^2-1)$. This gives the result. ∎

LEMMA 2.2. *Let ℓ be the hexagonal lattice, and let Q be its elementary unit cell (i.e., the regular hexagon centered at 0). Let*

$$g(x) = \ln|x| - \frac{1}{|Q|}\int_Q \ln|x-y|\;dy. \tag{2.1}$$

Then we have, for some constant $C > 0$,

$$\forall x \in B_1^c, \quad |g(x)| \leq \frac{C}{|x|^3}. \tag{2.2}$$

Hence, the function

$$v(x) = \sum_{j \in \ell} g(x-j) \tag{2.3}$$

is such that $e^{v(x)}$ exists, is continuous on $\mathbb{R}^2$, and ℓ-periodic.

Proof. We first point out that g is continuous on $\mathbb{R}^2 \setminus \{0\}$. Hence, we need only to show (2.2) on B_a^c for some $a > 0$. We fix $a > 0$ such that $Q \subset B_{\frac{a}{2}}$. For any $x \in B_a^c$ and any $y \in Q$, we have $\frac{|x-y|}{|x|} \geq \frac{|x|-\frac{a}{2}}{|x|} \geq \frac{1}{2}$. Hence,

$$\frac{|x-y|^2}{|x|^2} - 1 \geq -\frac{3}{4}.$$

For any $t > -\frac{3}{4}$, we have

$$t - \frac{t^2}{2} - |t|^3 \leq \ln(1+t) \leq t - \frac{t^2}{2} + \frac{t^3}{3}.$$

Hence, writing $g(x) = -\frac{1}{|Q|}\int_Q \frac{1}{2} \ln\left(1 - \frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2}\right) dy$, we infer

$$\frac{1}{2|Q|}\int_Q \left(-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2} - \frac{1}{2}\left(-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2}\right)^2 - \left|-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2}\right|^3\right) dy \leq -g(x)$$
$$\leq \frac{1}{2|Q|}\int_Q \left(-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2} - \frac{1}{2}\left(-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2}\right)^2 + \frac{1}{3}\left(-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2}\right)^3\right) dy.$$

Since Q is symmetric with respect to the origin, $\int_Q y\cdot x\, dy = 0$. In addition, Q is invariant under the rotation of angle $\pi/3$, so one easily shows that $\int_Q \left(|y|^2 - 2(y\cdot x)^2\right) dy = 0$. We thus have

$$\frac{1}{2|Q|}\int_Q \left(\frac{2y\cdot x|y|^2}{|x|^4} - \frac{|y|^4}{|x|^4} - \left|-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2}\right|^3\right) dy \leq g(x)$$
$$\leq \frac{1}{2|Q|}\int_Q \left(\frac{2y\cdot x|y|^2}{|x|^4} - \frac{|y|^4}{|x|^4} + \frac{1}{3}\left(-\frac{2y\cdot x}{|x|^2} + \frac{|y|^2}{|x|^2}\right)^3\right) dy.$$

Using $|y| \leq \frac{a}{2}$, we end up with

$$|g(x)| \leq \frac{1}{|Q|}\int_Q \left(\frac{|y|^3}{|x|^3} + \frac{|y|^4}{2|x|^4} + \frac{1}{2}\left|\frac{2|y|}{|x|} + \frac{|y|^2}{|x|^2}\right|^3\right) dy \leq \frac{\frac{a^3}{8} + \frac{a^3}{16} + 2a^3}{|x|^3}.$$

This ensures that the series (2.3) converges normally on any set of the form $(\bigcup_{j\in\ell} B_\varepsilon(j))^c$, which implies that v exists, is ℓ-periodic, and continuous on $\mathbb{R}^2 \setminus \ell$. Near a point $k \in \ell$, we write

$$e^{v(x)} = |x-k| e^{-\frac{1}{|Q|}\int_Q \ln|x-y|\ dy} e^{\sum_{j\in\ell\setminus\{k\}} g(x-j)}$$

and the conclusion follows. ∎

Remark 2.3. It is in this lemma that we have used the symmetry properties of the lattice. However, the same proof applies to a general lattice (which does not necessarily have the above symmetries). In this case, one needs to use instead of $g = \ln|\cdot| * (\delta_0 - \frac{1}{|Q|}\mathbf{1}_Q)$ the convolution of $\ln|\cdot|$ with a distribution g_0 such that $\sum_{k\in\ell} g_0(x-k) = \sum_{k\in\ell}\delta_k - \frac{1}{|Q|}$ and such that the first harmonic moments of g_0 cancel. This is always possible (in such a case g_0 is not supported inside Q).

Remark 2.4. The estimate (2.2) is valid for a fixed hexagonal lattice ℓ_0. Now, $g = g_\ell$ depends on ℓ in the following way: if $\ell = \lambda\ell_0$, then $g_\ell(x) = g_{\ell_0}\left(\frac{x}{\lambda}\right)$. Hence, $|g_\ell(x)| \le \frac{C_0\lambda^3}{|x|^3}$ if $|x| \ge \lambda$ for some constant C_0 independent of ℓ.

We finally state the following lemma, which may be seen as the simplest case of two-scale convergence [8, 29]. The proof is standard and may be found, for example, in [8].

LEMMA 2.5. *Let $v \in L^1_{\mathrm{loc}}\left(\mathbb{R}^2\right)$, which is periodic, and set $v_\sigma(x) = v(\sigma x)$ for any $\sigma > 0$. Then*

$$v_\sigma \xrightarrow[\sigma\to\infty]{*} \fint v \text{ in } L^\infty(\mathbb{R}^2).$$

Proof of Theorem 1.1. Let $f_R(z) = \ln|\psi_R(z)|$. We split f_R into

$$f_R(z) = v_R(z) + w_R(z) \tag{2.4}$$

with

$$v_R(z) = \sum_{j\in l\cap B_R} \ln|z-j| - \frac{1}{V}\int_Q \ln|z-y-j|\,dy, \tag{2.5}$$

$$w_R(z) = \ln A_R - \frac{|z|^2}{2} + \frac{1}{V}\sum_{j\in\ell\cap B_R}\int_Q \ln|z-y-j|\,dy. \tag{2.6}$$

Let v be given by (2.3). We have

$$v_R(z) - v(z) = \sum_{j\in\ell\cap B_R^c} g(z-j).$$

Hence, if $z \in B_R$, we deduce from Lemma 2.2 that

$$|v_R(z) - v(z)| \le \sum_{j\in\ell\cap B_R^c} \frac{C}{|z-j|^3}$$

for some constant C independent of R and z. One can thus find a constant C independent of R such that

$$\forall A \in (0,R), \quad \|v_R - v\|_{L^\infty(B_{R-A})} \le \frac{C}{A}. \tag{2.7}$$

In addition, we have, for any $z \in \mathbb{R}^2$, denoting by j_z the unique point of ℓ such that $|j_z - z| < 1$,

$$v_R(z) \le \ln|z-j_z| + C + \sum_{j\in\ell\setminus\{j_z\}} \frac{C}{|z-j|^3} \le \ln|z-j_z| + C,$$

for various constants C independent of z and R. Hence, e^{v_R} is bounded in $L^\infty(\mathbb{R}^2)$ independently of R. Next, using the inequality $|e^a - e^b| \leq \frac{1}{2}(e^a + e^b)|a-b|$ and (2.7), we infer that e^{v_R} converges to e^v in $L^\infty_{\rm loc}(\mathbb{R}^2)$.

Let us call $\tilde{w}_R(z) = w_R(z) - w_R(0) - \ln(A_R) + \frac{1}{2\sigma^2}|z|^2$. Applying Lemma 2.1, we have

$$\tilde{w}_R(z) \leq -\frac{|z|^2}{2} + \left(\frac{\pi}{2V} + \frac{C}{R}\right)|z|^2 + \frac{1}{2\sigma^2}|z|^2 = \frac{C}{R}|z|^2. \tag{2.8}$$

In addition, $\tilde{w}_R$ is a harmonic function in $Q_R = \bigcup_{j\in\ell\cap B_R}(Q+j)$ and vanishes at 0. Hence, using Harnack inequality, $\tilde{w}_R$ is bounded and we may extract convergence of $\tilde{w}_R$ in $L^\infty_{\rm loc}(\mathbb{R}^2)$ to some $\tilde{w}$, which is harmonic nonpositive, and vanishes at 0. Applying Liouville theorem, we find that $\tilde{w} = 0$. Gathering all the previous results, we thus have

$$\frac{|\psi_R(z)|}{A_R e^{w_R(0)}} \longrightarrow e^{v(z)}e^{-|z|^2/(2\sigma^2)} \text{ almost everywhere in } \mathbb{R}^2. \tag{2.9}$$

For R large enough, (2.8) also implies

$$\frac{|\psi_R(z)|}{A_R e^{w_R(0)}} \leq Ce^{-\frac{|z|^2}{4\sigma^2}}. \tag{2.10}$$

From (2.9) and (2.10), we apply the dominated convergence theorem and find that, as R tends to infinity,

$$\frac{|\psi_R|}{A_R e^{w_R(0)}} \longrightarrow e^{v(z)}e^{-|z|^2/(2\sigma^2)} \text{ in } L^p(\mathbb{R}^2) \quad \forall p \geq 1.$$

Using the fact that $\|\psi_R\|_{L^2} = 1$, we deduce that $1/(A_R e^{w_R(0)})$ converges to the appropriate constant, so that (1.13) holds.

Then we write the limiting energy, z being identified with a vector in $\mathbb{R}^2$:

$$\begin{aligned} E_{LLL}(\psi) &= \Omega + \int_{\mathbb{R}^2}\left(\frac{1-\Omega^2}{2\Omega}|z|^2|\eta(z)|^2e^{-|z|^2/\sigma^2} + \frac{G\Omega}{2\sigma^2}|\eta(z)|^4e^{-2|z|^2/\sigma^2}\right)\frac{dz}{\pi\sigma^2} \\ &= \Omega + \int_{\mathbb{R}^2}\left(\frac{1-\Omega^2}{2\Omega}\sigma^2|\eta(\sigma\xi)|^2|\xi|^2e^{-|\xi|^2} + \frac{G\Omega}{2\sigma^2}|\eta(\sigma\xi)|^4e^{-2|\xi|^2}\right)\frac{d\xi}{\pi}. \end{aligned}$$

We want to find the limit as σ tends to infinity. Since the function η is periodic, Lemma 2.5 implies that $|\eta(\sigma\xi)|^2$ and $|\eta(\sigma\xi)|^4$, respectively, converge L^∞-weak-$*$ to $⨍|\eta|^2$ and $⨍|\eta|^4$, and (1.14) holds. □

3. Distorted lattice. In this section, we prove two theorems, which will imply Theorem 1.2. Theorem 3.1 consists of studying a distorted lattice analogous to Figure 2 and find the limit of the wave function with an infinite number of vortices. The proof is similar to the regular lattice case, since only the central vortices are displaced from their regular location. Theorem 3.2 identifies the limit of the modulus of the wave function as Ω tends to 1: a periodic, rapidly oscillating part which converges in L^∞-weak-$*$ and a slowly varying profile. The proof is more involved since this is precisely where the distortion of the lattice appears.

THEOREM 3.1. *Let ℓ be a hexagonal lattice, and let Q be the regular hexagon of area π centered at zero. Let $\gamma > 0$ and let f be a positive Lipschitz continuous function defined in $[0,\gamma)$ such that*

$$\lim_{t\to\gamma} f(t) = +\infty \quad \textit{and} \quad \lim_{t\to\gamma}\int_0^t f(s)ds = \infty. \tag{3.1}$$

Let us define

$$\nu_R(t) = 1 + \frac{1}{2R^2} f\left(\frac{t^2}{R^2}\right) + O\left(\frac{1}{R^4}\right), \tag{3.2}$$

where $O\left(\frac{1}{R^4}\right)$ *is uniform with respect to* $t \in \mathbb{R}^+$. *Let* $\alpha \in (0, \gamma)$

$$\lambda_R(t) = \begin{cases} \nu_R(t) & \text{if} \quad t \le \alpha R, \\ \nu_{\alpha,R} = \nu_R(\alpha R) & \text{if} \quad t > \alpha R. \end{cases} \tag{3.3}$$

For $R' > R$, *we define*

$$\psi_{R,R'}(x) = A_{R,R'} \prod_{j \in \ell \cap B_{R'}} (x - \lambda_R(|j|)j)\, e^{-|x|^2/2}, \tag{3.4}$$

where $A_{R,R'}$ *is such that* $\|\psi_{R,R'}\|_{L^2(\mathbb{R}^2)} = 1$. *Then, we have the following convergence in* $L^p(\mathbb{R}^2, (1+|x|^2)dx)$ *for any* $p < +\infty$*:*

$$|\psi_{R,R'}| \underset{R' \to +\infty}{\longrightarrow} A_R e^{v_R\left(\frac{x}{\nu_{\alpha,R}}\right) + w_R(x) + \left(\frac{1}{{\nu_{\alpha,R}}^2} - 1\right)\frac{|x|^2}{2}}, \tag{3.5}$$

where

$$w_R(x) = \sum_{j \in \ell \cap B_{\alpha R}} \frac{1}{\nu_{\alpha,R}^2 |Q|} \int_{\nu_{\alpha,R} Q} \ln\left(\frac{|x - y - \nu_R(|j|)j|}{|x - y - \nu_{\alpha,R} j|}\right) dy, \tag{3.6}$$

and

$$v_R(y) = \sum_{j \in \ell} g\left(y - \frac{\lambda_R(|j|)}{\nu_{\alpha,R}} j\right), \tag{3.7}$$

the function g *being defined by* (2.1).

Then, we let Ω tend to 1, or equivalently R to infinity.

THEOREM 3.2. *With the same definitions as in Theorem* 3.1, *we have*

$$\forall n \ge 1, \quad e^{n v_R(Rx)} \underset{R \to +\infty}{\overset{*}{\longrightarrow}} \fint e^{nv} \quad \text{in} \quad L^\infty(\mathbb{R}^2), \tag{3.8}$$

$$e^{2w_R(Rx) + \left(\frac{1}{\lambda_R(\alpha R)^2} - 1\right) R^2 |x|^2} \underset{R \to +\infty}{\longrightarrow} \rho(x) \tag{3.9}$$

in $L^p(\mathbb{R}^2, (1+|x|^2)dx)$ $\forall p \ge 1$, *where* v *is given by* (2.3),

$$\rho(x) = e^{-F(|x|^2)} \mathbf{1}_{B_\alpha}(x) + e^{\alpha^2 f(\alpha^2) - F(\alpha^2) - f(\alpha^2)|x|^2} \mathbf{1}_{B_\alpha^c}(x), \tag{3.10}$$

and F *is such that* $F' = f$ *and* $\fint e^{2v} \int \rho = 1$.

Proof of Theorem 1.2. We let Ω tend to 1, let R be given by (1.21), and take a diagonal sequence in R'. Theorems 3.1 and 3.2 provide the convergence of $\int |\psi_{R,R'}(Rz)|^2$ to $\fint e^{2v} \int \rho$, and similarly for the energy:

$$E_{LLL}(\psi_{R,R'}(Rz)) - \Omega \underset{\Omega \to 1}{\sim} \sqrt{\frac{2Gb(1-\Omega)}{\pi}} \left(\fint e^{2v} \int_0^\infty s\rho(\sqrt{s})\, ds + \frac{1}{4} \fint e^{4v} \int_0^\infty \rho^2(\sqrt{s})\, ds \right), \tag{3.11}$$

where ρ is defined by (3.10), F is such that $F' = f$ and $\int_0^\gamma e^{-F(s)}\, ds = 1$ and $\fint e^{2v} = 1$. If one lets α tend to γ: the contribution to ρ in the outer part B_γ^c vanishes and the energy is given by (1.27).

We want to find which type of distortion f provides the optimal energy. The minimizer of (1.27) under $\int_0^\gamma e^{-F(s)}\, ds = 1$ is reached when

$$\gamma = 1 \text{ and } e^{-F(r^2)} = 2(1 - r^2). \tag{3.12}$$

Thus, the decay of the wave function is asymptotically an inverted parabola. The corresponding value of f is $f(s) = 1/(1-s)$. The limiting value of the energy is (1.18).

Proof of Theorem 3.1. This proof is a mere adaptation of section 2. Indeed, up to normalization by a constant, the function $\ln|\psi_{R,R'}|^2$ is equal to

$$\ln|\psi_{R,R'}(x)|^2 = 2\sum_{j\in\ell\cap B_{R'}} \Bigg(\ln|x - \lambda_R(|j|)j| \tag{3.13}$$

$$-\frac{1}{{\nu_{\alpha,R}}^2|Q|}\int_{\nu_{\alpha,R}Q} \ln|x - y - \lambda_R(|j|)j|dy\Bigg) \tag{3.14}$$

$$+\sum_{j\in\ell\cap B_{R'}} \frac{2}{{\nu_{\alpha,R}}^2|Q|}\int_{\nu_{\alpha,R}Q} \ln|x - y - \lambda_R(|j|)j|dy \tag{3.15}$$

$$-|x|^2. \tag{3.16}$$

The sum (3.13)–(3.14) may be written

$$\sum_{j\in\ell\cap B_{R'}} g\left(\frac{x}{\nu_{\alpha,R}} - \frac{\lambda_R(|j|)}{\nu_{\alpha,R}}j\right), \tag{3.17}$$

where g is defined by (2.1). Now, R being fixed, Lemma 2.2 ensures that the above sum converges as R' goes to infinity to $v_R(\frac{x}{\nu(\alpha R)})$, where v_R is defined by (3.7). Moreover, the convergence of the exponential of (3.17) to $e^{v_R\left(\frac{x}{\nu(\alpha R)}\right)}$ is the same as in Theorem 1.1, that is, $L^\infty_{\mathrm{loc}}(\mathbb{R}^2)$. Next, the sum (3.15) is equal to

$$\begin{aligned}
&\sum_{j\in\ell\cap B_{R'}} \frac{2}{{\nu_{\alpha,R}}^2|Q|}\int_{\nu_{\alpha,R}Q} \ln|x - y - \lambda_R(|j|)j|dy \\
&= \sum_{j\in\ell\cap B_{\alpha R}} \frac{2}{{\nu_{\alpha,R}}^2|Q|}\int_{\nu_{\alpha,R}Q} \ln\frac{|x - y - \nu_R(|j|)j|}{|x - y - \nu_{\alpha,R}\, j|}dy \\
&+ \sum_{j\in\ell\cap B_{R'}} \frac{2}{{\nu_{\alpha,R}}^2|Q|}\int_{\nu_{\alpha,R}Q} \ln|x - y - \nu_{\alpha,R}\, j|dy.
\end{aligned} \tag{3.18}$$

The first sum in the left-hand side of (3.18) is $w_R(x)$, while the second sum is the one appearing in (2.6), with $\nu(\alpha R)\ell$ replacing ℓ. Since this lattice is also a hexagonal one (with a different volume for its unit cell), the proof of its convergence applies, using Lemma 2.1. ∎

Proof of Theorem 3.2. For simplicity, we will give the proof in the case where the $O(1/R^4)$ is zero. We start with the proof of (3.9). We define $\varepsilon > 0$ depending on R such that, as R tends to infinity,

$$\begin{cases} R\varepsilon \longrightarrow +\infty, \\ R\varepsilon^2 \longrightarrow 0. \end{cases} \tag{3.19}$$

For instance, $\varepsilon = R^{-3/4}$ is a suitable choice. Writing

$$w_R(Rx) = \sum_{k \in \frac{\ell}{R} \cap B_\alpha} \frac{R^2}{{\nu_{\alpha,R}}^2 |Q|} \int_{\frac{\nu_{\alpha,R}}{R} Q} \ln \left(\frac{|x - z - \nu_R(R|k|)k|}{|x - z - \nu_{\alpha,R} k|} \right) dz,$$

we split this sum into terms for which $|k - x| < \varepsilon$, and terms for which $|x - k| \geq \varepsilon$: in the first case, we use the inequality

$$\forall a, b > 0, \quad |\ln a - \ln b| \leq \frac{1}{2} \left(\frac{1}{a} + \frac{1}{b} \right) |b - a|,$$

and the fact that $|\nu_{\alpha,R} - \nu_R(|k|R)| \leq \frac{C}{R^2}$ for some constant C independent of R and x. Hence,

$$\begin{aligned}
&\left| \sum_{|k-x|<\varepsilon} \frac{R^2}{\nu_{\alpha,R}|Q|} \int_{\frac{\nu_{\alpha,R}}{R} Q} \ln \left(\frac{|x - z - \nu_R(R|k|)k|}{|x - z - \nu_{\alpha,R} k|} \right) dz \right| \\
&\leq \sum_{|k-x|<\varepsilon} \frac{R^2}{2\nu_{\alpha,R}|Q|} \int_{\frac{\nu_{\alpha,R}}{R} Q} \Bigg(\frac{1}{|x - z - \nu_R(R|k|)k|} \\
&+ \frac{1}{|x - z - \nu_{\alpha,R} k|} \Bigg) |k| |\nu_{\alpha,R} - \nu_R(|k|R)| dz \\
&\leq \frac{C}{R^2} \sum_{|k-x|<\varepsilon} \frac{R^2}{2\nu_{\alpha,R}|Q|} \int_{B_{\frac{3}{R}}} \frac{dy}{|y|} \leq C \# \left(\frac{\ell}{R} \cap B_\varepsilon(x) \right) \frac{1}{R} = CR\varepsilon^2,
\end{aligned}$$

which tends to zero as $R \to +\infty$. Next, we deal with $|k - x| \geq \varepsilon$, and denote the corresponding sum by $T_R(x)$:

$$T_R(x) = \sum_{|k-x|\geq\varepsilon} \frac{R^2}{{\nu_{\alpha,R}}^2 |Q|} \int_{\frac{\nu_{\alpha,R}}{R} Q} \ln \left(\frac{|x - z - \nu_R(R|k|)k|}{|x - z - \nu_{\alpha,R} k|} \right) dz.$$

Using the equality $\nu_R(R|k|) = 1 + \frac{f(|k|^2)}{2R^2}$, valid for any $|k| \leq \alpha$, we deduce

$$T_R(x) = \sum_{|k-x|\geq\varepsilon} \frac{R^2}{{\nu_{\alpha,R}}^2 |Q|} \int_{\frac{\nu_{\alpha,R}}{R} Q} \ln \left(\frac{|x - z - k - \frac{f(|k|^2)}{2R^2} k|}{|x - z - k - \frac{f(\alpha^2)}{2R^2} k|} \right) dz.$$

We have $|x - z - k| \geq |x - k| - |z| \geq \varepsilon - \frac{C}{R} = \varepsilon(1 - \frac{C}{\varepsilon R})$ for $z \in \frac{\nu_{\alpha,R}}{R} Q$, so that for R large enough, we get $|x - z - k| \geq \frac{\varepsilon}{2}$. Hence, developing the quotient in the logarithm, we get

$$T_R(x) = \sum_{|k-x|\geq\varepsilon} \frac{R^2}{2{\nu_{\alpha,R}}^2 |Q|} \int_{\frac{\nu_{\alpha,R}}{R} Q} \ln \left(\frac{1 - \frac{f(|k|^2)}{R^2 |x-z-k|^2} k \cdot (x - z - k) + O\left(\frac{1}{\varepsilon^2 R^4}\right)}{1 - \frac{f(\alpha^2)}{R^2 |x-z-k|^2} k \cdot (x - z - k) + O\left(\frac{1}{\varepsilon^2 R^4}\right)} \right) dz,$$

where the $O\left(\frac{1}{\varepsilon^2 R^4}\right)$ are uniform with respect to x and k. Developing the logarithm, we thus find

$$T_R(x) = \sum_{|k-x|\geq\varepsilon} \left(\frac{R^2}{{\nu_{\alpha,R}}^2 |Q|} \int_{\frac{\nu_{\alpha,R}}{R} Q} \frac{f(\alpha^2) - f(|k|^2)}{R^2} \frac{k \cdot (x - z - k)}{|x - z - k|^2} dz \right) + O\left(\frac{1}{\varepsilon^2 R^2} \right).$$

Using the fact that f is smooth in $[0,\alpha]$, and recalling that the sum is a sum over the set $\frac{\ell}{R}\cap B_\alpha\cap B_\varepsilon(x)^c$, we find that it converges to the corresponding integral, namely

$$\lim_{R\to+\infty} T_R(x)=\frac{1}{|Q|}\int_{B_\alpha}\left(f(\alpha^2)-f(|y|^2)\right)\frac{y\cdot(x-y)}{|x-y|^2}dy.$$

We then point out that $\frac{x-y}{|x-y|^2}=-\nabla_y \ln|x-y|$, so that, integrating by parts, we have

$$\lim_{R\to+\infty} w_R(Rx)=\frac{1}{|Q|}\int_{B_\alpha}\operatorname{div}\left(f(\alpha^2)-f(|y|^2)y\right)\ln|x-y|dy.$$

This limit is a radially symmetric function, which solves the partial differential equation $\Delta u=\frac{2\pi}{|Q|}\operatorname{div}\left(f(\alpha^2)-f(|y|^2)y\right)$ in B_α, $\Delta u=0$ elsewhere. For any F such that $F'=f$, the function $\frac12\left(f(\alpha^2)|y|^2-F(|y|^2)\right)\mathbf{1}_{B_\alpha}(y)+\frac12\left(\alpha^2 f(\alpha^2)-F(\alpha^2)\right)\mathbf{1}_{B_\alpha^c}(y)$ is such a solution, so we have (3.9) almost everywhere. In addition, the above proof allows one to bound $2w_R(Rx)+(\frac{1}{\nu_{\alpha,R}^2}-1)R^2|x|^2$ by $C-\frac{f(\alpha^2)}{4}|x|^2$ for some constant C independent of R and x, which allows one to apply the dominated convergence theorem.

We now prove (3.8). We fix $n=1$, the general proof following exactly the same pattern. It is sufficient to show that the following convergence holds for any measurable bounded set D:

$$\int_D e^{v_R(Rx)}\underset{R\to+\infty}{\longrightarrow}|D|\fint e^v. \tag{3.20}$$

Hence, we are going to prove that for any $a>0$,

$$\left|e^{v_R(x)}-e^{v(x)}\right|\le C\frac{1+\sqrt{|x|}}{R}\quad\text{for}\quad|x|\le aR. \tag{3.21}$$

Since, according to Lemma 2.5, $e^{v(Rx)}$ converges in L^∞ weak-$*$ to $\fint e^v$ (because e^v is continuous and periodic), (3.21) will give (3.20). Let j_x be the point of ℓ which is the closest to x. As R goes to infinity, $|(\lambda_R(|j_x|)-1)j_x|=O\left(\frac1R\right)$ uniformly with respect to x since $|x|\le aR$. Hence, for R large enough, $\lambda_R(|j_x|)j_x$ is the closest to x among all $\lambda_R(|j|)j$, $j\in\ell$. Hence, for $j\in\ell\setminus\{j_x\}$, we have, for some $\varepsilon>0$,

$$\forall y\in Q,\quad |x-j-y|\ge\varepsilon\quad\text{and}\quad|x-j-\lambda_R(|j|)(j+y)|\ge\varepsilon. \tag{3.22}$$

We then isolate j_x in the sum defining v_R and write

$$\left|e^{v_R(x)}-e^{v(x)}\right|\le\left|e^{g_R(x-\lambda_R(|j_x|)j_x)}-e^{g_R(x-j_x)}\right|e^{\sum_{j\neq j_x}g_R(x-\lambda_R(|j|)j)} \tag{3.23}$$

$$+e^{g_R(x-j_x)}\left|e^{\sum_{j\neq j_x}g_R(x-\lambda_R(|j|)j)}-e^{\sum_{j\neq j_x}g_R(x-j)}\right|, \tag{3.24}$$

where $g_R(z)=g(\frac{z}{\nu_{\alpha,R}})$. We first bound (3.23). For this purpose, we point out that, according to Lemma 2.2, one can find a constant C_ε such that

$$\forall z\in B_\varepsilon^c,\quad |g(z)|\le\frac{C_\varepsilon}{|z|^3}. \tag{3.25}$$

Hence, the sum appearing in (3.23) may be bounded as follows:

$$\sum_{j\neq j_x}|g_R(x-\lambda_R(|j|)j)|\le\sum_{j\neq j_x}\frac{C_\varepsilon{\nu_{\alpha,R}}^3}{|x-\lambda_R(|j|)j|^3},$$

which is bounded independently of R. Moreover, we have

$$\left|e^{g_R(x-\lambda_R(|j_x|)j_x)} - e^{g_R(x-j_x)}\right| \le C\,||x-\lambda_R(|j_x|)j_x| - |x-j_x||$$
$$+C\left|\int_Q \left(\ln\frac{|x-y-j_x|}{|x-y-\lambda_R(|j_x|)j_x|}\right)dy\right|;$$

thus

$$\left|e^{g_R(x-\lambda_R(|j_x|)j_x)} - e^{g_R(x-j_x)}\right| \le C\,|(1-\lambda_R(|j_x|)j_x|$$
$$+C\int_Q \left(\frac{1}{|x-y-j_x|}+\frac{1}{|x-y-\lambda_R(|j_x|)j_x|}\right)|1-\lambda_R(|j_x|)||j_x|dy.$$

This implies

$$\left|e^{g_R(x-\lambda_R(|j_x|)j_x)} - e^{g_R(x-j_x)}\right| \le C\frac{|x|}{R^2} + C\frac{|x|}{R^2}\int_{B_3}\frac{dy}{|y|} \le C\frac{|x|}{R^2}.$$

Hence, the left-hand side of (3.23) is bounded by $C\frac{|x|}{R^2}$. Next, we deal with (3.24). Since g is bounded from above, it is sufficient to show the following:

$$\left|\sum_{j\neq j_x} g_R(x-\lambda_R(|j|)j) - \sum_{j\neq j_x} g_R(x-j)\right| \le C\frac{1+\sqrt{|x|}}{R}. \tag{3.26}$$

In order to prove (3.26), we define $A>0$ depending on R and x, to be fixed later on, and distinguish in the above sum between terms for which $|j-j_x|\le A$ and those for which $|j-j_x|>A$. We have

$$\sum_{0<|j-j_x|\le A} |g_R(x-\lambda_R(|j|)j) - g_R(x-j)| \le \|\nabla g\|_{L^\infty(B_\varepsilon^c)} \sum_{0<|j-j_x|\le A} |j||\lambda_R(|j|)-1|$$
$$\le \frac{C}{R^2}\sum_{0<|j-j_x|\le A}|j| \le \frac{C}{R^2}A^2(|x|+A).$$

We have used here the fact that g is Lipschitz continuous in B_ε^c. Considering the case $|j-j_x|>A$, we have, using (3.25),

$$\sum_{A>|j-j_x|} |g_R(x-\lambda_R(|j|)j) - g_R(x-j)| \le \sum_{A>|j-j_x|}\frac{C}{|x-j|^3} \le \frac{C}{A}.$$

We thus may bound the left-hand side of (3.26) by $\frac{C}{A}+\frac{CA^2|x|}{R^2}+\frac{CA^3}{R^2}$. Choosing $A=\frac{\sqrt{R}}{1+|x|^{1/4}}$, we thus find (3.26), thereby concluding the proof of (3.21). ∎

4. Other trapping potentials. In the previous sections, we have studied a harmonic confinement, which is the case of most current experiments. As pointed out in [2], one could imagine a more general trapping potential, where in (1.2), $(1-\Omega^2)|x|^2/2$ is replaced by $(1-\Omega^2)|x|^2/2+W(x)$, and perform a similar analysis. Then, the limiting distribution replacing the inverted parabola should be

$$|\psi(x)|^2 = \left(\frac{\mu - \frac{1-\Omega^2}{2\Omega}|x|^2 - W\left(\frac{x}{\sqrt{\Omega}}\right)}{Gb\Omega}\right)_+, \tag{4.1}$$

where μ is such that $\int |\psi|^2 = 1$. There are two necessary conditions to apply our previous analysis: we need $E_{LLL}(\psi) - \Omega$ to be small and the extent of the condensate (where $|\psi|^2$ is nonzero) to be large. The first condition is required so that the lowest Landau level is indeed a good approximation. The second condition then allows one to write the wave function as the product of a rapidly oscillating function (which is averaged in the limit), multiplied by a slowly varying profile.

In recent experiments [13, 36], $W(x) = k|x|^4/4$. One can check that if $\Omega > \Omega_c = 1+\sqrt{\Delta}$, where $\Delta = (3k^2Gb/8\pi)^{2/3}$, then the limiting distribution (4.1) has its support in an annulus of inner and outer radii $R_\pm = 2(\Omega - 1 \pm \sqrt{\Delta})/k$. An interesting regime to study is when k is small and $\Omega - 1 = \alpha k^{2/3}$, with α such that $\Omega > \Omega_c$. Then the large scaling parameter replacing R is $k^{-1/6}$, which is the order of magnitude of $R_\pm$. The vortex lattice is located in the annulus (R_-, R_+) and is distorted towards the inner and outer edges, the inner disk corresponding to a giant vortex.

This approach does not allow one to study the case where Ω is large and the annulus gets thin [19]: vortices are arranged on concentric circles and there is only a few number of them in the condensate. The extent of the condensate is no longer much larger than a vortex cell so that we are no longer in the setting to use the averaging effect.

Acknowledgments. This work was motivated by many fruitful discussions with Jean Dalibard and the joint paper [2]. We would like to thank him warmly for all his comments and suggestions.

REFERENCES

[1] J. R. Abo-Shaeer, C. Raman, J. M. Vogels, and W. Ketterle, *Observation of vortex lattices in Bose-Einstein condensates*, Science, 292 (2001), pp. 476–479.

[2] A. Aftalion, X. Blanc, and J. Dalibard, *Vortex patterns in a fast rotating Bose-Einstein condensate*, Phys. Rev. A, 71 (2005), p. 023611.

[3] A. Aftalion and X. Blanc, *Reduced energy functionals for a three dimensional fast rotating Bose-Einstein condensate*, in preparation, 2006.

[4] A. Aftalion, X. Blanc, and F. Nier, *Lowest Landau level functional and Bargmann spaces for Bose-Einstein condensates*, to appear in J. Funct. Anal., 2006.

[5] A. Aftalion, X. Blanc, and F. Nier, *Vortex distribution in the lowest Landau level*, Phys. Rev. A, 73 (2006), p, 011601(R).

[6] A. Aftalion and R. L. Jerrard, *Properties of a single vortex solution in a rotating Bose Einstein condensate*, C. R. Acad. Sci. Paris, Ser. I, 336 (2003).

[7] A. Aftalion and T. Riviere, *Vortex energy and vortex bending for a rotating Bose-Einstein condensate*, Phys. Rev. A, 64 (2001), p. 043611.

[8] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[9] G. Baym and C. J. Pethick, *Vortex core structure and global properties of rapidly rotating Bose-Einstein condensates*, Phys. Rev. A, 69 (2004), p. 043619.

[10] F. Bethuel, H. Brezis, and F. Helein, *Ginzburg-Landau Vortices*, Progr. Nonlinear Differential Equations Appl. 13, Birkhäuser Boston, Inc., Boston, 1994.

[11] P. M. Bleher and J. L. Lebowitz, *Energy-level statistics of model quantum systems: Universality and scaling in a lattice-point problem*, J. Statist. Phys., 74 (1994), pp. 167–217.

[12] D. Butts and R. Rokhsar, *Predicted signatures of rotating Bose-Einstein condensates*, Nature, 397 (1999), p. 327.

[13] V. Bretin, S. Stock, Y. Seurin, and J. Dalibard, *Fast rotation of a Bose-Einstein condensate*, Phys. Rev. Lett., 92 (2004), p. 050403.

[14] I. Coddington, P. C. Haljan, P. Engels, V. Schweikhard, S. Tung, and E. A. Cornell, *Experimental studies of equilibrium vortex properties in a Bose-condensed gas*, Phys. Rev. A, 70 (2004), p. 063607.

[15] N. R. Cooper, S. Komineas, and N. Read, *Vortex lattices in the lowest Landau level for confined Bose-Einstein condensates*, Phys. Rev. A, 70 (2004) p. 033604.

[16] N. R. Cooper, N. K. Wilkin, and J. M. F. Gunn, *Quantum phases of vortices in rotating Bose-Einstein condensates*, Phys. Rev. Lett., 87 (2001), p. 120405.
[17] R. J. Donnelly, *Quantized Vortices in Helium* II, Cambridge University Press, Cambridge, UK, Chaps. 4 and 5, 1991.
[18] P. Engels, I. Coddington, P. C. Haljan, V. Schweikhard, and E. A. Cornell, *Observation of long-lived vortex aggregates in rapidly rotating Bose-Einstein condensates*, Phys. Rev. Lett., 90 (2003), p. 170405.
[19] A. L. Fetter, B. Jackson, and S. Stringari, *Rapid rotation of a Bose-Einstein condensate in a harmonic plus quartic trap*, Phys. Rev. A, 71 (2005), p. 013605.
[20] U. R. Fischer and G. Baym, *Vortex states of rapidly rotating dilute Bose-Einstein condensates*, Phys. Rev. Lett., 90 (2003), p. 140402.
[21] T. L. Ho, *Bose-Einstein condensates with large number of vortices*, Phys. Rev. Lett., 87 (2001), p. 060403.
[22] W. H. Kleiner, L. M. Roth, and S. H. Autler, *Bulk solution of Ginzburg-Landau equations for Type* II *Superconductors: Upper critical field region*, Phys. Rev., 133 (1964), p. A1226.
[23] L. D. Landau and E. M. Lifschitz, *Quantum Mechanics*, Pergamon, Oxford, 1965.
[24] E. H. Lieb and R. Seiringer, *Derivation of the Gross-Pitaevskii equation for rotating Bose gases*, Arxiv:math-ph/0504042, 2005.
[25] E. H. Lieb, R. Seiringer, and J. Yngvason, *Bosons in a trap: A rigorous derivation of the Gross-Pitaevskii energy functional*, Phys. Rev. A, 61 (2000), p. 0436021.
[26] K. Lu and X. B. Pan, *Eigenvalue problems of Ginzburg-Landau operator in bounded domains*, J. Math. Phys., 40 (1999), pp. 2647–2670.
[27] K. Madison, F. Chevy, V. Bretin, and J. Dalibard, *Vortex formation in a stirred Bose-Einstein condensate*, Phys. Rev. Lett., 84 (2000), p. 806.
[28] M. R. Matthews et al., *Vortices in a Bose-Einstein condensate*, Phys. Rev. Lett., 83 (1999), p. 2498.
[29] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[30] M. Olshanii, *Atomic scattering in the presence of an external confinement and a gas of impenetrable bosons*, Phys. Rev. Lett., 81 (1998), pp. 938–941.
[31] C. Raman, J. R. Abo-Shaeer, J. M. Vogels, K. Xu, and W. Ketterle, *Vortex nucleation in a stirred Bose-Einstein condensate*, Phys. Rev. Lett., 87 (2001), p. 210402.
[32] K. Schnee and J. Yngvason, *Bosons in Disc-Shaped Traps: From* 3*D to* 2*D*, preprint, 2004.
[33] V. Schweikhard, I. Coddington, P. Engels, V. P. Mogendorff, and E. A. Cornell, *Rapidly rotating Bose-Einstein condensates in and near the lowest Landau level*, Phys. Rev. Lett., 92 (2004), p. 040404.
[34] D. E. Sheehy and L. Radzihovsky, *Vortices in spatially inhomogeneous superfluids*, Phys. Rev. A, 70 (2004), p. 063620.
[35] J. Sinova, C. B. Hanna, and A. H. MacDonald, *Quantum melting and absence of Bose-Einstein condensation in two-dimensional vortex matter*, Phys. Rev. Lett., 89 (2002), p. 030403.
[36] S. Stock, V. Bretin, F. Chevy, and J. Dalibard, *Shape oscillation of a rotating Bose-Einstein condensate*, Europhys. Lett., 65 (2004), p. 594.
[37] G. Watanabe, G. Baym, and C. J. Pethick, *Landau levels and the Thomas-Fermi structure of rapidly rotating Bose-Einstein condensates*, Phys. Rev. Lett., 93 (2004), p. 190401.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 894–905

SPECTRAL CHARACTERIZATION OF THE TRACE SPACES $H^s(\partial\Omega)$*

GILES AUCHMUTY†

Abstract. This paper defines certain scales of trace spaces $H^s(\partial\Omega)$ using harmonic Steklov eigenfunction expansions. The approach is intrinsic and applies to bounded regions in $\mathbb{R}^n$ for which standard imbedding results hold. In particular it suffices that the boundary of the region be a finite disjoint union of Lipschitz surfaces. The definition generalizes the classical definitions that require the boundary to consist of smooth manifolds. The description depends on a special inner product on $H^1(\Omega)$, certain completeness theorems for Steklov eigenfunctions, and special properties of the harmonic Steklov eigenfunctions. The characterization provides explicit formulae for the inner products and norms of a function in $H^s(\partial\Omega)$ and allows the description of specific orthonormal bases for these spaces. For $s < 0$, the spaces are obtained by duality from the case for $s > 0$.

Key words. Hilbert trace spaces, Steklov eigenproblems, boundary trace operator, normal derivative operator

AMS subject classifications. 46E35, 31B25

DOI. 10.1137/050626053

1. Introduction. This paper develops a spectral characterization of real Hilbert trace spaces associated with bounded regions in $\mathbb{R}^n$ whose boundary may consist of Lipschitz surfaces—or slightly more general surfaces of finite area. The description uses harmonic Steklov eigenfunction expansions and provides concrete formulae for the inner products, norms, and orthonormal bases.

The approach used here is based on the use of special inner products and related decompositions for $H^1(\Omega)$. A remarkable property of the harmonic Steklov eigenfunctions is that they provide an orthogonal set in $H^1(\Omega)$ with respect to a natural inner product as well as a maximal orthogonal set in $L^2(\partial\Omega, d\sigma)$. This property is used to develop a description of $H^s(\partial\Omega)$ for all real s. The resulting spaces form an interpolatory family of spaces, and explicit formulae for the inner products in $H^{1/2}(\partial\Omega)$ and $H^{-1/2}(\partial\Omega)$ are obtained. Moreover it enables the use of Hilbert space approximation results using these harmonic Steklov eigenfunctions.

The usual theory of trace spaces as described in Adams and Fournier [2], Dautray and Lions [8], Lions and Magenes [12], and McLean [14] requires the use of local diffeomorphisms of domains onto a half-space. Then the allowable functions in the trace spaces are described either via properties of their Fourier transforms, or else using double integrals over the boundary of certain potential theoretic expressions. These definitions required various regularity assumptions for the boundary; in [8] and [12] they are required to be C^∞-manifolds.

Here an intrinsic definition of these spaces is given which does not require any mappings of the underlying domain. This approach permits weaker regularity of the

*Received by the editors March 5, 2005; accepted for publication (in revised form) February 7, 2006; published electronically September 5, 2006.

http://www.siam.org/journals/sima/38-3/62605.html

†Division of Mathematical Sciences, National Science Foundation, Arlington, VA 22230 (gauchmut @nsf.gov), and Department of Mathematics, University of Houston, Houston, TX 77204-3008.

boundary than does the conventional theory. We require only that condition (B2) of section 2 hold. In particular this is satisfied when (B1) holds. The results here depend on the characterization, and completeness, of Steklov eigenfunctions proved in Auchmuty [4].

2. Definitions and notation. This paper will develop a spectral characterization of the trace spaces $H^s(\partial\Omega)$, where Ω is a bounded region in $\mathbb{R}^n$, and show that these spaces are Hilbert spaces with respect to a natural inner product. A region is a nonempty, connected, open subset of $\mathbb{R}^n$. Its closure is denoted $\overline{\Omega}$ and its boundary is $\partial\Omega := \overline{\Omega} \setminus \Omega$. A standard assumption on the region is the following.

(B1) *Ω is a bounded region in $\mathbb{R}^n$ and its boundary $\partial\Omega$ is the union of a finite number of disjoint closed Lipschitz surfaces, each surface having finite surface area.*

When this holds there is an outward unit normal ν defined at σ for a.e. point of $\partial\Omega$. The definitions and terminology of Evans and Gariepy [10] will be followed except that $\sigma, d\sigma$ will represent the Hausdorff $(n-1)$-dimensional measure and integration with respect to this measure, respectively. All functions in this paper will take values in $\overline{\mathbb{R}} := [-\infty, \infty]$ and derivatives should be taken in a weak sense.

The real Lebesgue spaces $L^p(\Omega)$ and $L^p(\partial\Omega, d\sigma)$, $1 \le p \le \infty$, are defined in the standard manner and have the usual p-norm denoted by $\|u\|_p$ and $\|u\|_{p,\partial\Omega}$, respectively. Their inner products are defined by

$$\langle u, v\rangle := \int_\Omega u(x)\, v(x)\, dx \qquad \text{and} \qquad \langle u, v\rangle_{\partial\Omega} := |\partial\Omega|^{-1} \int_{\partial\Omega} u\, v\, d\sigma.$$

Let $H^1(\Omega)$ be the usual real Sobolev space of functions on Ω. It is a real Hilbert space under the standard H^1-inner product

$$[u, v]_1 := \int_\Omega [u(x).v(x) \;+\; \nabla u(x) \cdot \nabla v(x)]\, dx. \tag{2.1}$$

Here ∇u is the gradient of the function u and the corresponding norm is denoted $\|u\|_{1,2}$.

When Ω satisfies (B1), then the *Gauss–Green* theorem holds in the form

$$\int_\Omega u(x) D_j\, v(x)\, dx \;=\; \int_{\partial\Omega} u\, v\nu_j\, d\sigma \;-\; \int_\Omega v(x) D_j\, u(x)\, dx \quad \text{for } 1 \le j \le n \tag{2.2}$$

and for all $u,\ v$ in $H^1(\Omega)$.

The region Ω is said to satisfy *Rellich's theorem* provided the imbedding of $H^1(\Omega)$ into $L^p(\Omega)$ is compact for $1 \le p < p_S$, where $p_S(n) := 2n/(n-2)$ when $n \ge 3$, or $p_S(2) = \infty$ when $n = 2$.

There are a number of different criteria on Ω and $\partial\Omega$ that imply this result. When (B1) holds it is Theorem 1 in section 4.6 of [10]; see also Amick [1]. DiBenedetto [9], in Theorem 14.1 of Chapter 9, shows that the result holds when Ω is bounded and satisfies a "cone property." Adams and Fournier give a thorough treatment of conditions for this result in Chapter 6 of [2] and show that it also holds for some classes of unbounded regions.

When (B1) holds and $u \in W^{1,1}(\Omega)$, then the trace of u on $\partial\Omega$ is well-defined and is a Lebesgue integrable function with respect to σ; see [10, section 4.2] for details. The region Ω is said to satisfy a *compact trace theorem*, provided the trace mapping $\Gamma : H^1(\Omega) \to L^2(\partial\Omega, d\sigma)$ is compact. The trace map is the linear extension of the

map restricting Lipschitz continuous functions on $\overline{\Omega}$ to $\partial\Omega$. Occasionally u will be used in place of Γu for the trace of a function on $\partial\Omega$.

Evans and Gariepy [10, section 4.3] show that Γ is continuous when $\partial\Omega$ satisfies (B1). Theorem 1.5.1.10 of Grisvard [11] proves an inequality that implies the compact trace theorem when $\partial\Omega$ satisfies (B1). This inequality is also proved in [9, Chapter 9, section 18], under stronger regularity conditions on the boundary.

The results in this paper require only that the region satisfy the following condition.

(B2) *Ω is a bounded region with a boundary $\partial\Omega$ such that the Gauss–Green, Rellich, and compact trace theorems hold.*

Condition (B2) holds when (B1) does. Discussion of general regions for which (B2) holds may be found in Maz'ya and Poborchi [13], especially section 6.3, and also section 3 of [7]. Our analysis uses the following (equivalent) inner product on $H^1(\Omega)$:

$$[u,v]_\partial := \int_\Omega \nabla u \cdot \nabla v \; dx \; + \; |\partial\Omega|^{-1} \int_{\partial\Omega} u \; v \; d\sigma. \tag{2.3}$$

The corresponding norm will be denoted by $\|u\|_\partial$. The proof that this norm is equivalent to the usual $(1,2)$-norm on $H^1(\Omega)$ when (B2) holds is Corollary 6.2 of [4] and also is part of Theorem 21A of [15].

A function $u \in H^1(\Omega)$ is said to be *harmonic on* Ω, provided it is a solution of Laplace's equation in the usual weak sense. Namely,

$$\int_\Omega \nabla u \cdot \nabla \varphi \; dx \; = \; 0 \qquad \text{for all } \varphi \in C_c^1(\Omega). \tag{2.4}$$

Here $C_c^1(\Omega)$ is the set of all C^1-functions on Ω with compact support in Ω.

Define $\mathcal{H}(\Omega)$ to be the space of all such harmonic functions on Ω. When (B1) holds, the closure of $C_c^1(\Omega)$ in the H^1-norm is the usual Sobolev space $H_0^1(\Omega)$. Then (2.4) is equivalent to saying that $\mathcal{H}(\Omega)$ is ∂-orthogonal to $H_0^1(\Omega)$. This may be expressed as

$$H^1(\Omega) \; = \; H_0^1(\Omega) \oplus_\partial \mathcal{H}(\Omega), \tag{2.5}$$

where $\oplus_\partial$ indicates that this is a ∂-orthogonal decomposition. This result is also discussed in section 22.4 of [15].

In this paper we shall use various standard results from the calculus of variations and convex analysis. Background material on such methods may be found in Blanchard and Brüning [6] or Zeidler [16], both of which have discussions of the variational principles for the Dirichlet eigenvalues and eigenfunctions of second order elliptic operators. The variational principles used here are variants of the principles described there and are analogous to those for the Laplacian described in section 5 of Auchmuty [3].

All the variational principles and functionals in this paper will be defined on (closed convex subsets of) $H^1(\Omega)$. When $\mathcal{F} : H^1(\Omega) \to (-\infty, \infty]$ is a functional, then $\mathcal{F}$ is said to be *G-differentiable* at a point $u \in H^1(\Omega)$ if there is an $\mathcal{F}'(u)$ such that

$$\lim_{t\to 0} \; t^{-1} \left[\mathcal{F}(u+tv) - \mathcal{F}(u)\right] \; = \; \mathcal{F}'(u)(v) \qquad \text{for all } v \in H^1(\Omega),$$

with $\mathcal{F}'(u)$ a continuous linear functional on $H^1(\Omega)$. In this case, $\mathcal{F}'(u)$ is called the G-derivative of $\mathcal{F}$ at u.

In this paper we will say that a real sequence $\{x_m : m \geq 1\}$ is said to be increasing if $x_{m+1} \geq x_m$ for all m; it is strictly increasing if strict inequality holds here for all m. Similarly a function u is said to be (strictly) positive on a set E, if $u(x) \geq (>)\ 0$ on E.

3. The harmonic Steklov eigenproblem. Assume Ω is a region in $\mathbb{R}^n$ which satisfies (B2). A nonzero function $s \in H^1(\Omega)$ is said to be a *harmonic Steklov eigenfunction* on Ω corresponding to the Steklov eigenvalue δ, provided s satisfies

$$\int_\Omega \nabla s \cdot \nabla v \, dx \;=\; \delta \, |\partial\Omega|^{-1} \int_{\partial\Omega} s \, v \, d\sigma \qquad \text{for all} \quad v \in H^1(\Omega). \tag{3.1}$$

This is the weak form of the boundary value problem

$$\Delta \, s \;=\; 0 \qquad \text{on } \Omega \quad \text{with } D_\nu \, s \;=\; \delta \, |\partial\Omega|^{-1} \, s \quad \text{on } \partial\Omega. \tag{3.2}$$

Here Δ is the Laplacian and $D_\nu \, s(x) := \nabla s(x) \cdot \nu(x)$ is the unit outward normal derivative of s at a point on the boundary.

$\delta_0 = 0$ is the least eigenvalue of this problem corresponding to the eigenfunction $s_0(x) \equiv 1$ on Ω. This eigenvalue is simple as Ω is connected. All other eigenvalues of (3.1) are strictly positive.

These eigenvalues and a corresponding family of ∂-orthonormal eigenfunctions may be found using variational principles as described in sections 6 and 7 of [4]. A different variational description is developed in Bandle [5, Chapter 3]. Let the first k Steklov eigenvalues be $0 = \delta_0 < \delta_1 \leq \delta_2 \leq \cdots \leq \delta_{k-1}$ and let $s_0, s_1, \ldots, s_{k-1}$ be a corresponding set of ∂-orthonormal eigenfunctions. The kth eigenfunction s_k will be a maximizer of the functional

$$\mathcal{B}(u) \;:=\; |\partial\Omega|^{-1} \int_{\partial\Omega} u^2 \, d\sigma \tag{3.3}$$

over the subset B_k of functions in $H^1(\Omega)$ which satisfy

$$\|s\|_\partial \;\leq\; 1 \qquad \text{and} \qquad [s, s_l]_\partial \;=\; 0 \quad \text{for} \quad 0 \leq l \leq k-1. \tag{3.4}$$

The existence and some properties of these eigenfunctions are described in sections 6 and 7 of [4]. In particular, that analysis shows that each δ_j is of finite multiplicity and $\delta_j \to \infty$ as $j \to \infty$; see Theorem 7.2 of [4]. Let $\mathcal{S} := \{s_j : j \geq 0\}$ be the maximal family of ∂-orthonormal eigenfunctions constructed inductively as above. For each $u \in H^1(\Omega)$, consider the series

$$P_H \, u(x) \;:=\; \sum_{j=0}^{\infty} [u, s_j]_\partial \; s_j(x). \tag{3.5}$$

THEOREM 3.1. *Assume $\Omega, \partial\Omega$ satisfy* (B2) *and P_H is defined by* (3.5)*; then P_H is the ∂-orthogonal projection of $H^1(\Omega)$ onto $\mathcal{H}(\Omega)$.*

Proof. This follows from standard results about orthogonal expansions and Theorem 7.3 of [4], which says that $\mathcal{S}$ is a maximal orthonormal subset of $\mathcal{H}(\Omega)$. □

An expression of the form

$$v(x) \;:=\; \sum_{j=0}^{\infty} c_j \, s_j(x) \qquad \text{with } c_j := [v, s_j]_\partial \tag{3.6}$$

will be called a harmonic Steklov expansion and, since $\mathcal{S}$ is a basis of $\mathcal{H}(\Omega)$, the Riesz–Fischer theorem implies that it represents an H^1-harmonic function on Ω if and only if

$$\sum_{j=0}^{\infty} |c_j|^2 \ < \ \infty. \tag{3.7}$$

As described in section 8 of [4], the Steklov eigenfunctions on the unit disc in $\mathbb{R}^2$ are the functions $r^k \cos k\theta$ and $r^k \sin k\theta$, so the above series are familiar from classical treatments of harmonic functions on a disc. Similarly when Ω is the unit ball in $\mathbb{R}^3$ the Steklov eigenfunctions are spherical harmonics and these series generalize some common expansions used in classical mathematical physics.

4. A spectral representation of the trace operator. The Steklov eigenfunctions s_j described in the preceding section have L^2 traces on the boundary $\partial\Omega$ whenever (B1) holds. Define

$$\hat{s_j}(x) \ := \ \sqrt{1+\delta_j}\ \Gamma\, s_j(x) \qquad \text{for } x \in \partial\Omega, \ \text{ and } \ j \geq 0. \tag{4.1}$$

Then (3.1) and (3.4) imply that the set $\hat{\mathcal{S}} := \{\hat{s_j} : j \geq 0\}$ will be an orthonormal set in $L^2(\partial\Omega, d\sigma)$ with respect to the inner product defined in section 2. The following result provides an explicit expression for the trace operator in terms of the harmonic Steklov expansion of a function $u \in H^1(\Omega)$.

THEOREM 4.1. *Assume $\Omega, \partial\Omega$ satisfy* (B2)*, with $\Gamma, \hat{\mathcal{S}}$ as above. Then $\hat{\mathcal{S}}$ is a maximal orthonormal set in $L^2(\partial\Omega, d\sigma)$ and*

$$\Gamma u \ = \ \sum_{j=0}^{\infty} (1+\delta_j)^{-1/2}\ [u, s_j]_\partial\ \hat{s_j} \quad \textit{for each } u \in H^1(\Omega). \tag{4.2}$$

Proof. The first claim is a special case of Theorem 9.4 in [4]. The null space of the operator Γ is $H_0^1(\Omega)$ from Theorem 3.40 of [14]. Hence, from (2.5) and Theorem 3.1, $\Gamma u = \Gamma P_H\, u$, where P_H is the projection onto the space $\mathcal{H}(\Omega)$. Equation (4.2) then follows from (4.1). □

Apply Parseval's identity to (4.2). Then

$$\|\Gamma\, u\|^2_{\partial\Omega} \ := \ |\partial\Omega|^{-1} \int_{\partial\Omega} |\Gamma\, u|^2\, d\sigma \ = \ \sum_{j=0}^{\infty} (1+\delta_j)^{-1}\ [u, s_j]_\partial{}^2 \tag{4.3}$$

for any $u \in H^1(\Omega)$, since $\hat{\mathcal{S}}$ is a basis of $L^2(\partial\Omega, d\sigma)$.

Suppose now that $g = \Gamma\, u$ for some $u \in H^1(\Omega)$. Then $g \in L^2(\partial\Omega, d\sigma)$ and

$$g(x) \ = \ \sum_{j=0}^{\infty} g_j\ \hat{s_j}(x) \qquad \text{with } g_j \ = \ \langle g, \hat{s_j}\rangle_{\partial\Omega}. \tag{4.4}$$

Equation (4.2) and the orthonormality of $\hat{\mathcal{S}}$ imply that

$$[u, s_j]_\partial \ = \ (1+\delta_j)^{1/2}\ g_j \qquad \text{for all } j \geq 0.$$

Let g_M be the Mth partial sum of the series (4.4) and consider the map $E_M : L^2(\partial\Omega, d\sigma) \to \mathcal{H}(\Omega)$ defined by

$$E_M\, g(x) \ := \ \sum_{j=0}^{M} (1+\delta_j)^{1/2}\ g_j\, s_j(x). \tag{4.5}$$

This is a harmonic function on Ω with boundary trace g_M. Define

$$E\,g(x) \;:=\; \lim_{M\to\infty} E_M\,g(x). \tag{4.6}$$

Then (3.7) shows that Eg is in $\mathcal{H}(\Omega)$ if and only if

$$\sum_{j=0}^{\infty} (1+\delta_j)\,|g_j|^2 \;<\; \infty. \tag{4.7}$$

Formally E defines a *harmonic extension* of the boundary data g to Ω.

5. A spectral definition of $H^s(\partial\Omega)$. The classical description of trace theorems for H^1-functions on a region requires the description of the boundary using local coordinates and mappings from canonical regions such as a half-space. See chapter 3 of McLean [14] for a detailed description under weak regularity conditions. A comparison of a number of methods for defining these spaces is given in the appendix to Chapter 4 of [8]. Here a very different definition will be described which should be much more useful for approximation theory and computational purposes. It is based on the use of a scale of spaces defined by weights that depend on the Steklov eigenvalues.

Specifically $H^s(\partial\Omega)$ is defined as that subspace of $L^2(\partial\Omega, d\sigma)$ of functions whose harmonic Steklov coefficients satisfy certain summability conditions. For $s \geq 0$, we define $H^s(\partial\Omega)$ to be the subspace of all functions $g \in L^2(\partial\Omega, d\sigma)$ with Steklov expansion (4.4) satisfying

$$\sum_{j=0}^{\infty} (1+\delta_j)^{2s}\,|g_j|^2 \;<\; \infty. \tag{5.1}$$

Define the *s-inner product* and *s-norm* on $H^s(\partial\Omega)$ by

$$[g,h]_{s,\partial\Omega} \;:=\; \sum_{j=0}^{\infty} (1+\delta_j)^{2s}\,g_j\,h_j \qquad \text{and} \qquad \|g\|^2_{s,\partial\Omega} \;:=\; \sum_{j=0}^{\infty} (1+\delta_j)^{2s}\,{g_j}^2. \tag{5.2}$$

When $s = 0$, one sees that $H^0(\partial\Omega) = L^2(\partial\Omega, d\sigma)$.

When $s = 1/2$, (4.7) shows that the space $H^{1/2}(\partial\Omega)$ will be precisely the class of all boundary values of H^1-functions on Ω—so this definition agrees with the classical definition based on Fourier methods when $\partial\Omega$ is a smooth manifold. Note that this definition of the spaces $H^s(\partial\Omega)$ requires only that $\partial\Omega$ be smooth enough for the Steklov eigenanalysis to hold.

For general $s > 0$, this definition corresponds to the spaces described via (complex) interpolation. The following results show that this definition satisfies the same intermediate space properties as the original definitions of Lions and Magenes [12, Chapter 1, section 7.3], which required that the boundary comprise C^∞-manifolds.

THEOREM 5.1. *Assume that $\Omega, \partial\Omega$ satisfy* (B2) *and $H^s(\partial\Omega)$ is defined as above. If $0 \leq s_1 < s_2$, then $H^{s_2}(\partial\Omega)$ is a dense subspace of $H^{s_1}(\partial\Omega)$ and the imbedding of $H^{s_2}(\partial\Omega)$ into $H^{s_1}(\partial\Omega)$ is compact.*

Proof. For $M \geq 1$, let $P_M : L^2(\partial\Omega, d\sigma) \to L^2(\partial\Omega, d\sigma)$ be the finite rank operator corresponding to the Mth partial sum of the Steklov expansion (4.4). That is,

$$P_M\,g(x) \;:=\; \sum_{j=0}^{M} \langle g, \hat{s_j}\rangle_{\partial\Omega}\;\hat{s_j}(x) \qquad \text{for } g \in L^2(\partial\Omega, d\sigma). \tag{5.3}$$

Obviously $P_M\, g \in H^s(\partial\Omega)$ for all $s \geq 0$ and the definition (5.2) yields that

$$\|g\|_{s_1,\partial\Omega} \;\leq\; \|g\|_{s_2,\partial\Omega} \qquad \text{whenever} \quad 0 \leq s_1 < s_2. \tag{5.4}$$

Given $g \in H^{s_1}(\partial\Omega)$, then the sequence $\{P_M\, g : M \geq 1\}$ is a subset of $H^{s_2}(\partial\Omega)$ which converges to g in $H^{s_1}(\partial\Omega)$. Thus $H^{s_2}(\partial\Omega)$ is a dense subspace of $H^{s_1}(\partial\Omega)$.

Consider the linear map $L_\theta : L^2(\partial\Omega, d\sigma) \to L^2(\partial\Omega, d\sigma)$ defined by

$$L_\theta\, g(x) \;:=\; \sum_{j=0}^{\infty} \,(1+\delta_j)^{-\theta}\; \langle g, \hat{s_j}\rangle_{\partial\Omega}\; \hat{s_j}(x). \tag{5.5}$$

For $\theta > 0$, using the fact that $\delta_j \to \infty$, L_θ is a compact linear operator, as it may be uniformly approximated by a finite rank operator. Moreover

$$\|L_\theta\, g\|^2_{s,\partial\Omega} \;=\; \sum_{j=0}^{\infty} \,(1+\delta_j)^{2(s-\theta)}\; \langle g, \hat{s_j}\rangle^2_{\partial\Omega}. \tag{5.6}$$

Thus L_θ is a linear isometry of $L^2(\partial\Omega, d\sigma)$ onto $H^\theta(\partial\Omega)$, so the imbedding of $H^s(\partial\Omega)$ into $L^2(\partial\Omega, d\sigma)$ is compact whenever $s > 0$. A translation in s then yields that the imbedding of $H^{s_2}(\partial\Omega)$ into $H^{s_1}(\partial\Omega)$ is compact whenever $s_1 < s_2$. □

The family of spaces $H^s(\partial\Omega)$ with $s \geq 0$ forms an interpolatory family (or scale) of real Hilbert spaces, as these s-norms satisfy the following log-convexity inequality.

THEOREM 5.2. *Assume that $\Omega, \partial\Omega$ satisfy* (B2) *and $H^s(\partial\Omega)$ is defined as above. If $0 \leq s_1 < s_2$ and $s = (1-\theta)s_1 + \theta s_2$ with $0 \leq \theta \leq 1$, then*

$$\|g\|_{s,\partial\Omega} \;\leq\; \|g\|^{1-\theta}_{s_1,\partial\Omega}\; \|g\|^{\theta}_{s_2,\partial\Omega} \qquad \textit{for all} \quad g \in H^{s_2}(\partial\Omega). \tag{5.7}$$

Proof. This is obviously true when $\theta = 0$ or 1. Assume $0 < \theta < 1$. Then from (5.2),

$$\|g\|^2_{s,\partial\Omega} \;:=\; \sum_{j=0}^{\infty} \,(1+\delta_j)^{2s}\, {g_j}^2. \tag{5.8}$$

Factor each term in the sum, so that $(1+\delta_j)^{2s}\, {g_j}^2 \;=\; c_j\, d_j$ with

$$c_j := (1+\delta_j)^{s_1(1-\theta)}\, {g_j}^{2(1-\theta)}, \qquad d_j := (1+\delta_j)^{s_2\theta}\, {g_j}^{2\theta}.$$

Apply Holder's inequality to (5.8) with $p := 1/(1-\theta), p^* := \theta^{-1}$. Then inequality (5.7) follows. □

Suppose F is a continuous linear functional on $H^s(\partial\Omega)$ with $s \;\geq\; 0$. F will be *represented by* a function $f \in L^2(\partial\Omega, d\sigma)$, provided

$$F(g) \;=\; |\partial\Omega|^{-1} \int_{\partial\Omega} f\, g\; d\sigma \;=\; \langle f, g\rangle_{\partial\Omega} \qquad \text{for all} \quad g \in H^s(\partial\Omega). \tag{5.9}$$

When f, g have Steklov expansions of the form (4.4) with Steklov coefficients f_j, g_j, then this becomes

$$F(g) \;=\; \sum_{j=0}^{\infty} \, f_j\, g_j. \tag{5.10}$$

Use of Schwarz's inequality here shows that each $f \in L^2(\partial\Omega, d\sigma)$ represents a continuous linear functional on such $H^s(\partial\Omega)$.

For $s < 0$, define the space $H^s(\partial\Omega)$ to be the completion of the space $L^2(\partial\Omega, d\sigma)$ with respect to the inner product and norm of (5.2). Elements of this space will be called *generalized functions* on $\partial\Omega$. Below we shall show that $H^{-s}(\partial\Omega)$ is precisely the dual space of $H^s(\partial\Omega)$ with respect to the pairing induced by the L^2-inner product on $\partial\Omega$. When $\partial\Omega$ is a C^∞-manifold, these will be spaces of distributions on $\partial\Omega$.

It is straightforward to verify that $H^s(\partial\Omega)$ with $s < 0$ is a real Hilbert space under the inner product of (5.2). The following theorem specifies the duality relationship.

THEOREM 5.3. *Assume that* $\Omega, \partial\Omega$ *satisfy* (B2), $H^s(\partial\Omega)$ *is defined as above with* $s > 0$, *and* F *is a continuous linear functional on* $H^s(\partial\Omega)$. *Then there is a unique generalized function* $f \in H^{-s}(\partial\Omega)$ *such that*

$$F(g) \;=\; \langle f, g\rangle_{\partial\Omega} \qquad \textit{for all} \quad g \in H^s(\partial\Omega). \tag{5.11}$$

Moreover the dual norm of F *is* $\|f\|_{-s,\partial\Omega}$.

Proof. Rewrite each term in the sum (5.10) as the product of

$$c_j := \mu_j^{-1}\, f_j \quad \text{and} \quad d_j := \mu_j\, g_j \qquad \text{with} \quad \mu_j := (1+\delta_j)^s.$$

Apply Schwarz's inequality to (5.10). Then the definitions of the norms yield

$$|F(g)| \;\le\; \|f\|_{-s,\partial\Omega}\, \|g\|_{s,\partial\Omega}. \tag{5.12}$$

Moreover equality holds here whenever $f_j = (1+\delta_j)^2\, g_j$ for all $j \ge 0$. Since F is continuous if and only if it is bounded, we see that each continuous linear functional on $H^s(\partial\Omega)$ will be represented by a generalized function in $H^{-s}(\partial\Omega)$. The dual norm is defined by

$$\|F\|_{*s} \;:=\; \sup_{\|g\|_{s,\partial\Omega}\,\le\,1} |F(g)|,$$

so (5.12) shows that it is given by the norm on $H^{-s}(\partial\Omega)$. ∎

6. Spectral representation of the normal derivative operator. When $\partial\Omega$ is locally a C^1-manifold, then the exterior unit normal ν is a continuous vector field. The exterior normal derivative of a C^1-function u on $\overline{\Omega}$ is then given by

$$D_\nu\, u(x) \;:=\; \nabla u(x) \cdot \nu(x) \qquad \text{for } x \in \partial\Omega. \tag{6.1}$$

When the Steklov eigenfunctions are sufficiently smooth, then (3.2) and (4.1) imply that

$$D_\nu\, s_j(x) \;=\; \frac{\delta_j}{|\partial\Omega|\,\sqrt{(1+\delta_j)}}\; \hat{s_j}(x) \qquad \text{for } x \in \partial\Omega \text{ and } j \ge 0. \tag{6.2}$$

Take this to hold for each Steklov eigenfunction. When $v \in \mathcal{H}(\Omega)$ has a Steklov expansion of the form (3.6), define the linear extension of D_ν to be

$$D_\nu\, v \;=\; |\partial\Omega|^{-1} \sum_{j=1}^{\infty} \frac{\delta_j}{\sqrt{(1+\delta_j)}}\; [v, s_j]_\partial\; \hat{s_j}. \tag{6.3}$$

The H^s-norm of this generalized function on $\partial\Omega$ is

$$\|D_\nu\, v\|^2_{s,\partial\Omega} \;=\; |\partial\Omega|^{-2}\; \sum_{j=1}^{\infty}\; (1+\delta_j)^{2s-1}{\delta_j}^2\; [v, s_j]^2_\partial. \tag{6.4}$$

In view of this calculation, this operator satisfies the following.

THEOREM 6.1. *Assume that $\Omega, \partial\Omega$ satisfy* (B2)*, and $H^s(\partial\Omega)$ is defined in section* 5*. Then the operator D_ν defined by* (6.3) *is a continuous map from $\mathcal{H}(\Omega)$ to $H^s(\partial\Omega)$ for $s \le -1/2$.*

Proof. From (6.4), the operator will be continuous if and only if there is a constant $C > 0$ such that

$$(1+\delta_j)^{2s-1}{\delta_j}^2 \;\le\; C \qquad \text{for all } j \ge 0.$$

Since $\delta_j \to \infty$, this holds if and only if $s \le -1/2$. □

7. Explicit inner product on $H^{1/2}(\partial\Omega)$. When $f, g \in H^{1/2}(\partial\Omega)$, the inner product on $H^{1/2}(\partial\Omega)$ was defined in section 5 in terms of a Steklov series expansion. Here it will be shown to have an expression in terms of the boundary trace and a normal derivative.

Given $g \in H^{1/2}(\partial\Omega)$, let Eg be its harmonic extension in $\mathcal{H}(\Omega)$ defined by (4.6). Then the outward normal derivative $D_\nu\, Eg$ will be in $H^{-1/2}(\partial\Omega)$ from Theorem 6.1.

THEOREM 7.1. *Assume that $\Omega, \partial\Omega$ satisfy* (B2) *and $H^{1/2}(\partial\Omega)$ is defined as above. Then E is a linear isometry from $H^{1/2}(\partial\Omega)$ to $\mathcal{H}(\Omega)$ and*

$$[f, g]_{1/2,\partial\Omega} \;=\; \langle f, g \,+\, |\partial\Omega|\, D_\nu\, Eg\rangle_{\partial\Omega} \qquad \textit{for all } f, g \in H^{1/2}(\partial\Omega). \tag{7.1}$$

Proof. From (4.6),

$$E\, g(x) \;=\; \sum_{j=0}^{\infty}\; (1+\delta_j)^{1/2}\;\; g_j\, s_j(x),$$

so

$$\|Eg\|^2_\partial \;=\; \sum_{j=0}^{\infty} (1+\delta_j)\, |g_j|^2 \;=\; \|g\|^2_{1/2,\partial\Omega} \qquad \text{for all } g \in H^{1/2}(\partial\Omega). \tag{7.2}$$

Hence E is an isometry as claimed. Substitute for Eg in (6.3). Then

$$D_\nu\, E\, g(x) \;=\; |\partial\Omega|^{-1}\; \sum_{j=0}^{\infty}\; \delta_j\; g_j\, \hat{s_j}(x).$$

This and the orthonormality of $\hat{\mathcal{S}}$ yield that

$$\langle f, g \,+\, |\partial\Omega|\, D_\nu\, Eg\rangle_{\partial\Omega} \;=\; \sum_{j=0}^{\infty}\; f_j\, (1+\delta_j)\, g_j, \tag{7.3}$$

which is (7.1). □

This result (7.1) may be written formally as

$$[f, g]_{1/2,\partial\Omega} \;=\; \int_{\partial\Omega}\; f\, (|\partial\Omega|^{-1}\, g \,+\, D_\nu g)\; d\sigma,$$

so the (1/2)-norm is defined by the quadratic form

$$\|g\|^2_{1/2,\partial\Omega} \;=\; \int_{\partial\Omega} \left[\, |\partial\Omega|^{-1}\, g^2 \;+\; g\, D_\nu g \,\right]\, d\sigma. \tag{7.4}$$

That is, $H^{1/2}(\partial\Omega)$ is the space of all functions in $L^2(\partial\Omega, d\sigma)$ for which this quadratic form is finite. Here $D_\nu g$ is actually the outward normal derivative of the harmonic extension of g to Ω.

8. The inner product on $H^{-1/2}(\partial\Omega)$. The space $H^{-1/2}(\partial\Omega)$ was defined as the completion of $L^2(\partial\Omega, d\sigma)$ with respect to the inner product defined by (5.2) with $s = -1/2$. In this section, this inner product will be characterized in terms of the solution of a Robin boundary value problem for Laplace's equation. More specifically, it will be described using a variational principle for such solutions.

Given $g \in H^{-1/2}(\partial\Omega)$, define the functional $\mathcal{D} : H^1(\Omega) \to \mathbb{R}$ by

$$\mathcal{D}(u) \;:=\; \int_\Omega |\nabla u|^2 \, dx \;+\; |\partial\Omega|^{-1} \int_{\partial\Omega} |\Gamma u|^2 \, d\sigma \;-\; 2\, \langle g, \Gamma u\rangle_{\partial\Omega}. \tag{8.1}$$

Consider the variational principle of minimizing $\mathcal{D}$ on $H^1(\Omega)$. The essential results about this problem can be stated as follows.

THEOREM 8.1. *Assume that $\Omega, \partial\Omega$ satisfy* (B2)*, $g \in H^{-1/2}(\partial\Omega)$, and $\mathcal{D}$ is defined by* (8.1)*. Then there is a unique minimizer $\hat{u}$ of $\mathcal{D}$ on $H^1(\Omega)$ and it satisfies*

$$\int_\Omega \nabla u \cdot \nabla v \; dx \;+\; \langle (\Gamma\, u \;-\; g), \Gamma\, v\rangle_{\partial\Omega} \;=\; 0 \qquad \textit{for all} \quad v \in H^1(\Omega). \tag{8.2}$$

Proof. The existence of a unique minimizer of $\mathcal{D}$ on $H^1(\Omega)$ is Theorem 9.2 of [4] with $\tau = 1/2$ and $g \in L^2(\partial\Omega, d\sigma)$. The extension to $g \in H^{-1/2}(\partial\Omega)$ is straightforward.

This functional $\mathcal{D}$ is convex and G-differentiable on $H^1(\Omega)$ and its derivative can be obtained from Theorems 3.1 and 6.1 of [4]. The left-hand side of (8.2) is the directional derivative of $\mathcal{D}$, so the extremality conditions imply that $\hat{u}$ will be a minimizer of $\mathcal{D}$ on $H^1(\Omega)$ if and only if it is a solution of (8.2). □

Note that (8.2) is the weak form of the Robin boundary value problem

$$\Delta\, u \;=\; 0 \qquad \text{on } \Omega \quad \text{ with } |\partial\Omega|\, D_\nu\, u \;+\; u \;=\; g \quad \text{on } \partial\Omega. \tag{8.3}$$

When $v \in H^1_0(\Omega)$, then $\Gamma v \equiv 0$, so (8.2) implies that (2.4) holds or the solution $\hat{u}$ is harmonic on Ω. Let $\hat{u}$ have a Steklov expansion of the form

$$\hat{u}(x) \;:=\; \sum_{j=0}^{\infty} \; u_j\, s_j(x) \qquad \text{on } \Omega. \tag{8.4}$$

Then, from (4.2), the boundary trace $\Gamma\, \hat{u}$ is given by

$$\Gamma\, \hat{u} \;:=\; \sum_{j=0}^{\infty} \; \frac{u_j}{\sqrt{(1+\delta_j)}}\, \hat{s_j}. \tag{8.5}$$

Assume that $g \in H^{-1/2}(\partial\Omega)$ has the Steklov representation

$$g \;:=\; \sum_{j=0}^{\infty} \; g_j\, \hat{s_j} \qquad \text{with } g_j \;:=\; \langle g, \hat{s_j}\rangle_{\partial\Omega}. \tag{8.6}$$

Substitute $v = s_0$ in (8.2) to find that $u_0 = g_0$.

For $k \geq 1$, put $v = s_k$ in (3.1). Then

$$\int_\Omega \nabla u \cdot \nabla s_k \, dx = \delta_k \, \langle s_k, \Gamma u \rangle_{\partial\Omega} \qquad \text{for all} \quad u \in H^1(\Omega). \tag{8.7}$$

Substitute this in (8.2) with $v = s_k$, to obtain

$$(1 + \delta_k) \, \langle \Gamma \hat{u}, \hat{s_k} \rangle_{\partial\Omega} = \langle g, \hat{s_k} \rangle_{\partial\Omega} \qquad \text{for all } k \geq 1. \tag{8.8}$$

The expression (8.5) for $\Gamma\hat{u}$ yields that the Steklov coefficients u_k of the solution of this variational problem are given by

$$u_k = (1 + \delta_k)^{-1/2} \, g_k \qquad \text{for each } k \geq 0. \tag{8.9}$$

Define $G_R : H^{-1/2}(\partial\Omega) \to \mathcal{H}(\Omega)$ to be the solution operator of this variational problem. Equations (8.4) and (8.9) show that G_R has the Steklov spectral representation

$$\hat{u}(x) := G_R \, g(x) = \sum_{k=0}^{\infty} (1 + \delta_k)^{-1/2} \, g_k \, s_k(x) \tag{8.10}$$

for any $g \in H^{-1/2}(\partial\Omega)$ as in (8.6). The boundary trace of this function will be

$$\Gamma G_R \, g = \sum_{k=0}^{\infty} (1 + \delta_k)^{-1} \, g_k \, \hat{s_k}. \tag{8.11}$$

Moreover a straightforward computation shows that

$$\|\Gamma G_R \, g\|_{1/2,\partial\Omega} = \|g\|_{-1/2,\partial\Omega}, \tag{8.12}$$

so this operator ΓG_R is an isometric linear mapping of $H^{-1/2}(\partial\Omega)$ onto $H^{1/2}(\partial\Omega)$. More generally ΓG_R will be an isometry from any space $H^s(\partial\Omega)$ onto $H^{s+1}(\partial\Omega)$.

This, together with the orthonormality of $\hat{\mathcal{S}}$, proves the following theorem.

THEOREM 8.2. *Assume that $\Omega, \partial\Omega$ satisfy* (B2) *and G_R is the operator defined by* (8.11). *Then the inner product on $H^{-1/2}(\partial\Omega)$ obeys*

$$[f, g]_{-1/2,\partial\Omega} = \langle f, \Gamma \, G_R \, g \rangle_{\partial\Omega} \qquad \textit{for all } f, g \in H^{-1/2}(\partial\Omega). \tag{8.13}$$

For other negative values of s, the inner products on $H^s(\partial\Omega)$ may be defined using fractional powers of the operator ΓG_R.

REFERENCES

[1] C. J. AMICK, *Some remarks on Rellich's theorem and the Poincaré inequality*, J. London Math. Soc., 18 (1973), pp. 81–93.

[2] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev Spaces*, 2nd ed., Academic Press, New York, 2003.

[3] G. AUCHMUTY, *The main inequality of vector analysis*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1–25.

[4] G. AUCHMUTY, *Steklov eigenproblems and the representation of solutions of elliptic boundary value problems*, Numer. Funct. Anal. Optim., 25 (2004), pp. 321–348.

[5] C. BANDLE, *Isoperimetric Inequalities and Applications*, Pitman, London, 1980.

[6] P. Blanchard and E. Brüning, *Variational Methods in Mathematical Physics*, Springer-Verlag, Berlin, 1992.
[7] D. Daners, *Robin boundary value problems on arbitrary domains*, Trans. Amer. Math. Soc., 352 (2002), pp. 4207–4236.
[8] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Springer-Verlag, Berlin, 1988.
[9] E. DiBenedetto, *Real Analysis*, Birkhäuser, Boston, 2001.
[10] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[12] J. L. Lions and E. Magenes, *Problemes aux limites non homogenes et applications*, Vol. 1, Dunod, Paris, 1968.
[13] V. Maz'ya and S. Poborchi, *Differentiable Functions on Bad Domains*, World Scientific, River Edge, NJ, 1997.
[14] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[15] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*. II/A. *Linear Monotone Operators*, Springer-Verlag, New York, 1990.
[16] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*. III. *Variational Methods and Optimization*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 906–920

VARIATIONAL PRINCIPLES FOR WATER WAVES*

B. KOLEV† AND D. H. SATTINGER‡

Abstract. We describe the Hamiltonian structures, including the Poisson brackets and Hamiltonians, for free boundary problems for incompressible fluid flows with vorticity. The Hamiltonian structure is used to obtain variational principles for stationary gravity waves for both irrotational flows and flows with vorticity.

Key words. water waves, variational principles, Poisson structures

AMS subject classifications. 35Q20, 35Q35

DOI. 10.1137/050647220

1. Introduction.

In 1933 Friedrichs [9] proposed the functional

$$J(\psi) = \iiint\limits_{0\le\psi\le 1} \left[(\nabla\psi)^2 + v^2(x,y)\right]\, d^2\mathbf{x},$$

where ψ is the stream function for an incompressible flow, as a variational method of obtaining solutions to free boundary value problems. Critical points of J are harmonic functions which satisfy the condition

$$(\nabla\psi)^2 = v^2$$

on the free boundary, given by $\psi = 1$. The free boundary condition relevant to theory of gravity waves, however, is the Bernoulli equation

$$\frac{(\nabla\varphi)^2}{2} + g\zeta = \text{constant},$$

where φ is either the velocity potential for irrotational flow or the stream function in the case of flows with vorticity. Thus some other variational principle is needed for the study of gravity waves.

Recently, a variational principle for gravity waves with vorticity was given by Constantin, Sattinger, and Strauss [7], using a direct, "hands-on" approach. More generally, a variational principle for a stationary wave may be obtained for systems possessing a Hamiltonian structure by minimizing the Hamiltonian computed in a Galilean frame moving with the wave. We illustrate that approach in this study.

We begin with a brief review of Euler's equations of incompressible flows and the associated free boundary value problems; in section 3 we describe the Hamiltonian structure of these problems, for irrotational flows and flows with vorticity, as given by Lewis et al. [15]. All the functions under consideration in this article, including the free boundaries, are assumed to be smooth.

*Received by the editors December 11, 2005; accepted for publication (in revised form) March 21, 2006; published electronically October 10, 2006.
http://www.siam.org/journals/sima/38-3/64722.html
†Centre de Mathématiques et Informatique, Université de Provence, Marseille, France (bkolev@wanadoo.fr).
‡Department of Mathematics, Yale University, New Haven, CT 06520 (david.sattinger@yale.edu).

2. Incompressible fluid flows. Let the velocity field of an incompressible fluid in a fixed region $\mathcal{D}$ be denoted by $\mathbf{v}$. The incompressibility of the fluid is expressed by the condition div $\mathbf{v} = 0$. We must have $\mathbf{v} \cdot \nu = 0$ on the boundary of $\mathcal{D}$, where ν is the outward unit normal at the boundary. Euler's equations of motion for the flow of an inviscid, incompressible fluid are

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}, \qquad \rho \frac{d\mathbf{v}}{dt} = \rho(\mathbf{v}_t + (\mathbf{v} \cdot \nabla)\mathbf{v}) = -\nabla(p - \mathbf{g} \cdot \mathbf{x}), \tag{2.1}$$

$$\text{div } \mathbf{v} = 0,$$

where ρ is the density, p is the hydrodynamic pressure, and $\mathbf{g} \cdot \mathbf{x}$ is the gravitational potential. Henceforth we take $\rho = 1$.

Given a manifold $\mathcal{D} \in \mathbb{R}^3$ with smooth boundary, we denote by $\mathcal{L}^2(\mathcal{D})$ the Hilbert space of vector fields on $\mathcal{D}$ with the inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle = \iiint_{\mathcal{D}} \mathbf{v} \cdot \mathbf{w} \, d^3\mathbf{x}.$$

We denote by $\mathcal{L}_\pi$ the closed subspace of $\mathcal{L}^2(\mathcal{D})$ generated by vector fields, of the form $\mathbf{w} = \nabla p$ for some function p with finite Dirichlet norm. The orthogonal complement $\mathcal{L}_\sigma = \mathcal{L}_\pi^\perp$ is the space of all vector fields $\mathbf{v}$ for which $\langle \mathbf{v}, \nabla p \rangle = 0$ for all $p \in W^{1,2}(\mathcal{D})$. By applying the Gauss divergence theorem, we see that if $\mathbf{v} \in \mathcal{L}^2(\mathcal{D})$ and is smooth, say C^1, then div $\mathbf{v} = 0$ and $\mathbf{v} \cdot \nu = 0$ on $\Sigma = \partial\mathcal{D}$, where ν denotes the outward unit normal on Σ. The Hilbert space $\mathcal{L}_\sigma$ is the space of weakly divergence-free vector fields. We denote the orthogonal projections onto $\mathcal{L}_\sigma$ and $\mathcal{L}_\pi$ by P_σ and P_π, respectively.

In many applications the fluid is not confined to a fixed region but instead carries the region with it. In such cases, the region $\mathcal{D}$ occupied by the fluid must also be determined. Such problems are called free boundary problems and occupy a substantial part of the literature on incompressible flows.

Given an irrotational flow (curl $\mathbf{v} = 0$) on a simply connected domain, there is velocity potential φ for which $\mathbf{v} = \nabla\varphi$. The velocity potential is defined only up to an arbitrary function of time; the transformation $\varphi \mapsto \varphi + k(t)$ is called a *gauge transformation* and will play a role in what follows.

The equation div $\mathbf{v} = 0$ implies that φ is harmonic. Substituting $\mathbf{v} = \nabla\varphi$ into the second equation in (2.1) we obtain

$$\nabla \left(\varphi_t + \frac{1}{2}(\nabla\varphi)^2 + gz + p \right) = 0;$$

hence

$$\varphi_t + \frac{1}{2}(\nabla\varphi)^2 + gz + p = k(t)$$

for some function of time, which can be eliminated by a gauge transformation of the velocity potential. We always choose the gauge to be such that

$$\varphi_t + \frac{1}{2}(\nabla\varphi)^2 + gz + p = 0$$

everywhere in the fluid.

An interface between the fluid and another medium—for example, air—is called a *free surface*. If the pressure is constant in the air, then it is also constant at the

surface of the fluid, and we may normalize the pressure to be zero at the free surface. Hence we obtain Bernoulli's equation

$$\varphi_t + \frac{1}{2}(\nabla\varphi)^2 + gz = 0,$$

where gz is the gravitational potential on the free surface.

The free surface is given in space-time by $\phi = 0$, where $\phi(x, y, z, t) = z - \zeta(x, y, t)$. The free surface moves with the fluid; hence the material derivative of ϕ vanishes, and

$$0 = \frac{d\phi}{dt} = \frac{d}{dt}(z - \zeta) = v^3 - \zeta_t - v^1\zeta_x - v^2\zeta_y.$$

Thus

$$\zeta_t + v^1\zeta_x + v^2\zeta_y - v^3 = 0.$$

This is called the *kinematic* condition on the free surface.

This collection of equations for gravity waves on a free surface is known as Euler's equations for waves on the surface of an inviscid, incompressible fluid with irrotational flow in the region $\mathcal{D} = \{(x, y, z) : 0 \le z \le h + \zeta(x, y, t)\}$. They are

$$\Delta\varphi = 0, \qquad 0 \le y \le h + \zeta,$$

$$\zeta_t + \varphi_x\zeta_x + \varphi_y\zeta_y = \varphi_z \qquad \text{on } S, \tag{2.2}$$

$$\varphi_t + \frac{1}{2}|\nabla\varphi|^2 + g\,z = 0 \qquad \text{on } S, \tag{2.3}$$

$$\varphi_z = 0 \qquad \text{on } z = 0.$$

Here, φ is the velocity potential of the flow, and $\zeta(x, y, t)$ the displacement of the fluid surface from equilibrium. We have neglected surface tension. The second equation is known as the *kinematic equation*; the third equation is Bernoulli's equation. At rest, the fluid lies in the region $0 \le z \le h$; g is the acceleration due to gravity. The free surface is denoted by $S = \{(x, y, z) : z = h + \zeta(x, y, t)\}$.

The two physical constants in the theory are g and h. Let c denote a characteristic velocity (e.g., the velocity of a gravity wave); then h/c is a characteristic time. We introduce dimensionless variables

$$(x, y, z) = h(x', y', z'), \qquad t = ht'/c, \qquad \varphi = ch\varphi'.$$

Equation (2.3) now becomes

$$\varphi'_{t'} + \frac{1}{2}(\nabla'\varphi')^2 + \lambda\,\zeta = 0, \qquad \lambda = \frac{gh}{c^2}, \tag{2.4}$$

where λ is the inverse square of the *Froude* number. The other equations in Euler's system are unchanged under the rescaling. From now on we drop the primes and understand that we are working in nondimensional variables.

The Euler equations are invariant under the one parameter subgroup of Galilean boosts along the x axis, given by

$$(x', y', t') = (x - ct, y, t), \qquad \mathbf{v}'(x', y', t') = \mathbf{v}(x, y, t) - (c, 0).$$

The velocity potential, however, is determined only up to a function of time. Thus the Galilean boosts on the velocity potential are given by

$$\varphi'(x', y', t') = \varphi(x, y, t) - cx + q(t).$$

Under these Galilean boosts,

$$\frac{\partial \varphi'}{\partial t'} + \frac{1}{2}(\nabla' \varphi')^2 = \frac{\partial \varphi}{\partial t} + \frac{1}{2}(\nabla \varphi)^2 + q'(t) - \frac{1}{2}c^2. \tag{2.5}$$

The result follows by direct calculation, noting that

$$\frac{\partial}{\partial t'} = \frac{\partial}{\partial t} + c\frac{\partial}{\partial x}, \qquad \frac{\partial}{\partial x'} = \frac{\partial}{\partial x}.$$

PROPOSITION 2.1. *Suppose the solutions of Euler's equations are stationary in a Galilean frame moving with speed c. Then* $\zeta_{t'} = \varphi'_{t'} = 0$, *and, choosing* $q(t) = c^2 t$, *the conditions on the free surface are* (*dropping the primes*)

$$\varphi_x \zeta_x = \varphi_y, \qquad \frac{\varphi_x^2 + \varphi_y^2}{2} + \lambda\, \zeta = \frac{c^2}{2}. \tag{2.6}$$

Proof. By (2.5) the Bernoulli equation in the moving frame is

$$\varphi'_{t'} + \frac{1}{2}(\nabla' \varphi')^2 + \lambda \zeta' = q'(t') - \frac{1}{2}c^2.$$

As $x \to \pm\infty$, $\zeta' \to 0$ while $(\nabla' \varphi')^2 \to c^2$. Moreover, $\varphi'_{t'} = 0$ by the assumption of stationarity. These conditions force the choice $q' = c^2$, and the result follows. The kinematic equation in the moving frame is immediate. ∎

PROPOSITION 2.2. *Let* $\mathbf{v}$ *be a divergence-free vector field in a domain* $\mathcal{D}$. *There is a unique orthogonal decomposition, known as the Weyl–Hodge decomposition,*

$$\mathbf{v} = \mathbf{w} + \nabla \varphi, \tag{2.7}$$

$$\Delta \varphi = 0, \quad \varphi_\nu = \mathbf{v} \cdot \nu; \qquad \operatorname{div} \mathbf{w} = 0, \quad \mathbf{w} \cdot \nu = 0. \tag{2.8}$$

The proof is left to the reader.

3. Poisson structures. Let M be a C^∞ manifold of dimension n, and let $F, G \in C^\infty(M)$. A bilinear form $\{F, G\}$ is said to be a *Poisson bracket* if

- $\{F, G\} = -\{G, F\}$;
- $\{F, GH\} = \{F, G\}H + G\{F, H\}$;
- $\{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} = 0$.

The second property implies that the Poisson bracket is a derivation in each of its entries. Hence any $H \in C^\infty(M)$ generates a vector field X_H, called a *Hamiltonian* vector field on M, defined by $X_H F = \{H, F\}$. The Hamiltonian vector field X_H generates a flow on M; if x^i are a set of local coordinates on M, then the time evolution of the x^i on that chart is given by the ordinary differential equations

$$\dot{x}^i = \{H, x^i\}.$$

Due to the fact that the bracket acts as a derivation on each of its entries, we may represent a Poisson bracket in the form

$$\{F, G\} = \sum_{j,k=1}^{n} W^{jk} \frac{\partial F}{\partial x^j} \frac{\partial G}{\partial x^k},$$

where $W^{jk}(x)$ is a skew-symmetric matrix.

If $\det W \neq 0$, then it is easily seen that n must be even. A classical theorem of Darboux states that in this case it is always possible to find a set of local coordinates, called *canonical* coordinates q^i, p^i $(1 \leq i \leq n/2)$, in which the Poisson brackets take the form

$$\{F, G\} = \sum_{j=1}^{n} \frac{\partial F}{\partial p^j} \frac{\partial G}{\partial q^j} - \frac{\partial F}{\partial q^j} \frac{\partial G}{\partial p^j}.$$

A manifold with a Poisson bracket is called a *Poisson manifold*; if the brackets are nondegenerate, the manifold is called a *symplectic* manifold. On a symplectic manifold, the Hamiltonian flow takes the form

$$\dot{q}^i = \frac{\partial H}{\partial p^i}, \qquad \dot{p}^i = -\frac{\partial H}{\partial q^i}.$$

In this paper we shall restrict ourselves to the case in which M is a linear vector space with an inner product $\langle\ ,\ \rangle$, and we shall write the Poisson brackets in the form

$$\{F, G\} = \langle \nabla F, J_x \nabla G \rangle,$$

where J_x is a skew-symmetric linear transformation on M and ∇F is the gradient of the function F. The gradient is characterized as follows. Differentiating $F(x(t))$ along a curve $x(t)$ on M, we have

$$\frac{d}{dt} F = \langle \nabla F, \dot{x} \rangle.$$

If J_x is nonsingular, then the Poisson brackets are nondegenerate and have locally a canonical system of coordinates. In many problems of physical interest, however, the Poisson brackets are degenerate, i.e., $\det J_x = 0$. For example, in the study of rigid motions about a fixed point in $\mathbb{R}^3$, the Poisson bracket is

$$\{F, G\} = \langle \nabla F, \mathbf{x} \times \nabla G \rangle. \tag{3.1}$$

The operator $J_{\mathbf{x}}$ is defined by $J_{\mathbf{x}}\mathbf{v} = \mathbf{x} \times \mathbf{v}$; hence $\ker(J_{\mathbf{x}}) = \mathbb{R}\mathbf{x}$.

The bracket (3.1) vanishes for all regular functions G whenever F is spherically symmetric. Such a function F is called a *Casimir*. It is invariant under any Hamiltonian flow generated by these brackets.

Any Poisson bracket on an odd-dimensional manifold must be degenerate and therefore have Casimirs. The bracket (3.1) is an example of a noncanonical Poisson bracket.

The formalism of Poisson brackets and Hamiltonian flows can be extended to infinite dimensions—for example, in the study of continuum mechanics—although a number of technical difficulties arise. In particular, Poisson structures play a useful role in the theory of the Euler equations for an incompressible fluid. Two important such brackets are the Poisson bracket introduced by Arnold [3, 4, 5] in his study of incompressible fluids on fixed domains, and the Poisson bracket implicit in Zakharov's fundamental discovery [18] of the Hamiltonian structure of the Euler equations of gravity waves.

3.1. Arnold's Poisson brackets. Arnold observed that Euler's equations for an incompressible fluid in a fixed domain $\mathcal{D}$ are directly analogous to his equations for rigid body motion and that they have a Hamiltonian structure with the Hamiltonian and Poisson brackets given respectively by

$$H = \iiint_{\mathcal{D}} \frac{1}{2}\mathbf{v}\cdot\mathbf{v}\, d^3\mathbf{x} \tag{3.2}$$

and

$$\{F, G\} = \iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot \left(\operatorname{curl} \mathbf{v} \times \frac{\delta G}{\delta \mathbf{v}}\right) d^3\mathbf{x}. \tag{3.3}$$

Here, F and G are functionals on $\mathcal{L}_\sigma$ with gradients in $\mathcal{L}_\sigma$. The gradient of F is $\delta F/\delta \mathbf{v}$, the Euler–Lagrange derivative of F with respect to $\mathbf{v}$. For example, $\frac{\delta H}{\delta \mathbf{v}} = \mathbf{v}$. The operator $J_{\mathbf{v}}$ in this case is

$$J_{\mathbf{v}}\mathbf{w} = P_\sigma(\operatorname{curl} \mathbf{v} \times \mathbf{w}).$$

Let us show that (2.1) are the Hamiltonian equations generated by (3.2) and (3.3). We have

$$\dot{F} = \iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot \mathbf{v}_t\, d^3\mathbf{x},$$

$$\{H, F\} = \iiint_{\mathcal{D}} \frac{\delta H}{\delta \mathbf{v}} \cdot \left(\operatorname{curl} \mathbf{v} \times \frac{\delta F}{\delta \mathbf{v}}\right) d^3\,\mathbf{x} = \iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot (\mathbf{v} \times \operatorname{curl} \mathbf{v})\, d^3\,\mathbf{x}.$$

The Hamiltonian flow $\dot{F} = \{H, F\}$ implies that

$$\iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot (\mathbf{v}_t + (\operatorname{curl} \mathbf{v}) \times \mathbf{v})\, d^3\mathbf{x} = 0 \tag{3.4}$$

for all admissible F on $\mathcal{L}_\sigma$.

All linear functionals of the form $F_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{w}, \mathbf{v}\rangle$ are admissible, and the gradient of $F_{\mathbf{w}}$ is the vector $\mathbf{w}$. Therefore $\mathbf{v}_t + (\operatorname{curl} \mathbf{v}) \times \mathbf{v}$ belongs to $\mathcal{L}_\sigma^\perp = \mathcal{L}_\pi$. Hence it is a gradient, and

$$\mathbf{v}_t + (\operatorname{curl} \mathbf{v}) \times \mathbf{v} = \nabla f$$

for some function f. The Euler momentum equations (2.1) follow from this and the vector identity

$$(\operatorname{curl} \mathbf{v}) \times \mathbf{v} = (\mathbf{v}\cdot\nabla)\mathbf{v} - \frac{1}{2}\nabla|\mathbf{v}|^2 \tag{3.5}$$

if we let $f = -p + \frac{1}{2}\nabla|\mathbf{v}|^2$.

Just as in the case of rigid motion, the Arnold bracket is degenerate. This degeneracy is related to the action of the (formal) group of volume preserving diffeomorphisms acting on $\mathcal{D}$. Arnold's Poisson bracket is an example of a *Lie–Poisson* bracket.

3.2. Zakharov's Poisson brackets. In 1968, Zakharov made a striking observation: Euler's equations for *irrotational gravity waves* have a canonical Hamiltonian structure. The Hamiltonian (in nondimensional variables) is

$$H = \frac{1}{2}\iiint_{\mathcal{D}} (\nabla\varphi)^2\, d^3\mathbf{x} + \frac{1}{2}\lambda \iint_{\mathbb{R}^2} \zeta^2(x,y,t)\, d^2\mathbf{x}.$$

The Poisson brackets implicit in Zakharov's observation are the canonical brackets

$$\{F,G\} = \iint_{\mathbb{R}^2} \left(\frac{\delta F}{\delta\varphi}\frac{\delta G}{\delta\zeta} - \frac{\delta F}{\delta\zeta}\frac{\delta G}{\delta\varphi}\right) d^2\mathbf{x};$$

the Hamiltonian flow is then the canonical flow

$$\zeta_t = \frac{\delta H}{\delta\varphi}, \qquad \varphi_t = -\frac{\delta H}{\delta\zeta}.$$

The Hamiltonian H is regarded as a functional of $(\widetilde{\varphi},\, \zeta)$, where $\zeta = \zeta(x,y,t)$ is the height of the free surface and $\widetilde{\varphi} = \varphi|_S$ is the trace of the harmonic function φ on the free surface, with $\varphi_\nu = 0$ on the bottom. The evolution takes place in the space of harmonic functions on $\mathcal{D}$.

Zakharov's result is verified by calculating the gradients of H with respect to ζ and φ. Now

$$\frac{d}{d\varepsilon}H(\varphi,\zeta_\varepsilon)\Big|_{\varepsilon=0} = \iint_{\mathbb{R}^2}\left[\frac{1}{2}(\nabla\widetilde{\varphi})^2 + \lambda\zeta\right]\delta\zeta\, d^2\mathbf{x},$$

where $\nabla\widetilde{\varphi}$ denotes $\nabla\varphi\big|_{\mathbb{R}^2}$. By identification,

$$\frac{\delta H}{\delta\zeta} = \frac{1}{2}(\nabla\widetilde{\varphi})^2 + \lambda\zeta.$$

Similarly,

$$\begin{aligned}\frac{d}{d\varepsilon}H(\widetilde{\varphi}_\varepsilon,\zeta)\Big|_{\varepsilon=0} &= \iiint_{\mathcal{D}} \nabla\varphi\cdot\nabla\delta\varphi\, d^3\mathbf{x}\\ &= -\iiint_{\mathcal{D}} \delta\varphi\Delta\varphi\, d^3\mathbf{x} + \iint_{\Sigma}\delta\varphi\frac{\partial\varphi}{\partial\nu}dS = \iint_{\Sigma}\delta\varphi\frac{\partial\varphi}{\partial\nu}dS,\end{aligned}$$

since φ is harmonic in $\mathcal{D}$ and $\varphi_\nu = 0$ on the bottom.

On the free surface

$$\widetilde{\varphi}_\nu dS = \nabla\widetilde{\varphi}\cdot\frac{(-\zeta_x,-\zeta_y,1)}{\sqrt{1+\zeta_x^2+\zeta_y^2}}\sqrt{1+\zeta_x^2+\zeta_y^2}\,d^2\mathbf{x},$$

so

$$\frac{\delta H}{\delta\widetilde{\varphi}} = \widetilde{\varphi}_z - \widetilde{\varphi}_x\zeta_x - \widetilde{\varphi}_y\zeta_y.$$

The free boundary equations (2.2) and (2.3) are thus precisely the Hamiltonian equations for this system.

Remark. The effects of surface tension can be obtained by simply adding the boundary integral

$$\sigma \iint_{\Sigma} dS$$

to the Hamiltonian, where σ is the coefficient of surface tension and dS is the element of surface area on the free surface S. The inclusion of surface tension leads to an additional term in the Bernoulli equation; when the free surface is a graph $z = \zeta(x, y, t)$, it is

$$\varphi_t + \frac{1}{2}|\nabla\varphi|^2 + g\,z = \sigma\ \mathrm{div}\ \frac{\nabla\zeta}{\sqrt{1+(\nabla\zeta)^2}}, \qquad \nabla\zeta = (\zeta_x, \zeta_y).$$

The potential energy can also be written as the integral of the gravitational potential over the fluid domain, so that the Hamiltonian for gravity waves including the effects of surface tension is

$$H = \iiint_{\mathcal{D}} \left[\frac{(\nabla\varphi)^2}{2} + \lambda\, U_+(\mathbf{x})\right] d^3\mathbf{x} + \sigma \iint_{S} dS, \tag{3.6}$$

where $U_+(\mathbf{x})$ is the gravitational potential, truncated in such a way that the integral over the unbounded domain $\mathcal{D}$ converges. When the fluid is a horizontal layer and the gravity field is constant in the negative z direction, we take $U_+ = (z-1)_+$, where z_+ denotes the function given by z when $z > 0$ and by 0 when $z < 0$. The factor g has been absorbed into the pure parameter λ.

4. Free boundary flows with vorticity. Free boundary value flows with vorticity, with both gravitational forces and surface tension included, are generated by the Hamiltonian

$$H = \iiint_{\mathcal{D}} \mathcal{E}\, d^3\mathbf{x} + \sigma \iint_{\Sigma} dS, \qquad \mathcal{E} = \frac{\mathbf{v}\cdot\mathbf{v}}{2} + \lambda\, U_+(\mathbf{x}). \tag{4.1}$$

The corresponding Poisson brackets are [15]

$$\{F, G\} = \iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot \left(\mathrm{curl}\ \mathbf{v} \times \frac{\delta G}{\delta \mathbf{v}}\right) d^3\mathbf{x} + \iint_{\Sigma} \left(\frac{\delta F}{\delta\varphi}\frac{\delta G}{\delta\Sigma} - \frac{\delta F}{\delta\Sigma}\frac{\delta G}{\delta\varphi}\right) dS, \tag{4.2}$$

where Σ is the free boundary and dS is the element of surface area on Σ.

Admissible functionals are regarded as functions of $\mathbf{v}$ and Σ, the free boundary of $\mathcal{D}$, and their gradients are defined implicitly by the relation

$$\frac{d}{d\varepsilon} F(\mathbf{v}_\varepsilon, \Sigma_\varepsilon)\Big|_{\varepsilon=0} = \iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot \delta\mathbf{v}\, d^3\mathbf{x} + \iint_{\Sigma} \frac{\delta F}{\delta\Sigma}\delta\Sigma\, dS.$$

Variations with respect to the free surface are restricted to normal variations, in a sense explained below. Admissible functionals F are those for which $\delta F/\delta\mathbf{v}$ is a divergence free vector field. We require that $\iint_{\mathcal{D}} \delta\Sigma dS = 0$, reflecting the fact that only volume preserving variations are allowed. This means that the gradient of a functional with respect to Σ is determined only up to a constant.

Let $\mathcal{L}_d(\mathcal{D})$ be the space of divergence-free L^2 vector fields on $\mathcal{D}$. Let P_1 and P_2 be the orthogonal projections defined by $P_1\mathbf{v} = \mathbf{w}$ and $P_2\mathbf{v} = \nabla\varphi$ in the Weyl–Hodge decomposition.

LEMMA 4.1. *Let*

$$F(\mathbf{v}, \Sigma) = \iiint_{\mathcal{D}} \mathcal{F}(\mathbf{v}, \mathbf{x}) d^3\mathbf{x} + \sigma \iint_{\Sigma} dS$$

be an admissible functional. Then

$$\frac{\delta F}{\delta \mathbf{w}} = P_1 \frac{\delta F}{\delta \mathbf{v}} \in \mathcal{L}^2(\mathcal{D}, d^3\mathbf{x}).$$

The gradients with respect to φ and Σ lie in $\mathcal{L}^2(\Sigma, dS)$ and are given by

$$\frac{\delta F}{\delta \varphi} = \frac{\delta F}{\delta \mathbf{v}}\Big|_{\Sigma} \cdot \nu, \qquad \frac{\delta F}{\delta \Sigma} = \mathcal{F}(\mathbf{v}, \mathbf{x}) + \sigma\kappa\Big|_{\Sigma} \ \text{ mod constant},$$

where κ is the mean curvature function on Σ.

Proof. Applying the Weyl–Hodge decomposition to both $\delta\mathbf{v}$ and $\delta F/\delta\mathbf{v}$ we obtain

$$\left\langle \frac{\delta F}{\delta \mathbf{v}}, \delta\mathbf{v} \right\rangle = \iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot \delta\mathbf{v}\, d^3\mathbf{x} = \iiint_{\mathcal{D}} P_1 \frac{\delta F}{\delta \mathbf{v}} \cdot \delta\mathbf{w} + P_2 \frac{\delta F}{\delta \mathbf{v}} \cdot \delta\nabla\varphi\, d^3\mathbf{x}.$$

By the uniqueness of the Weyl–Hodge decomposition, we may conclude

$$\frac{\delta F}{\delta \mathbf{w}} = P_1 \frac{\delta F}{\delta \mathbf{v}}, \qquad \frac{\delta F}{\delta \varphi} = P_2 \frac{\delta F}{\delta \mathbf{v}}.$$

Since $\delta F/\delta\mathbf{v}$ is divergence-free, we have, by the divergence theorem,

$$\iiint_{\mathcal{D}} P_2 \frac{\delta F}{\delta \mathbf{v}} \cdot \delta\nabla\varphi\, d^3\mathbf{x} = \iiint_{\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot \nabla\delta\varphi\, d^3\mathbf{x} = \iint_{\partial\mathcal{D}} \frac{\delta F}{\delta \mathbf{v}} \cdot \nu\, \delta\varphi\, dS,$$

and the second relation follows.

Let Σ_ε be a one parameter family of surfaces parameterized by a vector valued map

$$\mathbf{X}(u, v, \varepsilon) = \mathbf{X}_0(u, v) + \varepsilon\, \delta\Sigma\, \mathbf{N}(u, v),$$

where $\mathbf{N}$ is the normal vector field to Σ. For ε sufficiently small, the symmetric difference $\mathcal{D}_\varepsilon \Delta \mathcal{D}$ of the domains bounded respectively by Σ_ε and Σ is contained in a tubular neighborhood of Σ. In this neighborhood, the volume element of the 3-space can be written as $d^3\mathbf{x} = dr dS$, where dS is the area element on Σ and dr corresponds to the normal coordinate in the tubular neighborhood. We get

$$\delta \iiint_{\mathcal{D}} \mathcal{F}(\mathbf{v}, \mathbf{x})\, d^3\mathbf{x} = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \iiint_{\mathcal{D}_\varepsilon \Delta \mathcal{D}} \mathcal{F}(\mathbf{v}, \mathbf{x})\, d^3\mathbf{x}$$

$$= \iint_{\Sigma} \left(\lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \int_0^{\varepsilon\delta\Sigma} \mathcal{F}(\mathbf{v}, \mathbf{x})\, dr \right) dS = \iint_{\Sigma} \mathcal{F}(\mathbf{v}, \mathbf{x}) \delta\Sigma\, dS.$$

On the other hand, by classical differential geometry,

$$\delta \iint_\Sigma dS = \iint_\Sigma \kappa \delta\Sigma \, dS,$$

where κ is the mean curvature function on Σ, and this completes the proof of Lemma 4.1. □

Let us derive the equations of motion from the Hamiltonian structure. We have

$$\frac{\delta H}{\delta \mathbf{v}} = \mathbf{v}, \qquad \frac{\delta H}{\delta \varphi} = \mathbf{v} \cdot \nu, \qquad \frac{\delta H}{\delta \Sigma} = \mathcal{E}\Big|_\Sigma + \sigma\kappa.$$

From $\dot{F} = \{H, F\}$, we get

$$\iiint_\mathcal{D} \frac{\delta F}{\delta \mathbf{v}} \cdot \mathbf{v}_t \, d^3\mathbf{x} + \iint_\Sigma \frac{\delta F}{\delta \Sigma} \Sigma_t \, dS \tag{4.3}$$

$$= \iiint_\mathcal{D} (\mathbf{v} \times (\operatorname{curl} \mathbf{v})) \cdot \frac{\delta F}{\delta \mathbf{v}} \, d^3\mathbf{x} + \iint_\Sigma \left(\mathbf{v} \cdot \nu \frac{\delta F}{\delta \Sigma} - (\mathcal{E} + \sigma\kappa) \frac{\delta F}{\delta \varphi} \right) dS.$$

Since

$$\iint_\Sigma \mathcal{E} \frac{\delta F}{\delta \varphi} \, dS = \iint_\Sigma \mathcal{E} \frac{\delta F}{\delta \mathbf{v}} \cdot \nu \, dS = \iiint_\mathcal{D} \nabla \mathcal{E} \cdot \frac{\delta F}{\delta \mathbf{v}} \, d^3\mathbf{x}$$

and $\delta F / \delta \mathbf{v}$ is divergence-free, we get from (4.3), using functionals for which $\delta F / \delta \varphi = 0$,

$$\mathbf{v}_t + (\operatorname{curl} \mathbf{v}) \times \mathbf{v} = \nabla(-p + \mathcal{E}), \qquad \Sigma_t = \mathbf{v} \cdot \nu\Big|_\Sigma.$$

The boundary condition on the bottom is $\mathbf{v} \cdot \nu = 0$, where ν is the outward normal.

The first equation, together with (3.5), implies that $\Delta p = -\operatorname{div}(\mathbf{v} \cdot \nabla)\mathbf{v}$. Substituting the two equations above into (4.3), we obtain

$$\iiint_\mathcal{D} \nabla p \cdot \frac{\delta F}{\delta \mathbf{v}} \, d^3\mathbf{x} - \iint_\Sigma \sigma\kappa \frac{\delta F}{\delta \varphi} \, dS = 0$$

for all admissible functionals F. Applying the divergence theorem to the integral over $\mathcal{D}$ we obtain

$$\iint_\Sigma (p - \sigma\kappa) \frac{\delta F}{\delta \varphi} \, dS = 0$$

for all admissible functionals F. But

$$\iint_\Sigma \frac{\delta F}{\delta \varphi} \, dS = \iint_\Sigma \frac{\delta F}{\delta \mathbf{v}} \cdot \nu \, dS = \iiint_\mathcal{D} \operatorname{div} \frac{\delta F}{\delta \mathbf{v}} \, d^3\mathbf{x} = 0,$$

and therefore

$$p\Big|_\Sigma = \sigma\kappa + \text{constant}. \tag{4.4}$$

Thus the Hamiltonian approach yields the dynamic conditions on the free boundary in the case of surface tension [7, 15, 12].

Remark. In the general theory one considers normal variations of the free surface, whereas in the theory of gravity waves on a free surface over a horizontal bottom, it is customary to use the height of the free surface, ζ. More generally, if the surface Σ is a graph over a fixed manifold $\mathcal{M}$, we may represent Σ by a "height" function ζ defined on $\mathcal{M}$. In that case we refer to $\delta\zeta$ as the "vertical" variation and $\delta\Sigma$ as the "normal" variation.

PROPOSITION 4.2. *Let $\delta\Sigma$ and $\delta\zeta$ denote the normal and vertical variations of a surface Σ in the case when Σ is a graph over a fixed manifold. Let Σ be given in local coordinates by $\phi = 0$, where $\phi = z - \zeta$. Then $\delta\zeta = |\nabla\phi|\,\delta\Sigma$.*

Proof. Let $\mathbf{X} : U \mapsto \mathbb{R}^3$ be a local embedding of Σ in $\mathbb{R}^3$, and let $\mathbf{X}_\varepsilon$ be a one parameter family of embeddings, with $\mathbf{X}_0 = \mathbf{X}$. Then

$$\delta\Sigma = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} (\mathbf{X}_\varepsilon(u,v)) \cdot \nu.$$

Let Σ be defined by $\phi = 0$, $\phi = z - \zeta$. Then $\mathbf{X}_\varepsilon = (u, v, \zeta_\varepsilon(u,v)$, and

$$\delta\Sigma = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \begin{pmatrix} u \\ v \\ \zeta_\varepsilon(u,v) \end{pmatrix} \cdot \nu = \begin{pmatrix} 0 \\ 0 \\ \delta\zeta \end{pmatrix} \cdot \frac{\nabla\phi}{|\nabla\phi|} = \frac{\delta\zeta}{|\nabla\phi|}. \qquad \square$$

5. Variational principles for traveling waves. The Hamiltonian structure of the equations for gravity waves can be used to obtain variational principles for traveling waves—waves of constant speed and shape. Such a wave is a stationary solution of the Hamiltonian system in a Galilean frame moving with the wave; thus the wave is a critical point for the Hamiltonian, computed in such a reference frame. We apply the method here to the general case of gravity waves on a horizontal surface. The variational principle for irrotational flows given below appears to be new.

A variational approach, if successful, would permit a global treatment of the existence of traveling waves by the direct methods of the calculus of variation, but so far, the existence of traveling waves for potential flows of low amplitude has been proved by perturbation methods. The first existence theorems were given independently for periodic wave trains by Levi-Civita [14] and Struik [17] in the case of finite depth. The existence of the solitary wave, which is a more difficult problem, was first proved by Friedrichs and Hyers [10], since the bifurcation problem in this case is a singular perturbation problem (see the discussion by Sattinger [16]). These authors used conformal mapping techniques. A dynamical systems approach to the existence of traveling waves has been developed by Kirchgässner [13]; Amick and Toland [2] have shown that periodic wave trains tend to a solitary wave in the limit as the period tends to infinity.

In the direct method, one first uses compactness properties of the functional to obtain a minimum from a minimizing sequence. In general, this guarantees only a weak solution of the associated Euler–Lagrange equations. In many cases, these are elliptic equations, and it is possible to prove sufficient regularity of the weak solution to show that in fact it is a classical solution to the problem. (See Alt and Caffarelli [1] for functionals of Friedrichs' type.) For the present, we simply indicate the method for the problems discussed here in the theorems below.

THEOREM 5.1. *Euler's equations for gravity waves are the Euler–Lagrange equations for the functional*

$$\mathcal{H}(\varphi,\zeta)=\iint\limits_{\mathcal{D}_\zeta}\left[\frac{1}{2}\left[(\nabla\varphi)^2-1\right]+\lambda(y-1)_+\right]d^2\mathbf{x}\,, \tag{5.1}$$

where

$$y_+=\begin{cases}0, & y\le 0,\\ y, & y\ge 0,\end{cases}$$

$$\mathcal{D}_\zeta=\{(x,y)\,:\,-\infty<x<\infty,\ 0\le y\le 1+\zeta(x)\},$$

and the minimum is taken over all functions φ for which

$$\iint\limits_{\mathcal{D}}\left[(\varphi_x-1)^2+\varphi_y^2\right]\,d^2\mathbf{x}<+\infty.$$

If (φ,ζ) is a local minimum of $\mathcal{H}$, then φ is harmonic on the interior of $\mathcal{D}_\zeta$; if ζ is C^1 and $\varphi\in H^2(\mathcal{D})$, then the kinematic and Bernoulli equations hold on the free surface.

Remark. The Hamiltonian (5.1) is the renormalization of the Hamiltonian in the moving frame. By carrying out the integration in y we obtain

$$\iint\limits_{\mathcal{D}}(y-1)_+\,d^2\mathbf{x}=\frac{1}{2}\int_{-\infty}^{\infty}\zeta^2\,dx;$$

thus $\mathcal{H}$ can also be written

$$\mathcal{H}(\varphi,\zeta)=\iint\limits_{\mathcal{D}}\frac{1}{2}\left[(\nabla\varphi)^2-1\right]\,d^2\mathbf{x}+\frac{\lambda}{2}\int_{-\infty}^{\infty}\zeta^2(x)\,dx.$$

If $\varphi\in H^2(\mathcal{D})$ and $\zeta\in C^1$, then $\nabla\varphi$ has an L^2 trace on the boundary $y=\zeta$, and Stokes' theorem applies.

Proof. Let (φ,ζ) be a minimizer of $\mathcal{H}$ and suppose that ζ is C^1 and $\varphi\in H^2(\mathcal{D})$. Let $(\varphi_\varepsilon,\zeta_\varepsilon)$ be a one parameter family of admissible functions and denote the corresponding domains by $\mathcal{D}_\varepsilon$. By the calculations in section 3.2 we have

$$\begin{aligned}\delta\mathcal{H}(\delta\varphi,\delta\zeta)&=\frac{\partial H(\varphi_\varepsilon,\zeta_\varepsilon)}{\partial\varepsilon}\Big|_{\varepsilon=0}\\ &=-\iint\limits_{\mathcal{D}}\Delta\varphi\delta\varphi\,d^2\mathbf{x}+\oint\limits_{\partial\mathcal{D}}\varphi_\nu\delta\varphi\,ds+\int_{-\infty}^{+\infty}\left[\frac{1}{2}(\nabla\varphi)^2-\frac{1}{2}+\lambda\zeta\right]\delta\zeta\,dx\\ &=0\end{aligned}$$

for all admissible $\delta\varphi$, $\delta\zeta$.

Since the bottom is fixed, $\varphi_\nu=0$ on $y=0$. We first restrict ourselves to variations for which $\delta\zeta=\delta\varphi\big|_S=0$. Then the double integral must vanish for a set of variations $\delta\varphi$ which are dense in $L^2(\mathcal{D})$; it follows that φ is harmonic in the interior of $\mathcal{D}$. As before, $\varphi_\nu ds=\nabla\varphi\cdot(-\zeta_x,1)dx=(\varphi_y-\zeta_x\varphi_x)dx$, and so

$$\delta\mathcal{H}=\int_{-\infty}^{+\infty}\left[\left(\frac{(\nabla\varphi)^2-1}{2}+\lambda\zeta\right)\delta\zeta+(\varphi_y-\varphi_x\zeta_x)\,\delta\varphi\right]dx.$$

Setting first $\delta\zeta=0$ and letting $\delta\varphi$ vary on Σ, we obtain the kinematic equation on the free surface. Therefore the second term always vanishes. Now allowing $\delta\zeta$ to vary, we see that Bernoulli's equation holds on Σ. ∎

6. A variational problem with constraint. Whereas Friedrich's paper shows that Bernoulli's equation is not obtained when the functional J is minimized with respect to the stream function, Constantin, Sattinger, and Strauss showed in [7] that traveling gravity waves in the rotational case are obtained as extremals of a variational problem for the stream function with constraints. The existence of traveling water waves with vorticity was established in [8] for the periodic case. In a recent Ph.D. thesis at Brown University, V. Hur [11] has constructed solitary waves with nonzero vorticity. Some of their qualitative properties were investigated in [6].

In the irrotational case we have the following theorem.

THEOREM 6.1. *Define the set of admissible functions $\mathcal{K} = \{\psi, \zeta\}$ with the following properties:*

$$\text{(i)} \qquad \int_{-\infty}^{+\infty} \zeta(x)dx = m, \quad \int_{-\infty}^{+\infty} \zeta^2\, dx < \infty,$$

$$\text{(ii)} \qquad \psi(x,0) = 0, \quad \psi(x, 1+\zeta(x)) = 1,$$

$$\text{(iii)} \qquad \iint\limits_{\mathcal{D}} \left[\psi_x^2 + (\psi_y - 1)^2\right] d^2\mathbf{x} < +\infty.$$

Consider the variational problem

$$\lambda = \inf_{\mathcal{K}} \frac{\iint_{\mathcal{D}} \left[(\nabla\psi)^2 - 1\right] d^2\mathbf{x}}{\int_{-\infty}^{+\infty} \zeta^2\, dx}.$$

Let (ψ, ζ) be a minimizer in $\mathcal{K}$ of the above variational principle. Then ψ is harmonic in the interior of $\mathcal{D}$. If ζ is C^1 and $\psi \in H^2(\mathcal{D})$, then the Bernoulli equation is satisfied on the free surface $\psi = 1$. Hence minima of the above variational problem provide an irrotational flow for the gravity wave problem.

Proof. Let (ψ, ζ) be a minimizer, and let $\psi_\varepsilon, \zeta_\varepsilon$ be a family of admissible functions with $\psi_0 = \psi$ and $\zeta_0 = \zeta$. Then $J(\varepsilon) \geq 0$ and $J(0) = 0$, where

$$J(\varepsilon) = \iint\limits_{\mathcal{D}_\varepsilon} \left[(\nabla\psi_\varepsilon)^2 - 1\right]\, d^2\mathbf{x} - \lambda \int_{-\infty}^{+\infty} \zeta_\varepsilon^2\, dx.$$

Then $\delta J(\delta\psi, \delta\zeta) = 0$ for all admissible variations, where

$$\begin{aligned}\delta J &= \iint\limits_{\mathcal{D}} 2\nabla\psi \cdot \nabla\, \delta\psi\, d^2\mathbf{x} + \int_{-\infty}^{+\infty} \left[(\nabla\psi)^2 - 1 - 2\lambda\zeta\right] \delta\zeta\, dx \\ &= -2 \iint\limits_{\mathcal{D}} \Delta\psi\, \delta\psi\, d^2\mathbf{x} + 2 \oint\limits_{\partial\mathcal{D}} \delta\psi\psi_\nu\, ds + \int_{-\infty}^{+\infty} \left[(\nabla\psi)^2 - 1 - 2\lambda\zeta\right] \delta\zeta\, dx.\end{aligned}$$

The integral over the bottom of the flow domain vanishes, since $\psi_\nu = 0$ there. On the free surface (see $\delta(2)$ [9, p. 65])

$$\delta\psi + \psi_y \delta\zeta = 0.$$

This follows immediately by differentiating the relation $\psi_\varepsilon(x, 1+\zeta_\varepsilon(x)) \equiv 1$ with respect to ε and setting ε equal to zero. Similarly, differentiating the expression $\psi(x, 1+\zeta(x)) \equiv 1$ with respect to x we find that $\psi_x/\psi_y = -\zeta_x$; hence

$$\psi_\nu = \nabla\psi \cdot \nu = \nabla\psi \cdot \frac{\nabla\psi}{||\nabla\psi||} = \frac{(\nabla\psi)^2}{\sqrt{\psi_x^2+\psi_y^2}} = \frac{(\nabla\psi)^2}{|\psi_y|\sqrt{1+\zeta_x^2}}.$$

Hence δJ reduces to

$$\delta J = -2\iint_{\mathcal{D}} \Delta\psi\,\delta\psi\,d^2\mathbf{x} - \int_{-\infty}^{+\infty} \left[(\nabla\psi)^2 + 1 + 2\lambda\zeta\right]\delta\zeta\,dx. \tag{6.1}$$

First restrict the variations to fixed domains, $\delta\zeta = 0$, and the first integral must vanish for all variations $\delta\psi$ which vanish on $\partial\mathcal{D}$. Hence ψ is harmonic in the interior of $\mathcal{D}$, and the double integral vanishes.

We next consider variations of the domain. Since $\int \zeta_\varepsilon\,dx = m$ for all variations, we have $\int \delta\zeta\,dx = 0$; then the condition

$$\int_{-\infty}^{+\infty} ((\nabla\psi)^2 + 1 + 2\lambda\zeta)\delta\zeta\,dx = 0$$

for all such $\delta\zeta$ implies that the integrand is a constant. We therefore have $(\nabla\psi)^2 + 2\lambda\zeta + 1 = C =$ constant on the line; letting $x \to \infty$ and noting that $\zeta \to 0$ while $(\nabla\psi)^2 \to 1$ we see that $C = 2$, and the Bernoulli equation is satisfied. □

Acknowledgment. The results in this paper were obtained during the authors' visit to the Mittag-Leffler Institute in October 2005 in conjunction with the Program on Wave Motion. The authors wish to extend their thanks to the institute for its generous sponsorship of the program, as well as to the organizers for their work.

REFERENCES

[1] H. W. Alt and L. A. Caffarelli, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105–144.

[2] C. J. Amick and J. F. Toland, *On periodic water-waves and their convergence to solitary waves in the long-wave limit*, Philos. Trans. Roy. Soc. London Ser. A, 303 (1981), pp. 633–669.

[3] V. I. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, Ann. Inst. Fourier (Grenoble), 16 (1966), pp. 319–361.

[4] V. I. Arnold, *The Hamiltonian nature of the Euler equations in the dynamics of a rigid body and of an ideal fluid*, Uspekhi Mat. Nauk, 24 (1969), pp. 225–226.

[5] V. I. Arnold and B. A. Khesin, *Topological Methods in Hydrodynamics*, Appl. Math. Sci. 125, Springer-Verlag, New York, 1998.

[6] A. Constantin and J. Escher, *Symmetry of steady deep-water waves with vorticity*, European J. Appl. Math., 15 (2004), pp. 755–768.

[7] A. Constantin, D. H. Sattinger, and W. Strauss, *Variational formulations for steady water waves with vorticity*, J. Fluid Mech., 548 (2006), pp. 151–163.

[8] A. Constantin and W. Strauss, *Exact steady periodic water waves with vorticity*, Comm. Pure Appl. Math., 57 (2004), pp. 481–527.

[9] K. O. Friedrichs, *Über ein Minimumproblem für Potentialströmung mit freiem Rand*, Math. Ann., 109 (1933), pp. 60–82.

[10] K. O. Friedrichs and D. H. Hyers, *The existence of solitary waves*, Comm. Pure Appl. Math., 7 (1954), pp. 517–550.

[11] V. Hur, *Exact solitary water waves with vorticity*, submitted.

[12] R. S. Johnson, *A Modern Introduction to the Mathematical Theory of Water Waves*, Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, UK, 1997.
[13] K. Kirchgässner, *Nonlinearly resonant surface waves and homoclinic bifurcation*, in Advances in Applied Mechanics, Vol. 26, Adv. Appl. Mech. 26, Academic Press, Boston, 1988, pp. 135–181.
[14] T. Levi-Civita, *Détermination rigoureuse des ondes permanentes d'ampleur finie*, Math. Ann., 93 (1925), pp. 264–314.
[15] D. Lewis, J. Marsden, R. Montgomery, and T. Ratiu, *The Hamiltonian structure for dynamic free boundary problems*, Phys. D, 18 (1986), pp. 391–404.
[16] D. H. Sattinger, *Tsunamis and barge canals*, J. Math. Fluid Mech., 7 (2006), to appear.
[17] D. J. Struik, *Determination rigoueuse des ondes irrotationnelles périodiques dans un canal à profondeur finie*, Math. Ann., 95 (1926), pp. 595–634.
[18] V. E. Zakharov, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, J. Appl. Mech. Tech. Phys., 2 (1968), pp. 190–194.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 921–943

STEADY PERIODIC CAPILLARY-GRAVITY WAVES WITH VORTICITY*

ERIK WAHLÉN†

Abstract. In this paper we prove the existence of steady periodic two-dimensional capillary-gravity waves on flows with an arbitrary vorticity distribution. The original free-surface problem is first transformed to a second-order quasi-linear elliptic equation with a second-order quasi-linear boundary condition in a fixed domain by a change of variables. We then use local bifurcation theory combined with the Schauder theory of elliptic equations with Venttsel boundary conditions and spectral theory in Pontryagin spaces to construct the solutions. We show that some bifurcation points are simple while others are double, a situation already known to occur in the case of irrotational capillary-gravity waves.

Key words. water waves, vorticity, capillarity, bifurcation theory

AMS subject classifications. 76B15, 76B03, 76B45, 35Q35, 35J65, 47J15

DOI. 10.1137/050630465

1. Introduction.

The mathematical studies of water waves have until recently been mainly restricted to irrotational flows (see, e.g., [8, 12, 16, 31, 38]). While the irrotational setting is regarded as appropriate for waves traveling into still water [19, 27], there are many situations in which it is necessary to take vorticity into account. For example, nonuniform currents generate water flows with vorticity [35, 36, 37] and the effect of a wind blowing in one direction results at first in the creation of capillary waves.

In the last few years there has been an increasing amount of research in the area of water waves with vorticity; see [4, 5] regarding the symmetry of rotational water waves, [14, 23, 24, 39] for questions of uniqueness, and [2, 3, 6, 7, 40] for existence results. However, in all of these recent investigations the surface tension is neglected. It is therefore an interesting task to study the effects of surface tension in the presence of vorticity. The recent paper [41] dealt with the existence of pure capillary waves with vorticity, that is, the force of gravity was neglected. In the present paper we take into account both gravity and capillarity. It is known that the irrotational capillary-gravity wave problem is a single problem in which many different bifurcation phenomena appear. One of the main differences from the theory of pure capillary and gravity waves is the presence of double bifurcation points. Closely connected with this is the existence of secondary bifurcation points. We refer the reader to [8, 20, 33, 34] (finite depth) and [8, 21, 22] (infinite depth) for a detailed presentation of the local bifurcation theory of irrotational capillary-gravity waves. As we shall see, double bifurcation points also occur when vorticity is present. However, we do not fully investigate the structure of the solution set at these points.

The main result of this paper is the proof of the existence of steady periodic capillary-gravity waves for an arbitrary vorticity distribution. The proof is inspired by the approach in [7], but there are several interesting differences. We will first formulate the water wave problem in terms of the stream function. After this we perform

*Received by the editors September 16, 2005; accepted for publication (in revised form) March 6, 2006; published electronically October 12, 2006.
http://www.siam.org/journals/sima/38-3/63046.html
†Department of Mathematics, Lund University, P.O. Box 118, 22100 Lund, Sweden (erik.wahlen@math.lu.se).

a change of variables which transforms the free-boundary problem into a second-order quasi-linear elliptic equation in a fixed domain. As in [41], one difference to the pure gravity problem is that this elliptic equation has a second-order boundary condition. This means that we cannot use the standard Schauder theory for elliptic equations, but instead have to rely on similar estimates for so-called Venttsel boundary conditions (see [29]). A perhaps more dramatic difference is that the eigenvalue problem associated with bifurcation is not a standard Sturm–Liouville problem. The operator T involved is, however, self-adjoint with respect to an indefinite inner product (see [18]). Unlike the pure capillary case, the operator $T(\lambda)$ is not positive definite for all values of the bifurcation parameter λ. This is the key to why double bifurcation points occur when we also include gravity. After the proof of the main theorem, we investigate the distribution of the bifurcation points and present a necessary and sufficient condition for the existence of double bifurcation points. We also consider what happens as we vary the surface tension and gravity. Finally, at the end of the paper we take a closer look at the interesting special cases of irrotational flows ($\gamma \equiv 0$) and flows with nonzero, constant vorticity.

2. Preliminaries. In this section we present the governing equations for capillary-gravity waves [11, 19]. We consider two-dimensional waves propagating over water with a flat bed. In its undisturbed state the equation for the flat surface is $y = 0$ and the flat bottom is given by $y = -d$ for some $d > 0$. The x-variable represents the direction of propagation and the wavelength is $2\pi/k$, where $k \in \mathbb{R}^+$ is the wave number. The equations of motion are the equation of mass conservation

$$u_x + v_y = 0 \tag{2.1}$$

and Euler's equation

$$\begin{cases} u_t + uu_x + vu_y = -P_x, \\ v_t + uv_x + vv_y = -P_y - g, \end{cases} \tag{2.2}$$

where $P(t, x, y)$ denotes the pressure and g is the gravitational constant. The boundary conditions for capillary-gravity waves are the dynamic boundary condition

$$P = P_0 - \sigma \frac{\eta_{xx}}{(1 + \eta_x^2)^{\frac{3}{2}}} \quad \text{on} \quad y = \eta(t, x), \tag{2.3}$$

P_0 being the constant atmospheric pressure and $\sigma > 0$ being the coefficient of surface tension, as well as the kinematic boundary conditions

$$v = \eta_t + u\,\eta_x \quad \text{on} \quad y = \eta(t, x) \tag{2.4}$$

and

$$v = 0 \quad \text{on} \quad y = -d; \tag{2.5}$$

cf. [19].

We are looking for steady periodic waves traveling at speed $c > 0$, that is, the space-time dependence of the free surface, the pressure, and the velocity field is of the form $(x - ct)$. The map $x - ct \mapsto x$ transforms (2.2)–(2.5) into the stationary problem

$$\begin{cases} (u - c)u_x + vu_y = -P_x \\ (u - c)v_x + vv_y = -P_y - g \end{cases} \quad \text{on} \quad -d < y < \eta(x)$$

and

$$\begin{cases} v = (u-c)\eta_x \quad \text{at} \quad y = \eta(x), \\ P = P_0 - \sigma \frac{\eta_{xx}}{(1+\eta_x^2)^{\frac{3}{2}}} \quad \text{at} \quad y = \eta(x), \\ v = 0 \quad \text{at} \quad y = -d. \end{cases}$$

The equation of mass conservation (2.1) allows us to introduce the (relative) stream function ψ, satisfying $\psi_x = -v$, $\psi_y = u - c$. The kinematic boundary condition then shows that ψ is constant on the free surface, and we determine it uniquely by requiring that the constant value be zero. Field evidence indicates that for waves not near the spilling or breaking state, the propagation speed c of the surface wave is considerably larger than the horizontal velocity u of each individual water particle [28]. It follows that ψ is a strictly decreasing function of y for each fixed x. Let

$$p_0 = \int_{-d}^{\eta(x)} [u(x,y) - c] dy$$

be the relative mass flux—it follows by differentiation, using (2.1) and (2.4), that this expression is independent of $x \in \mathbb{R}$. Then by construction $\psi = -p_0$ on the flat bottom. We can now pose problem (2.1)–(2.5) in terms of ψ:

$$\begin{cases} \psi_y \psi_{xy} - \psi_x \psi_{yy} = -P_x \\ -\psi_y \psi_{xx} + \psi_x \psi_{xy} = -P_y - g \end{cases} \qquad \text{on} \quad -d < y < \eta(x)$$

and

$$\begin{cases} \psi = 0 \quad \text{at} \quad y = \eta(x), \\ P = P_0 - \sigma \frac{\eta_{xx}}{(1+\eta_x^2)^{\frac{3}{2}}} \quad \text{at} \quad y = \eta(x), \\ \psi = -p_0 \quad \text{at} \quad y = -d, \end{cases}$$

where P, ψ, and η are $2\pi/k$-periodic in the x-variable and $\psi_y < 0$.

The vorticity ω is defined by $\omega = v_x - u_y$. The assumption $u < c$ guarantees that ω is a function of ψ, that is, $\omega = \gamma(\psi)$ (see [7]). Thus $\Delta\psi = -\omega = -\gamma(\psi)$. Introduce the function

$$\Gamma(p) = \int_0^p \gamma(-s) ds, \qquad p_0 \le p \le 0,$$

and let $\Gamma_{\min} \le 0$ be its minimum value. Using the equations of motion and the properties of ψ we obtain Bernoulli's law, which states that

$$E = \frac{(c-u)^2 + v^2}{2} + gy + P - \Gamma(-\psi)$$

is constant throughout the fluid. On the free surface we have

$$E = \frac{(c-u)^2 + v^2}{2} + g\eta + P_0 - \sigma \frac{\eta_{xx}}{(1+\eta_x^2)^{\frac{3}{2}}},$$

so that letting $Q = 2(E - P_0 + gd)$ we obtain

$$(\psi_x^2 + \psi_y^2) + 2g(y+d) - 2\sigma \frac{\eta_{xx}}{(1+\eta_x^2)^{\frac{3}{2}}} = Q$$

on the free surface. This condition is equivalent to the dynamic boundary condition.

We have now obtained the following formulation of our problem:

$$\begin{cases} \Delta\psi = -\gamma(\psi) \quad \text{on} \quad -d < y < \eta(x), \\ |\nabla\psi|^2 + 2g(y+d) - 2\sigma\frac{\eta_{xx}}{(1+\eta_x^2)^{\frac{3}{2}}} = Q \quad \text{at} \quad y = \eta(x), \\ \psi = 0 \quad \text{at} \quad y = \eta(x), \\ \psi = -p_0 \quad \text{at} \quad y = -d. \end{cases}$$

The main difficulty with this formulation lies in the fact that η is not known a priori. For this purpose, we make a change of variables due to Dubreil-Jacotin [13]. Since ψ is constant on the free surface and on the bottom and strictly decreasing as a function of y, we choose the new variables $q = x$, $p = -\psi(x, y)$. A domain of one wavelength is then transformed to $R = \{(q, p) \in \mathbb{R}^2 : 0 < q < 2\pi/k,\ p_0 < p < 0\}$. Introducing the height function $h(q, p) = y + d$, we have

$$h_q = \frac{v}{u - c}, \quad h_p = \frac{1}{c - u}.$$

Thus

$$v = -\frac{h_q}{h_p}, \quad u = c - \frac{1}{h_p}, \quad \partial_x = \partial_q - \frac{h_q}{h_p}\partial_p, \quad \partial_y = \frac{1}{h_p}\partial_p.$$

Note also that $\omega = \gamma(-p)$ and that $\eta(x) = h(q, 0) - d$. We obtain the following formulation of the capillary-gravity problem:

$$\begin{cases} (1 + h_q^2)h_{pp} - 2h_p h_q h_{pq} + h_p^2 h_{qq} = -\gamma(-p)h_p^3 \quad \text{in} \quad p_0 < p < 0, \\ 1 + h_q^2 + (2gh - Q)h_p^2 - 2\sigma\frac{h_p^2 h_{qq}}{(1+h_q^2)^{\frac{3}{2}}} = 0 \quad \text{on} \quad p = 0, \\ h = 0 \quad \text{on} \quad p = p_0, \end{cases} \tag{2.6}$$

where h is $2\pi/k$-periodic in the q-variable and $h_p > 0$ throughout $\overline{R}$.

So far we have derived (2.6) from (2.1)–(2.5). We shall now see that it is also possible to derive (2.1)–(2.5) starting with (2.6). Denote the fluid domain

$$D_\eta = \{(x, y) \in \mathbb{R}^2 : -d < y < \eta(x)\}.$$

For the Hölder-parameter $\alpha \in (0, 1)$, let $C^{m+\alpha}_{\text{per}}(\overline{D}_\eta)$ be the space of functions $f : \overline{D}_\eta \to \mathbb{R}$ with Hölder-continuous derivatives of exponent α up to order m, and with period $2\pi/k$ in the x-variable. Similarly, $C^{m+\alpha}_{\text{per}}(\mathbb{R})$ denotes the space of $2\pi/k$-periodic real-valued functions on $\mathbb{R}$ of class $C^{m+\alpha}$. A small modification of the argument in [7] yields the following.

PROPOSITION 2.1. *Problem* (2.6) *is equivalent to* (2.1)–(2.5). *Furthermore, if* $h \in C^{2+\alpha}_{per}(\overline{R})$, *then* $(u, v, \eta) \in C^{1+\alpha}_{per}(\overline{D_\eta}) \times C^{1+\alpha}_{per}(\overline{D_\eta}) \times C^{2+\alpha}_{per}(\mathbb{R})$, *and if* h *is even in the* q*-variable, then* u *and* η *are even in* x *while* v *is odd.*

3. Main result.

THEOREM 3.1. *Let the wave speed* $c > 0$, *the relative mass flux* $p_0 < 0$, *the wave number* $k \in \mathbb{R}^+$, *and the vorticity function* $\gamma \in C^\alpha[0, |p_0|]$, $0 < \alpha < 1$, *be given. Then there exist infinitely many* C^1 *curves of small amplitude traveling wave solutions* (u, v, η) *of* (2.1)–(2.5) *in the space* $C^{1+\alpha}_{per}(\overline{D_\eta}) \times C^{1+\alpha}_{per}(\overline{D_\eta}) \times C^{2+\alpha}_{per}(\mathbb{R})$, *with period* $2\pi/k$, *speed* c, *and relative mass flux* p_0, *satisfying* $u < c$ *throughout the fluid. In*

particular, for sufficiently large numbers $n \in \mathbb{Z}^+$ there exist waves of minimal period $2\pi/(kn)$ with the property that (i) *the wave profile η has precisely one crest and one trough per period of length $2\pi/(kn)$;* (ii) *the wave profile is strictly monotone between crest and trough; and* (iii) *the functions u and η are even, while v is odd.*

Remark.

(i) We would like to point out that the depth varies along the solution continuum.

(ii) In Theorems 3.11 and 4.1, we will present more details about the set of small amplitude solutions.

LEMMA 3.2. *The trivial solutions (parallel shear flows with flat surfaces) $h(q,p) = H(p)$ of* (2.6) *are of the form*

$$H(p) = H(p;\lambda) = \int_0^p \frac{ds}{\sqrt{\lambda + 2\Gamma(s)}} + \frac{Q-\lambda}{2g} = \int_{p_0}^p \frac{ds}{\sqrt{\lambda + 2\Gamma(s)}},$$

where $0 \le -2\Gamma_{min} < \lambda$ and

$$Q(\lambda) = \lambda + 2g \int_{p_0}^0 \frac{dp}{\sqrt{\lambda + 2\Gamma(p)}}.$$

Proof. A solution of the form $H(p)$ satisfies the ordinary differential equation

$$H_{pp} = -\gamma(-p) H_p^3.$$

Integrating gives

$$H_p(p) = (\lambda + 2\Gamma(p))^{-\frac{1}{2}}$$

for $\lambda > -2\Gamma_{\min}$. The surface boundary condition $1 + (2gH(0) - Q)H_p^2(0) = 0$ yields $H(p) = \int_0^p \frac{ds}{\sqrt{\lambda+2\Gamma(s)}} + \frac{Q-\lambda}{2g}$ and the bottom boundary condition $H(p_0) = 0$ gives the expression for Q. ∎

Note that the function $Q(\lambda)$ is strictly convex for $\lambda > -2\Gamma_{\min}$ with $\lim_{\lambda\to\infty} Q'(\lambda) = 1$, respectively, $\lim_{\lambda\to -2\Gamma_{\min}} Q'(\lambda) = -\infty$. It follows that there is a unique $\lambda_0 > -2\Gamma_{\min}$ such that $Q'(\lambda_0) = 0$, that is,

$$\int_{p_0}^0 \frac{dp}{(\lambda_0 + 2\Gamma(p))^{3/2}} = \frac{1}{g}. \tag{3.1}$$

For every $Q > Q(\lambda_0)$ there is exactly one $\lambda > \lambda_0$ satisfying $Q(\lambda) = Q$, while for some Q there is a second solution in $(-2\Gamma_{\min}, \lambda_0)$.

Our goal is to show that the curve of trivial solutions parameterized by λ bifurcates at certain points. To identify these bifurcation points we will use the following tool.

THEOREM 3.3 (see [9]). *Let X and Y be Banach spaces, I an open interval in $\mathbb{R}$ containing λ^*, and $\mathcal{F}\colon I \times X \mapsto Y$ a continuous map with the following properties:*

(i) *$\mathcal{F}(\lambda, 0) = 0$ for all $\lambda \in I$;*

(ii) *$\mathcal{F}_\lambda$, $\mathcal{F}_w$, $\mathcal{F}_{\lambda w}$ exist and are continuous;*

(iii) *$\mathcal{N}(\mathcal{F}_w(\lambda^*, 0))$ and $Y/\mathcal{R}(\mathcal{F}_w(\lambda^*, 0))$ are one-dimensional, with the null space generated by w^*; and*

(iv) $\mathcal{F}_{\lambda w}(\lambda^*, 0)w^* \notin \mathcal{R}(\mathcal{F}_w(\lambda^*, 0))$.

Then there exists a continuous local bifurcation curve $\{(\lambda(s), w(s)) : |s| < \varepsilon\}$ *with* ε *sufficiently small such that* $(\lambda(0), w(0)) = (\lambda^*, 0)$ *and*

$$\{(\lambda, w) \in \mathcal{U} : w \neq 0,\ \mathcal{F}(\lambda, w) = 0\} = \{(\lambda(s), w(s)) : 0 < |s| < \varepsilon\}$$

for some neighborhood $\mathcal{U}$ *of* $(\lambda^*, 0) \in I \times X$. *Moreover, we have*

$$w(s) = sw^* + o(s) \quad \text{in } X,\ |s| < \varepsilon.$$

(v) *If* $\mathcal{F}_{ww}$ *is also continuous, then the curve is of class* C^1.

We will also need a slightly different bifurcation theorem based on the Morse lemma [17].

THEOREM 3.4 (see [30]). *Let* X, Y, *and* I *be as in Theorem* 3.3, *and let* $\mathcal{F}\colon I \times X \to Y$ *be a* C^m*-map with* $m > 2$. *Suppose that for some* $\lambda^* \in I$ *we have*

(i) $\mathcal{F}(\lambda^*, 0) = \mathcal{F}_\lambda(\lambda^*, 0) = 0$;

(ii) $\mathcal{N}(\mathcal{F}_w(\lambda^*, 0))$ *and* $Y/\mathcal{R}(\mathcal{F}_w(\lambda^*, 0))$ *are one-dimensional, with the null space spanned by* w^*;

(iii) $\mathcal{F}_{\lambda\lambda}(\lambda^*, 0) \in \mathcal{R}(\mathcal{F}_w(\lambda^*, 0))$ *and* $\mathcal{F}_{\lambda w}(\lambda^*, 0)w^* \notin \mathcal{R}(\mathcal{F}_w(\lambda^*, 0))$.

Then the set of solutions of $\mathcal{F}(\lambda, w) = 0$ *near* $(\lambda^*, 0)$ *consists of two* C^{m-2} *curves* $\mathcal{C}_1, \mathcal{C}_2$ *intersecting only at* $(\lambda^*, 0)$. *Furthermore,* $\mathcal{C}_1$ *is tangent to the* λ*-axis at* $(\lambda^*, 0)$ *and may be parameterized by* λ:

$$(\lambda, w(\lambda)), \qquad |\lambda - \lambda^*| < \varepsilon;$$

$\mathcal{C}_2$ *may be parameterized by a variable* s, $|s| < \varepsilon$, *as*

$$(\lambda(s), sw^* + w(s)),$$

with $w(0) = w'(0) = 0$, $\lambda(0) = \lambda^*$.

If $\mathcal{F}$ is regular enough, Theorem 3.3 follows as a special case of Theorem 3.4. In order to apply Theorems 3.3 and 3.4, we formulate our problem as an abstract operator equation. Let R be the rectangle $(0, 2\pi/k) \times (p_0, 0)$, let $T = \{p = 0\}$ be the top, and let $B = \{p = p_0\}$ be the bottom of its closure $\overline{R}$, and define the spaces

$$X = \{h \in C^{2+\alpha}_{\text{per}}(\overline{R}) : h = 0 \text{ on } B\},$$
$$Y = C^{\alpha}_{\text{per}}(\overline{R}) \times C^{\alpha}_{\text{per}}(T),$$

where the subscript "per" means periodicity and evenness in the variable q. We also define the subspaces $X_n \subseteq X$ and $Y_n \subseteq Y$ for $n \in \mathbb{N}$, consisting of functions that are $2\pi/(kn)$-periodic. For $n = 0$ we take this to mean that X_n and Y_n consist of functions that are independent of q.

We define the nonlinear operator $\mathcal{G}\colon \mathbb{R} \times X \to Y$ by $\mathcal{G}(Q, h) = (\mathcal{G}_1(Q, h), \mathcal{G}_2(Q, h))$, where

$$\mathcal{G}_1(Q, h) = (1 + h_q^2)h_{pp} - 2h_p h_q h_{pq} + h_p^2 h_{qq} + \gamma(-p)h_p^3$$

and

$$\mathcal{G}_2(Q, h) = 1 + h_q^2 + (2gh - Q)h_p^2 - 2\sigma \frac{h_p^2 h_{qq}}{(1 + h_q^2)^{\frac{3}{2}}}.$$

Then (Q, h) solves (2.6) if and only if $\mathcal{G}(Q, h) = 0$.

In order to apply Theorem 3.3 we introduce the operator $\mathcal{F}(\lambda, w) = \mathcal{G}(Q(\lambda), H(\lambda)+w)$ for $w \in X$ and $\lambda \in I = (-2\Gamma_{\min}, \infty)$. Thus $\mathcal{F}(\lambda, w) = (\mathcal{F}_1(\lambda, w), \mathcal{F}_2(\lambda, w))$, where

$$\begin{aligned}\mathcal{F}_1(\lambda, w) = (1 + w_q^2)(H_{pp} + w_{pp}) - 2(H_p + w_p)w_q w_{pq}\\ +(H_p + w_p)^2 w_{qq} + \gamma(-p)(H_p + w_p)^3\end{aligned}$$

and

$$\mathcal{F}_2(\lambda, w) = 1 + w_q^2 + (2g(H + w) - Q)(H_p + w_p)^2 - 2\sigma\frac{(H_p + w_p)^2 w_{qq}}{(1 + w_q^2)^{\frac{3}{2}}}.$$

We have $\mathcal{F}(\lambda, 0) \equiv 0$ by construction.

The Fréchet derivative of $\mathcal{F}$ with respect to w at $w = 0$ is the pair $\mathcal{F}_w(\lambda, 0) = (\mathcal{F}_{1w}(\lambda, 0), \mathcal{F}_{2w}(\lambda, 0))$, where

$$\begin{aligned}\mathcal{F}_{1w}(\lambda, 0) &= \partial_p^2 + H_p^2\partial_q^2 + 3\gamma(-p)H_p^2\partial_p \quad \text{in } R,\\ \mathcal{F}_{2w}(\lambda, 0) &= 2(g\lambda^{-1} - \lambda^{\frac{1}{2}}\partial_p - \lambda^{-1}\sigma\partial_q^2)\Big|_T,\end{aligned}$$

since $H_p(0) = \lambda^{-\frac{1}{2}}$, so that the linearization of the problem (2.6) at $w = 0$ is $\mathcal{F}_w(\lambda, 0)w = 0$, i.e.,

$$\begin{cases} w_{pp} + H_p^2 w_{qq} + 3\gamma(-p)H_p^2 w_p = 0 \quad \text{in} \quad p_0 < p < 0,\\ gw - \lambda^{\frac{3}{2}}w_p - \sigma w_{qq} = 0 \quad \text{on} \quad p = 0,\\ w = 0 \quad \text{on} \quad p = p_0. \end{cases} \tag{3.2}$$

Introducing $a_\lambda = \{\lambda + 2\Gamma(p)\}^{\frac{1}{2}}$, we can write this in the form

$$\begin{cases} \{a_\lambda^3 w_p\}_p + a_\lambda w_{qq} = 0 \quad \text{in} \quad p_0 < p < 0,\\ a_\lambda^3 w_p + \sigma w_{qq} - gw = 0 \quad \text{on} \quad p = 0,\\ w = 0 \quad \text{on} \quad p = p_0. \end{cases} \tag{3.3}$$

For each $n \in \mathbb{N}$ we let $\mathcal{F}^{(n)}$ be the restriction of $\mathcal{F}$ to $I \times X_n$. Note that $\mathcal{F}^{(n)}(I \times X_n) \subseteq Y_n$.

LEMMA 3.5. *For each fixed λ, the map $\mathcal{F}_w^{(n)}(\lambda, 0)\colon X_n \to Y_n$ is a Fredholm operator of index 0.*

Proof. For a fixed $\lambda \in I$ and $\mu \in \mathbb{R}$, define the operator $S(\mu)\colon X_n \to Y_n$ by

$$\begin{aligned}S_1(\mu) &= \partial_p^2 + H_p^2\partial_q^2 + 3\gamma(-p)H_p^2\partial_p - \mu,\\ S_2(\mu) &= 2(g\lambda^{-1} - \lambda^{\frac{1}{2}}\partial_p - \lambda^{-1}\sigma\partial_q^2 + \mu\partial_p)\Big|_T.\end{aligned}$$

Then the map $\mu \mapsto S(\mu) \in B(X_n, Y_n)$ is continuous and the following estimate holds for some $C = C(\mu) > 0$:

$$\|w\|_X \le C(\|w\|_{C^0(\overline{R})} + \|S(\mu)w\|_Y); \tag{3.4}$$

cf. [29]. It follows at once that for a fixed μ, $S(\mu)$ is semi-Fredholm, that is, it has finite-dimensional null space and closed range. For $\mu > \lambda^{1/2}$ the map $S(\mu)$ is an

isomorphism (see [29]) and therefore has index 0. By continuity of the index, we infer that this is also true for $\mu = 0$, whence the assertion follows. □

For each λ we introduce the inner product

$$\langle (\mathcal{A}_1, \mathcal{B}_1), (\mathcal{A}_2, \mathcal{B}_2) \rangle_Y = \iint_R a^3 \mathcal{A}_1 \mathcal{A}_2 \, dq \, dp + \frac{1}{2} \int_T a^2 \mathcal{B}_1 \mathcal{B}_2 \, dq,$$

where $a = a_\lambda$, on Y. Note that $\mathcal{N} = \mathcal{N}(\mathcal{F}_w^{(n)}(\lambda, 0))$ can be identified with the subspace $\widehat{\mathcal{N}} = \{(\mathcal{A}, \mathcal{B}) \in Y_n : \mathcal{A} \in \mathcal{N}, \, \mathcal{B} = \mathcal{A}|_T\}$ of Y_n. We have the following characterization of the range of $\mathcal{F}_w^{(n)}(\lambda, 0) \colon X_n \to Y_n$.

LEMMA 3.6. *The range of $\mathcal{F}_w^{(n)}(\lambda, 0)$ is given by the orthogonal complement $\widehat{\mathcal{N}}^\perp$ of $\widehat{\mathcal{N}}$ in Y_n with respect to the inner product $\langle \cdot, \cdot \rangle_Y$.*

Proof. Let $(\mathcal{A}, \mathcal{B}) \in Y_n$ belong to the range of $\mathcal{F}_w^{(n)}(\lambda, 0)$, and let $\varphi \in \mathcal{N} = \mathcal{N}(\mathcal{F}_w^{(n)}(\lambda, 0))$. Then

$$\iint_R \mathcal{A} \cdot a^3 \varphi \, dq \, dp + \frac{1}{2} \int_T \mathcal{B} \cdot a^2 \varphi \, dq = 0.$$

Indeed, $(\mathcal{A}, \mathcal{B})$ belongs to $\mathcal{R} = \mathcal{R}(\mathcal{F}_w^{(n)}(\lambda, 0))$ if and only if $\mathcal{A} = a^{-3}\{a^3 v_p\}_p + a^{-2} v_{qq}$ in R, $\mathcal{B} = 2(ga^{-2} v - a v_p - \sigma a^{-2} v_{qq})$ on T. We have

$$\begin{aligned} \iint_R \mathcal{A} \cdot a^3 \varphi \, dq \, dp &= \iint_R \{\{a^3 v_p\}_p + a v_{qq}\} \varphi \, dq \, dp \\ &= \iint_R \{\{a^3 \varphi_p\}_p + a \varphi_{qq}\} v \, dq \, dp + \int_T a^3 \{v_p \varphi - v \varphi_p\} dq \\ &= \int_T a^3 \{v_p \varphi - v \varphi_p\} dq, \end{aligned}$$

where the integral over R vanishes because of the equation satisfied by φ.

On the top we have $2(v_p \varphi - v \varphi_p) = 2 v_p \varphi - 2v(g a^{-3} \varphi - \sigma a^{-3} \varphi_{qq}) = -a^{-1} \mathcal{B} \varphi + 2\sigma a^{-3}(v \varphi_{qq} - v_{qq} \varphi)$. Thus the last integral equals

$$-\frac{1}{2} \int_T a^2 \mathcal{B} \varphi \, dq + \sigma \int_T (v \varphi_{qq} - v_{qq} \varphi) \, dq,$$

where the last term disappears after an integration by parts. Thus $\mathcal{R}(\mathcal{F}_w^{(n)}(\lambda, 0)) \subseteq \widehat{\mathcal{N}}^\perp$. Since $\operatorname{codim} \mathcal{R} = \dim \mathcal{N} = \dim \widehat{\mathcal{N}} = \operatorname{codim} \widehat{\mathcal{N}}^\perp < \infty$ (see [1]), we have $\mathcal{R} = \widehat{\mathcal{N}}^\perp$. □

Note that any function $\varphi \in X_n$ can be naturally identified with an element $\widehat{\varphi} = (\varphi, \varphi|_T) \in Y_n$.

LEMMA 3.7. *Let $0 \neq \varphi \in \mathcal{N}(\mathcal{F}_w^{(n)}(\lambda, 0))$ be of the form $\varphi = W(p) \cos(knq)$, $n \in \mathbb{N}$. Then $\langle \mathcal{F}_{w\lambda}(\lambda, 0) \varphi, \widehat{\varphi} \rangle_Y \neq 0$ and thus $\mathcal{F}_{w\lambda}^{(n)}(\lambda, 0) \varphi \notin \mathcal{R}(\mathcal{F}_w^{(n)}(\lambda, 0))$.*

Proof. Throughout the proof we let $a = a_\lambda$. Let us first calculate $\mathcal{F}_{w\lambda}$. We have

$$\mathcal{F}_{w\lambda}(\lambda, 0) = \left(-a^{-4} \partial_q^2 - 3\gamma a^{-4} \partial_p, 2 \left(a^{-4} \sigma \partial_q^2 - \frac{1}{2} a^{-1} \partial_p - a^{-4} g \right) \Big|_T \right).$$

This yields

$$\begin{aligned} \langle \mathcal{F}_{w\lambda}(\lambda, 0) \varphi, \widehat{\varphi} \rangle_Y = &\iint_R a^3 \varphi \cdot (-a^{-4} \varphi_{qq} - 3\gamma(-p) a^{-4} \varphi_p) \, dq \, dp \\ &+ \int_T a^2 \varphi \cdot (a^{-4} \sigma \varphi_{qq} - (2a)^{-1} \varphi_p - a^{-4} g \varphi) \, dq. \end{aligned}$$

The first term equals

$$(kn)^2 \iint_R a^{-1}\varphi^2 \, dq\, dp$$

due to the cosine. The second term equals

$$-\frac{3}{2}\iint_R a\varphi_p^2 \, dq\, dp - \frac{3(kn)^2}{2}\iint_R a^{-1}\varphi^2 \, dq\, dp + \frac{3}{2}\int_T a\varphi\varphi_p \, dq. \tag{3.5}$$

Indeed, using $a_p = \gamma(-p)a^{-1}$ (by definition) and $(a^3\varphi_p)_p = (kn)^2 a\varphi$ throughout R, we have that

$$\begin{aligned}
&\iint_R \gamma(-p)a^{-1}\varphi\varphi_p \, dq\, dp = \iint_R a_p\varphi\varphi_p \, dq\, dp \\
&= -\iint_R (a\varphi_p^2 + a\varphi\varphi_{pp}) \, dq\, dp + \int_T a\varphi\varphi_p \, dq - \int_B a\varphi\varphi_p \, dq \\
&= -\iint_R a\varphi_p^2 \, dq\, dp - (kn)^2 \iint_R a^{-1}\varphi^2 \, dq\, dp \\
&\quad + 3\iint_R \gamma(-p)a^{-1}\varphi\varphi_p \, dq\, dp + \int_T a\varphi\varphi_p \, dq
\end{aligned}$$

since $\varphi = 0$ on B. We now obtain

$$\iint_R \gamma(-p)a^{-1}\varphi\varphi_p \, dq\, dp = \frac{(kn)^2}{2}\iint_R a^{-1}\varphi^2 \, dq\, dp + \frac{1}{2}\iint_R a\varphi_p^2 \, dq\, dp - \frac{1}{2}\int_T a\varphi\varphi_p \, dq,$$

proving (3.5). The total contribution of the last three terms is

$$-\frac{3}{2}\int_T a\varphi\varphi_p \, dq,$$

due to the boundary condition satisfied by φ on T.

Adding up all terms we find that

$$\langle \mathcal{F}_{w\lambda}(\lambda,0)\varphi, \widehat{\varphi}\rangle_Y = -\frac{(kn)^2}{2}\iint_R a^{-1}\varphi^2 \, dq\, dp - \frac{3}{2}\iint_R a\varphi_p^2 \, dp < 0,$$

so that $\mathcal{F}_{w\lambda}^{(n)}(\lambda,0)\varphi \notin \mathcal{R}(\mathcal{F}_w^{(n)}(\lambda,0))$. □

In order to prove bifurcation, we must investigate the null space of $\mathcal{F}_w(\lambda,0)$. Note that any function w in X can be expanded in a cosine series,

$$w(q,p) = \sum_{n=0}^{\infty} w_n(p)\cos(knq).$$

We thus obtain $\mathcal{F}_w(\lambda,0)w = 0$ if and only if each component w_n satisfies

$$\begin{cases} (a^3 w_n')' = (kn)^2 a w_n, \quad p_0 < p < 0, \\ a^3(0)w_n'(0) = ((kn)^2\sigma + g)w_n(0), \\ w_n(p_0) = 0, \end{cases} \tag{3.6}$$

where $a = a_\lambda$.

To study this spectral problem we introduce the (complex) Pontryagin space (see [1, 18]) $\mathbb{H} = L^2[p_0, 0] \times \mathbb{C}$, with the indefinite inner product

$$[\tilde{u}_1, \tilde{u}_2] = \langle a u_1, u_2 \rangle_{L^2} - \sigma b_1 \overline{b_2},$$

where $\tilde{u}_i = (u_i, b_i) \in \mathbb{H}$, $i = 1, 2$. It is clear that $\mathbb{H}$ is a π_1-space, that is, any maximal negative definite (or negative semidefinite) subspace of $\mathbb{H}$ has dimension one. On $\mathbb{H}$ there is also an associated Hilbert space inner product, given by $\langle \tilde{u}, \tilde{v} \rangle_{\mathbb{H}} = [J\tilde{u}, \tilde{v}]$, where

$$J = \begin{pmatrix} I & 0 \\ 0 & -1 \end{pmatrix}.$$

Define the linear operator T by

$$T\tilde{u} = \big(-a^{-1}(a^3 u')', -a^3(0)\sigma^{-1}u'(0) + g\sigma^{-1}u(0)\big),$$

where we take as domain of definition $D(T) = \{\tilde{u} = (u, b) : u \in H^2[p_0, 0],\ u(p_0) = 0,\ b = u(0)\}$. Then (3.6) is equivalent with the eigenvalue problem $T\tilde{u} = \mu\tilde{u}$ for $\mu = -(kn)^2$. It is clear that $D(T)$ is dense in $\mathbb{H}$ and that T is closed. The identity

$$[T\tilde{u}, \tilde{u}] = \langle a^3 u', u' \rangle_{L^2} - g|u(0)|^2 \in \mathbb{R} \tag{3.7}$$

shows that T is symmetric.

LEMMA 3.8. *T is self-adjoint with discrete spectrum $\Sigma(T)$. Moreover, $0 \in \Sigma(T)$ if and only if $\lambda = \lambda_0$. For $\lambda > \lambda_0$ the operator T is positive with one negative eigenvalue.*

Proof. Note that $(T - \mu I)\tilde{u} = \tilde{f} := (f, b)$ is equivalent to the system of equations

$$\begin{cases} -(a^3 u')' - \mu a u = af, \\ B_1(u) := u(p_0) = 0, \\ B_2(u) := -a^3(0)u'(0) + gu(0) - \mu\sigma u(0) = \sigma b. \end{cases} \tag{3.8}$$

Let u_1 and u_2 be solutions of the equation $-(a^3 u')' - \mu a u = 0$ with initial data $u_1(p_0) = 0$, $u_1'(p_0) = 1$, respectively, $u_2(p_0) = 1$, $u_2'(p_0) = 0$. The characteristic determinant

$$\Delta(\mu) = \begin{vmatrix} B_1(u_1) & B_1(u_2) \\ B_2(u_1) & B_2(u_2) \end{vmatrix}$$

is an entire function of μ. If μ is not a zero of $\Delta(\mu)$, (3.8) is solvable by means of the formula

$$u(p) = c_1 u_1(p) + c_2 u_2(p) + \int_{p_0}^{0} G(p, r, \mu) f(r)\, dr,$$

where G is the Green's function for (3.8) and c_1, c_2 are chosen so that u satisfies the boundary conditions. Clearly, $(u, u(0)) \in D(T)$. On the other hand any zero of $\Delta(\mu)$ is an eigenvalue of T. Since $\mathcal{N}(\mathcal{F}_w(\lambda, 0))$ is finite-dimensional, there exist integers n such that $-(kn)^2$ is not an eigenvalue of T. Thus $\Delta(\mu) \not\equiv 0$ and $\Delta(\mu)$ has only isolated zeros of finite multiplicity. Since T is symmetric and closed, it is self-adjoint with discrete spectrum. The eigenvalues are geometrically simple as an elementary

consequence of the theory of ordinary differential equations. Moreover, since $\mathbb{H}$ is a π_1-space, T has a maximal invariant negative semidefinite subspace which is of dimension one; cf. [18]. Accordingly, it has at least one negative semidefinite eigenvalue. There are two possible cases: either there is a unique real negative semidefinite eigenvalue or there are two complex conjugate negative semidefinite eigenvalues. In the latter case the corresponding eigenspaces are neutral, that is, $[\cdot,\cdot]$ vanishes on each eigenspace. If μ is a nonneutral eigenvalue, we have a decomposition $\mathbb{H} = \mathcal{N} \,[\dot{+}]\, \mathcal{N}^{[\perp]}$, where $\mathcal{N} = \mathcal{N}(T - \mu_- I)$ and $\mathcal{N}^{[\perp]} = \{\tilde{u} \in \mathbb{H} : [\tilde{u}, \mathcal{N}] = 0\}$, as an orthogonal direct sum. Since $\mathcal{R}(T - \mu_- I) \subset \mathcal{N}^{[\perp]}$, it follows that μ is simple. Thus μ can be nonsimple only if it is neutral.

Note that $\Delta(\mu) = -B_2(u_1) = a^3(0)u_1'(0) - g u_1(0) + \mu\sigma u_1(0)$, so that $\Delta(0) = 0$ if and only if

$$\int_{p_0}^{0} \frac{dp}{a^3(p)} = \frac{1}{g}.$$

By (3.1) this is true if and only if $\lambda = \lambda_0$.

For $\lambda > \lambda_0$ we use the Cauchy–Schwarz inequality to conclude that if $0 \neq \tilde{u} \in D(T)$,

$$\begin{aligned} |u(0)|^2 &= \Big|\int_{p_0}^{0} u_p(p)\,dp\Big|^2 = \Big|\int_{p_0}^{0} \frac{a^{3/2}(p)}{a^{3/2}(p)} u_p(p)\,dp\Big|^2 \\ &\leq \int_{p_0}^{0} a^{-3}(p)\,dp \int_{p_0}^{0} a^3(p)|u_p(p)|^2\,dp < \frac{1}{g}\int_{p_0}^{0} a^3(p)|u_p(p)|^2\,dp. \end{aligned}$$

It follows that $[T\tilde{u}, \tilde{u}] > 0$ for nonzero $\tilde{u}$, i.e., T is positive definite. Thus the negative semidefinite eigenvalue is actually negative definite. For any eigenvector $\tilde{u}$ corresponding to the eigenvalue μ, we have

$$\mu[\tilde{u}, \tilde{u}] = [T\tilde{u}, \tilde{u}] > 0.$$

It follows therefore that $\mu < 0$ if and only if μ is of negative definite type. Hence the eigenvalue of negative definite type is the unique negative eigenvalue for $\lambda > \lambda_0$. ∎

Remark. From the above proof it follows that the eigenvalues are always geometrically simple. Moreover, they can only be algebraically nonsimple if they are neutral. As we shall see in section 6, for $\lambda \leq \lambda_0$ one cannot exclude the existence of neutral eigenvalues that are algebraically nonsimple. Nor can one exclude the existence of nonreal neutral eigenvalues.

An immediate corollary of Lemma 3.8 is that for $\lambda > \lambda_0$ the space $\mathcal{N}(\mathcal{F}_w(\lambda, 0))$ is at most one-dimensional. As we shall see in the following lemma it is in general at most two-dimensional.

LEMMA 3.9. *The null space of $\mathcal{F}_w(\lambda, 0)$ is at most two-dimensional.*

Proof. Let $Z = \{\tilde{u} = (u, b) \in H^1[p_0, 0] \times \mathbb{C} : u(p_0) = 0\}$ be endowed with the indefinite inner product $[\tilde{u}_1, \tilde{u}_2]_Z = \langle a^3 u_1', u_2'\rangle_{L^2} - g\, b_1 \overline{b_2}$. Clearly Z is a π_1-space and on the subspace $D(T)$ of Z we have $[\tilde{u}, \tilde{v}]_Z = [T\tilde{u}, \tilde{v}]$. We already know that T has at most one negative semidefinite eigenvalue in $(-\infty, 0]$. Suppose that there are two different positive definite eigenvalues $\mu_1, \mu_2 \leq 0$ and let $\tilde{u}_1$, respectively, $\tilde{u}_2$, be corresponding eigenvectors. Then $[\tilde{u}_i, \tilde{u}_i]_Z = \mu_i[\tilde{u}_i, \tilde{u}_i] \leq 0$ for $i = 1, 2$. Noting that $[\tilde{u}_1, \tilde{u}_2] = 0$, we obtain a two-dimensional negative semidefinite subspace of Z. This is a contradiction. Hence, there is at most one positive definite eigenvalue in $(-\infty, 0]$ and in total at most two nonpositive eigenvalues. ∎

We next present a lemma on the behavior of $\mu_-(\lambda)$ as $\lambda \to \infty$, where $\mu_-(\lambda)$ is the negative definite eigenvalue of $T(\lambda)$.

LEMMA 3.10. *We have $\mu_-(\lambda) \to -\infty$ as $\lambda \to \infty$.*

Proof. As in [41] it can be seen that for each $\lambda > \lambda_0$

$$\mu_- = \max_{[\tilde{u},\tilde{u}]<0} \frac{[T\tilde{u},\tilde{u}]}{[\tilde{u},\tilde{u}]}, \quad \tilde{u} \in D(T).$$

For $n > 0$ let $\lambda \geq n\sigma + n^{-1}g - 2\Gamma_{\min}$. Then $a_\lambda(p) = \sqrt{\lambda + 2\Gamma(p)} \geq \sqrt{n\sigma + n^{-1}g}$. This yields

$$\int_{p_0}^{0} (n^2 a_\lambda |u|^2 + a_\lambda^3 |u_p|^2)\, dp \geq \sqrt{n\sigma + n^{-1}g} \int_{p_0}^{0} (n^2|u|^2 + (n\sigma + n^{-1}g)|u_p|^2)\, dp$$
$$\geq n(n\sigma + n^{-1}g) \int_{p_0}^{0} 2\mathrm{Re}\, u\overline{u}_p\, dp = (n^2\sigma + g)|u(0)|^2$$

for $\tilde{u} \in D(T)$. It follows that if $[\tilde{u},\tilde{u}] < 0$, $\frac{[T\tilde{u},\tilde{u}]}{[\tilde{u},\tilde{u}]} \leq -n^2$. Hence $\mu_-(\lambda) \leq -n^2$. ∎

Before proving Theorem 3.1, we give a theorem which describes what happens at every point $\lambda^* \in I$ where $\mathcal{F}_w(\lambda^*, 0)\colon X \to Y$ has a nontrivial null space. At certain points (cases 1a, 1b, and 2c) we can find the full structure of the solution set of (2.6) close to $(Q(\lambda^*), H(\lambda^*))$, while at others (case 2a) we can only get a partial result. Theorem 3.1 will follow by showing that case 1a occurs for infinitely many points $\lambda^* \in I$. In the next section we will show that all the other cases occur for special values of k and σ (cf. Theorem 4.1 and the remark following the proof of Theorem 4.1).

THEOREM 3.11. *Let $\mathcal{N}$ be the null space of the operator $\mathcal{F}_w(\lambda^*, 0)\colon X \to Y$, and let $(Q^*, H^*) = (Q(\lambda^*), H(\lambda^*))$ be the corresponding point on the curve of trivial solutions in the (Q, h)-variables. By Lemma 3.9 we know that $\mathcal{N}$ is at most two-dimensional.*

1. *Suppose that* $\dim \mathcal{N} = 1$.
   (a) *If $\lambda^* \neq \lambda_0$, then $\mathcal{N}$ is spanned by a function of the form $W(p)\cos(knq)$ with $n \in \mathbb{Z}^+$ (this follows from Lemma 3.8) and there exists a neighborhood $\mathcal{U}$ of (Q^*, H^*) in $\mathbb{R} \times X$, such that the solution set of the equation (2.6) in $\mathcal{U}$ consists of the trivial solution curve, and a curve of nontrivial solutions of minimal period $2\pi/(kn)$, intersecting the trivial solution curve at (Q^*, H^*) transversally.*
   (b) *If $\lambda^* = \lambda_0$, there exists a neighborhood $\mathcal{U}$ of (Q^*, H^*) in $\mathbb{R} \times X$, such that the solution set of (2.6) in $\mathcal{U}$ consists only of the trivial solution curve.*
2. *Suppose that* $\dim \mathcal{N} = 2$.
   (a) *If $\lambda^* \neq \lambda_0$, it follows from Lemma 3.8 that $\mathcal{N}$ is spanned by two functions $\varphi_1(q,p) = W_1(q)\cos(kn_1 q)$ and $\varphi_2(q,p) = W_2(q)\cos(kn_2 q)$ with $0 < n_1 < n_2$.*
      - *If $n_1 \mid n_2$, then the solution set of (2.6) in a neighborhood of (Q^*, H^*) consists of the trivial solution curve, and at least two curves of nontrivial solutions having minimal period $2\pi/(kn_1)$, respectively, $2\pi/(kn_2)$, intersecting the trivial curve at (Q^*, H^*) transversally (and transversally to each other).*
      - *If $n_1 \nmid n_2$, then the solution set of (2.6) in a neighborhood of (Q^*, H^*) consists of the trivial solution curve, and at least one curve of nontrivial solutions having minimal period $2\pi/(kn_2)$, intersecting the trivial curve at (Q^*, H^*) transversally.*

(b) *If $\lambda^* = \lambda_0$, then $\mathcal{N}$ is spanned by the function $\varphi_1(q,p) = H_\lambda(p;\lambda_0)$ and a function $\varphi_2(q,p) = W(p)\cos(knq)$. The solution set of (2.6) in a neighborhood of (Q^*, H^*) consists of the trivial solution curve, and a curve of nontrivial solutions having minimal period $2\pi/(kn)$, intersecting the trivial curve at (Q^*, H^*) transversally.*

Proof.

Case 1a. If $\mathcal{N}$ is one-dimensional and $\lambda \neq \lambda^*$, then $\mathcal{N}$ is spanned by a function $\varphi(q,p) = W(p)\cos(knq)$ for some $n \in \mathbb{Z}^+$. We can then apply Theorem 3.3 to the equation $\mathcal{F}(\lambda, w) = 0$ directly. The regularity conditions (ii) and (v) are obvious, while (iii) and (iv) follow from Lemmas 3.5 and 3.7 with $n = 1$. This provides us with a C^1 curve $(\lambda(s), w(s))$, $|s| < \varepsilon$, of solutions to $\mathcal{F}(\lambda, w) = 0$, with $(\lambda(0), w(0)) = (\lambda^*, 0)$ and $w'(0) = W(p)\cos(knq)$. Since the bifurcation analysis can be carried out in X_n and Y_n (using Lemmas 3.5 and 3.7 with this n), by restricting the bifurcation curve we can assume that the solutions are $2\pi/(kn)$-periodic. These solutions correspond to solutions of (2.6) by the transformations $Q = Q(\lambda)$ and $h = H(\lambda) + w$. Note that since $H_p > 0$ throughout $\overline{R}$ we obtain $h_p > 0$ throughout $\overline{R}$ close to the bifurcation point.

Case 1b. Note that $\mathcal{N}$ is in this case generated by the function $H_\lambda(p;\lambda_0)$. Indeed, differentiating $\mathcal{F}(\lambda, 0) \equiv 0$ gives

$$\begin{aligned} 0 = \mathcal{F}_\lambda(\lambda_0, 0) &= \mathcal{G}_Q(Q(\lambda_0), H(\lambda_0))Q'(\lambda_0) + \mathcal{G}_h(Q(\lambda_0), H(\lambda_0))H_\lambda(\lambda_0) \\ &= \mathcal{G}_h(Q(\lambda_0), H(\lambda_0))H_\lambda(\lambda_0) = \mathcal{F}_w(\lambda_0, 0)H_\lambda(\lambda_0). \end{aligned}$$

Again one can apply Theorem 3.3 to the equation $\mathcal{F}(\lambda, w) = 0$ to obtain a locally unique curve of "nontrivial" solutions (in the sense that $w \neq 0$). However, we will now show that this curve is just a copy of the trivial solution curve. Since Q attains its minimum value at λ_0, there exists for every λ close to λ_0 a point $\tilde{\lambda} \neq \lambda$ with $Q(\tilde{\lambda}) = Q(\lambda)$. Letting $w(\lambda) = H(\tilde{\lambda}) - H(\lambda)$, we see that $(\lambda, w(\lambda))$ maps to $(Q(\lambda), H(\tilde{\lambda})) = (Q(\tilde{\lambda}), H(\tilde{\lambda}))$, so that $\mathcal{F}(\lambda, w(\lambda)) = 0$. By uniqueness, this is the solution curve found by the Crandall–Rabinowitz theorem. From the above, we see that in the original (Q,h)-variables, $(\lambda, w(\lambda))$ corresponds to a point on the trivial solution curve. The reason why we get this copy of the trivial solution curve in the (λ, w)-variables is that the trivial solution curve has a turning point in the (Q,h)-variables. We refer to [40] for a more detailed discussion.

Case 2a. Let $\mathcal{N}$ be generated by the functions φ_1, φ_2 of modes n_1, respectively, n_2, where $0 < n_1 < n_2$ and $n_1 \nmid n_2$. Denote by $\mathcal{F}^{(n_i)}$ the restriction of $\mathcal{F}$ to $I \times X_{n_i}$. By hypothesis $\mathcal{N}(\mathcal{F}_w^{(n_i)}(\lambda^*, 0))$ is spanned by φ_i. By Lemma 3.7 with $n = n_i$ it follows that $\mathcal{F}_{w\lambda}^{(n_i)}(\lambda^*, 0)\varphi_i \notin \mathcal{R}(\mathcal{F}_w^{(n_i)}(\lambda^*, 0))$. Thus we can apply Theorem 3.3 to infer the existence of two bifurcation curves $\mathcal{C}_{n_i,\lambda^*} \subset X_{n_i}$, $i = 1, 2$. These curves are clearly contained in X. Again, we can pass to solutions of (2.6) by letting $Q = Q(\lambda)$ and $h = H(\lambda) + w$.

Suppose that $n_1 \mid n_2$. Then we can apply the same considerations as above to n_2 but not to n_1. Thus we can only infer the existence of the curve $\mathcal{C}_{n_2,\lambda^*}$ containing solutions of mode n_2.

Case 2b. In this case we must have $\mu_+(\lambda_0) = 0$ while $\mu_-(\lambda_0) = -(kn)^2 < 0$. Here μ_+ denotes the smallest positive definite eigenvalue of T. The null space is then two-dimensional and generated by the functions $\varphi_1(q,p) = H_\lambda(p;\lambda_0)$ (see case 1b) and $\varphi_2(q,p) = W(p)\cos(knq)$. By identifying X with a subset of $C^{2+\alpha}(\overline{R}) \times C^{2+\alpha}(T)$ in the canonical way, we have the inner product $\langle \cdot, \cdot \rangle_Y$ on X. The null space $\mathcal{N} = \mathcal{N}(\mathcal{F}_w(\lambda_0, 0))$ can then be decomposed as the orthogonal sum $\mathcal{N} = \mathcal{N}_1 \oplus \mathcal{N}_2$, where

$\mathcal{N}_i$ is the span of φ_i for $i = 1, 2$. The eigenfunction φ_1 comes from the fact that $(Q(\lambda_0), H(\lambda_0))$ is a turning point for the trivial solution curve $(Q(\lambda), H(\lambda))$. In order to avoid the kind of problems that appeared in case 1b, we shall not use the (λ, w)-coordinates.

We define the function

$$g(r, Q, v) = \mathcal{G}(Q(\lambda_0) + Q, H(\lambda_0) + r\varphi_1 + v)$$

for $r, Q \in \mathbb{R}$ and $v \in \mathcal{N}_1^\perp$. Our aim is to apply Theorem 3.4 to the function g, regarding r as the bifurcation parameter. We have to verify properties (i)–(iii) of the theorem. For notational convenience we let $Q_0 = Q(\lambda_0)$ and $H_0 = H(\lambda_0)$. The regularity of g is obvious. We have $g(0,0,0) = 0$ and $g_r(0,0,0) = \mathcal{G}_h(Q_0, H_0)\varphi_1 = \mathcal{F}_w(\lambda_0, 0)\varphi_1 = 0$, so that (i) is satisfied. Furthermore,

$$g_{(Q,v)}(0,0,0)(Q, v) = \mathcal{G}_Q(Q_0, H_0)Q + \mathcal{G}_h(Q_0, H_0)v.$$

Note that $\mathcal{G}_Q(Q_0, H_0) = (0, -H_p^2(\lambda_0)) \in \widehat{\varphi}_2^\perp \backslash \mathcal{R}(\mathcal{F}_w(\lambda_0, 0))$ and thus $\mathcal{R}(g_{(Q,v)}(0,0,0))$ $= \widehat{\mathcal{N}}_1 \oplus \mathcal{R}(\mathcal{F}_w(\lambda_0, 0))$. It follows that $\mathcal{R}(g_{(Q,v)}(0,0,0))$ has codimension one. Moreover, $\mathcal{N}(g_{(Q,v)}(0,0,0)) = \{0\} \times \mathcal{N}_2$, since $\mathcal{G}_Q(Q_0, H_0) \notin \mathcal{R}(\mathcal{F}_w(\lambda_0, 0))$, so that the null space is one-dimensional. This proves (ii). As for (iii), we clearly have $g_{rr}(0,0,0) = \mathcal{G}_{hh}(Q_0, H_0)[\varphi_1, \varphi_1] \in \mathcal{R}(g_{(Q,v)}(0,0,0))$ since $\mathcal{G}_{hh}(Q_0, H_0)[\varphi_1, \varphi_1]$ is independent of q and thus orthogonal to $\widehat{\varphi}_2$. Finally, we have

$$g_{r(Q,v)}(0,0,0)(0, \varphi_2) = \mathcal{G}_{hh}(Q_0, H_0)[\varphi_1, \varphi_2].$$

On the other hand, using that $Q'(\lambda_0) = 0$ we obtain

$$\mathcal{F}_{w\lambda}(\lambda_0, 0)\varphi_2 = \mathcal{G}_{hh}(Q(\lambda_0), H(\lambda_0))[\varphi_1, \varphi_2].$$

By Lemma 3.7, $\langle \mathcal{F}_{w\lambda}(\lambda_0, 0)\varphi_2, \widehat{\varphi}_2 \rangle_Y \neq 0$, so that

$$g_{r(Q,v)}(0,0,0)(0, \varphi_2) \notin \mathcal{R}(g_{(Q,v)}(0,0,0)).$$

This takes care of (iii). We can now conclude that the zero set of g close to the origin consists of two curves $\mathcal{C}_1, \mathcal{C}_2$ intersecting only at the origin, such that $\mathcal{C}_1$ is of the form $(r, Q(r), v(r))$ and $\mathcal{C}_2$ is of the form $(r(s), Q(s), s\varphi_2 + v(s))$, $|s| < \varepsilon$, with $r(0) = Q(0) = Q'(0) = v(0) = v'(0) = 0$. Clearly the first curve is a simply a reparameterization of the trivial solution curve, while the second curve gives all nontrivial solutions near (Q_0, H_0). Thus, even if the null space of $\mathcal{F}_w(\lambda_0, 0)$ is two-dimensional, the point (Q_0, H_0) is in some sense only a simple bifurcation point. The first eigenfunction simply comes from the fact that the trivial solution curve turns at this point. □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Note that $\mu_-(\lambda)$ is a real-analytic function for $\lambda > \lambda_0$. This is a simple consequence of the fact that, picking an appropriate branch of the square root, $T(\lambda)$ is an analytic family of unbounded operators on $L^2[p_0, 0] \times \mathbb{C}$ and that μ_- is the only eigenvalue in the lower half plane for each $\lambda > \lambda_0$ (cf. [25]). Using Lemma 3.10, we see that for any sufficiently large integer n there exists a $\lambda^* > \lambda_0$ such that $\mu_-(\lambda^*) = -(kn)^2$. Thus, we at least know that there are infinitely many points λ^* such that case 1a of Theorem 3.11 applies.

Proposition 2.1 allows us to pass from solutions of the problem (2.6) to solutions of the water wave problem (2.1)–(2.5). For any mode n-bifurcation curve (that is, with

solutions of minimal period $2\pi/(kn)$), we have $h(q,p) = H(p;\lambda(s)) + sW(p)\cos(knq) + o(s)$ in X along the bifurcation curve close to (Q^*, H^*). It follows that $h_q(q,0) = -knsW(0)\sin(knq) + o(s)$ in $C^{1+\alpha}_{\mathrm{per}}(T)$, while $h_{qq}(q,0) = -(kn)^2 sW(0)\cos(knq) + o(s)$ in $C^{\alpha}_{\mathrm{per}}(T)$. Choosing $q_0 \in (0, \pi/(2kn))$, we can find an ε such that for $0 < s < \varepsilon$, $h_q(q,0) < 0$ for $q \in (q_0, \pi/(kn) - q_0)$, $h_{qq}(q,0) < 0$ for $q \in [0, q_0)$, while $h_{qq}(q,0) > 0$ for $q \in (\pi/(kn) - q_0, \pi]$. Since $h(q,0)$ is even and $2\pi/(kn)$-periodic, it follows that $h_q(0,0) = h_q(\pi/(kn),0) = 0$ and thus by construction $h_q < 0$ in $(0, \pi/(kn))$ for $0 < s < \varepsilon$. Due to the antisymmetry of h_q with respect to $q = \pi/(kn)$, it follows for small positive s that $h_q > 0$ in $(\pi/(kn), 2\pi/(kn))$ and that $h|_T$ is strictly decreasing in $(0, \pi/(kn))$ and strictly increasing in $(\pi/(kn), 2\pi/(kn))$. For small $s < 0$ we have instead that $h|_T$ is strictly increasing in $(0, \pi/(kn))$ and strictly decreasing in $(\pi/(kn), 2\pi/(kn))$. The nodal properties now follow from $\eta = h|_T$. □

Remark. The above result is incomplete in the sense that it doesn't present the full structure of the solution set at the double bifurcation points. What makes a double bifurcation point different from a simple one is the presence of solutions of mixed modes, that is, solutions that do not rise and fall once per period. Such phenomena have been thoroughly studied in the irrotational case (see [20, 21, 22]), and the study of the full structure of the solution set for other vorticity distributions is an interesting problem, but is out of the scope of the present paper.

4. The distribution of the bifurcation points. We know from the previous section that there are only simple bifurcation points for $\lambda > \lambda_0$. In order to find double bifurcation points, we must look in $\lambda < \lambda_0$. The aim of this section is to prove the following theorem, which gives a necessary and sufficient condition for the existence of double bifurcation points.

THEOREM 4.1.

1. *If*

$$\int_{p_0}^{0} a_{\lambda_0}(p) \left(\int_{p_0}^{p} a_{\lambda_0}^{-3}(s)\, ds \right)^2 dp \le \frac{\sigma}{g^2}, \tag{4.1}$$

there are only simple bifurcation points which lie in $\lambda > \lambda_0$, one for each mode $n \in \mathbb{Z}^+$.

2. *If*

$$\int_{p_0}^{0} a_{\lambda_0}(p) \left(\int_{p_0}^{p} a_{\lambda_0}^{-3}(s)\, ds \right)^2 dp > \frac{\sigma}{g^2}, \tag{4.2}$$

there exists for every sufficiently large rational number n_2/n_1, $n_1, n_2 \in \mathbb{Z}^+$, a $k \in \mathbb{R}^+$ such that $\mathcal{F}$ has a double bifurcation point λ^* of mode (n_1, n_2) ($\mathcal{N}(\mathcal{F}_w(\lambda^*, 0))$ is spanned by two solutions of mode n_1, respectively, n_2). If (4.2) holds and $\gamma \le 0$, there exists for an arbitrary rational number $n_2/n_1 > 1$ (and some $k \in \mathbb{R}^+$ depending on n_1, n_2) a double bifurcation point of mode (n_1, n_2).

The proof of this theorem relies on three lemmas.

LEMMA 4.2. *The function μ_- is strictly decreasing for $\lambda > \lambda_0$.*

Proof. We denote $Lu = -(a^3 u_p)_p$, where $a = a_\lambda$, and let $W = W(p;\lambda)$ be defined as a real-analytic curve of solutions to the problem

$$LW = \mu_- aW, \quad W(p_0) = 0, \quad W_p(0) = (g - \mu_-\sigma)\lambda^{-\frac{3}{2}} W(0).$$

Denoting differentiation with respect to λ by a dot, $\dot{}$, we have $\dot{a} = 1/(2a)$. Furthermore,

$$L\dot{W} - \frac{3}{2}(aW_p)_p = \dot{\mu}_- aW + \frac{\mu_-}{2a} W + \mu_- a\dot{W},$$

$$\dot{W}(p_0) = 0,$$

and

$$\dot{W}_p(0) = -\dot{\mu}_-\lambda^{-\frac{3}{2}}\sigma W(0) - \frac{3}{2}(g-\mu_-\sigma)\lambda^{-\frac{5}{2}}W(0) + (g-\mu_-\sigma)\lambda^{-\frac{3}{2}}\dot{W}(0).$$

Multiplying the W equation by $\dot{W}$ and vice versa yields, after integrating,

$$\langle \dot{W}, LW\rangle_{L^2} = \mu_-\langle \dot{W}, aW\rangle_{L^2}$$

and

$$\begin{aligned}&\langle L\dot{W}, W\rangle_{L^2} + \frac{3}{2}\int_{p_0}^0 aW_p^2\,dp - \frac{3}{2}aW_pW\Big|^0\\ &= \dot{\mu}_-\int_{p_0}^0 aW^2\,dp + \int_{p_0}^0 \frac{\mu_-}{2a}W^2\,dp + \mu_-\langle a\dot{W}, W\rangle_{L^2},\end{aligned}$$

where $\langle\cdot,\cdot\rangle_{L^2}$ denotes the $L^2[p_0,0]$ inner product. On the other hand

$$\begin{aligned}\langle \dot{W}, LW\rangle_{L^2} - \langle L\dot{W}, W\rangle_{L^2} &= \int_{p_0}^0 \{-\dot{W}(a^3W_p)_p + (a^3\dot{W}_p)_pW\}dp\\ &= \{-a^3\dot{W}W_p + a^3\dot{W}_pW\}\big|^0.\end{aligned}$$

Combining the last three equations, we obtain

$$\begin{aligned}&\frac{3}{2}\int_{p_0}^0 aW_p^2\,dp - \frac{3}{2}aW_pW|^0\\ &= \dot{\mu}_-\int_{p_0}^0 aW^2\,dp + \int_{p_0}^0 \frac{\mu_-}{2a}W^2\,dp + a^3[\dot{W}_pW - \dot{W}W_p]\Big|^0.\end{aligned}$$

The boundary terms, evaluated at $p=0$, are

$$\begin{aligned}&a^3(\dot{W}_pW - \dot{W}W_p) + \frac{3}{2}aW_pW\\ &= \lambda^{\frac{3}{2}}\Big[-\dot{\mu}_-\lambda^{-\frac{3}{2}}\sigma W - \frac{3}{2}(g-\mu_-\sigma)\lambda^{-\frac{5}{2}}W + (g-\mu_-\sigma)\lambda^{-\frac{3}{2}}\dot{W} - (g-\mu_-\sigma)\lambda^{-\frac{3}{2}}\dot{W}\Big]W\\ &\quad + \frac{3}{2}(g-\mu_-\sigma)\lambda^{-1}W^2\\ &= -\dot{\mu}_-\sigma W^2.\end{aligned}$$

We thus have

$$\dot{\mu}_-[\tilde{W},\tilde{W}] = \dot{\mu}_-\Big(\int_{p_0}^0 aW^2\,dp - \sigma W^2(0)\Big) = -\mu_-\int_{p_0}^0 \frac{1}{2a}W^2\,dp + \frac{3}{2}\int_{p_0}^0 aW_p^2\,dp > 0,$$

so that μ_- is strictly decreasing in view of $[\tilde{W},\tilde{W}] < 0$. □

LEMMA 4.3. *Let $W = W(p;\lambda)$ be the solution of $-(a^3W')' = \mu aW$ with $W(p_0) = 0$ and $W'(p_0) = 1$ for a fixed $\mu \le 0$. Then*

$$B(\lambda) = a_\lambda^3(0)\frac{W'(0;\lambda)}{W(0;\lambda)}$$

is a strictly increasing function of $\lambda \in I$.

Proof. We proceed as in Lemma 4.2 to establish the formula

$$a^3(\dot{W}_p W - \dot{W} W_p) + \frac{3}{2} a W_p W \Big|^0 = \frac{3}{2} \int_{p_0}^0 a W_p^2 \, dp - \mu \int_{p_0}^0 \frac{1}{2a} W^2 \, dp,$$

where the dot denotes differentiation with respect to λ. Noting that the left-hand side is $W^2(0)B'(\lambda)$, the assertion follows. ∎

Note that by continuity, the negative semidefinite eigenvalue $\mu_-(\lambda_0)$ of $T(\lambda_0)$ must be nonpositive. Since $T(\lambda_0)$ has the eigenvalue 0, we either have $\mu_-(\lambda_0) = 0$, in which case $T(\lambda_0)$ has no negative eigenvalues, or $\mu_-(\lambda_0) < 0$, in which case $\mu_+(\lambda_0) = 0$, where $\mu_+(\lambda_0)$ is the first positive definite eigenvalue of $T(\lambda_0)$.

LEMMA 4.4. *If* $\mu_-(\lambda_0) = 0$, *there are no negative eigenvalues of* $T(\lambda)$ *for* $\lambda < \lambda_0$. *If* $\mu_-(\lambda_0) < 0$, *there is a minimal point* $\lambda_1 \in [-2\Gamma_{min}, \lambda_0)$ *such that for each* $\lambda \in I' = (\lambda_1, \lambda_0)$ *there is exactly two negative eigenvalues of* T. *These are the negative definite eigenvalue* μ_- *and the lowest positive definite eigenvalue* μ_+. *Moreover,* μ_- *is a strictly decreasing function of* $\lambda \in I'$, *while* μ_+ *is strictly increasing. If* $\lambda_1 > -2\Gamma_{min}$, *then* $\mu_-(\lambda_1) = \lim_{\lambda \to \lambda_1} \mu_{+/-}(\lambda)$ *and there are no negative eigenvalues for* $\lambda < \lambda_1$. *If* $\mu_-(\lambda_0) < 0$ *and* $\gamma \le 0$, *then* $\lambda_1 > -2\Gamma_{min}$.

Proof. We begin with the first case. Note that μ is an eigenvalue of $T(\lambda)$ if and only if $B(\lambda; \mu) = -\sigma\mu + g$, where B is the function in Lemma 4.3. By Lemma 3.10 and the assumption $\mu_-(\lambda_0) = 0$, every $\mu \le 0$ is an eigenvalue for some $\lambda \ge \lambda_0$. It follows from Lemma 4.3 that $B(\lambda; \mu) < -\sigma\mu + g$ for $\lambda < \lambda_0$. Hence, there are no negative eigenvalues in this region.

In the second case we note that at λ_0 the operator T has two simple nonpositive eigenvalues $\mu_-(\lambda_0) < 0$ and $\mu_+(\lambda_0) = 0$. The fact that $\mu_-(\lambda_0)$ is simple follows from the preservation of total algebraic multiplicity (see [25]) and from the simplicity of $\mu_-(\lambda)$ for $\lambda > \lambda_0$. By analytic perturbation theory it follows that the two eigenvalues can be represented as real-analytic functions μ_- and μ_+ of λ in a neighborhood of λ_0. Using the same argument as in Lemma 4.2, we see that

$$\dot{\mu}[\tilde{W}, \tilde{W}] = -\mu \int_{p_0}^0 \frac{1}{2a} W^2 \, dp + \frac{3}{2} \int_{p_0}^0 a W_p^2 \, dp > 0 \tag{4.3}$$

for $\mu = \mu_{+/-}$. We see from this that μ_- is negative definite and that it is a strictly decreasing function of λ close to λ_0. Similarly μ_+ is a strictly increasing function of λ close to λ_0. Thus $\mu_-(\lambda) < \mu_+(\lambda) < 0$ for $\lambda < \lambda_0$ close to λ_0.

We may now continue μ_- and μ_+ for $\lambda < \lambda_0$ until eventually either (i) we reach $-2\Gamma_{\min}$ or (ii) we reach a point $\lambda_1 > -2\Gamma_{\min}$ such that $\lim_{\lambda \to \lambda_1^+} \mu_-(\lambda) = \lim_{\lambda \to \lambda_1^+} \mu_+(\lambda)$. In the latter case $\mu_-(\lambda_1) = \lim_{\lambda \to \lambda_1^+} \mu_{+/-}(\lambda)$ is neutral and has algebraic multiplicity two. Note that it follows from (4.3) that $\dot{\mu}_{+/-}(\lambda)$ diverges at λ_1. Proceeding as for $\mu_-(\lambda_0) = 0$, we see that there are no negative eigenvalues for $\lambda < \lambda_1$.

If $\gamma \le 0$, it can be shown (see [40]) that $\lim_{\lambda \to \infty} B(\lambda; \mu) = \infty$, while $B(\lambda; \mu) \to 0$ as $\lambda \to -2\Gamma_{\min}$ for every $\mu < 0$. Consequently, every negative μ is an eigenvalue of $T(\lambda)$ for some $\lambda > -2\Gamma_{\min}$. It follows that $\lambda_1 > -2\Gamma_{\min}$, since otherwise there would exist a μ with $\lim_{\lambda \to -2\Gamma_{\min}} \mu_-(\lambda) \le \mu \le \lim_{\lambda \to -2\Gamma_{\min}} \mu_+(\lambda)$, which was never an eigenvalue of $T(\lambda)$. ∎

Remark. In section 6 we will see examples with $\mu_-(\lambda_0) = 0$ as well as $\mu_-(\lambda_0) < 0$. In the latter case we will show that both $\lambda_1 = -2\Gamma_{\min}$ and $\lambda_1 > -2\Gamma_{\min}$ can happen. Thus all the alternatives in Lemma 4.4 are possible.

Proof of Theorem 4.1. Since the eigenfunction corresponding to the eigenvalue 0 at λ_0 is

$$W(p) = \int_{p_0}^{p} a_{\lambda_0}^{-3}(s)\, ds,$$

with $W(0) = 1/g$ (cf. (3.6)), we see that the integral condition in the statement of the theorem corresponds to $[\tilde{W}, \tilde{W}] \leq 0$, respectively, $[\tilde{W}, \tilde{W}] > 0$. In the first case the eigenvalue 0 is of negative semidefinite type, that is, $\mu_-(\lambda_0) = 0$, while in the second case 0 is of positive definite type, that is, $\mu_-(\lambda_0) < 0$ and $\mu_+(\lambda_0) = 0$.

Suppose that the first case holds, that is, $\mu_-(\lambda_0) = 0$. The first statement in Theorem 4.1 is then a direct consequence of Lemma 4.2, Lemma 4.4, and the fact that the bifurcation points for $\lambda > \lambda_0$ are all simple.

Suppose instead that the second case holds, that is, $\mu_-(\lambda_0) < 0$. With μ_- and μ_+ as in Lemma 4.4, we then have $\mu_-(\lambda)/\mu_+(\lambda) \to \infty$ as $\lambda \to \lambda_0^-$. Hence there exists for every sufficiently large rational number n_2/n_1 a point $\lambda^* < \lambda_0$ with $\mu_-(\lambda^*) = (n_2/n_1)^2 \mu_+(\lambda^*)$. Letting $k = \sqrt{|\mu_+(\lambda^*)|}/n_1$, we obtain $\mu_+(\lambda^*) = -(kn_1)^2$ and $\mu_-(\lambda^*) = -(kn_2)^2$. Thus λ^* is a mode (n_1, n_2)-bifurcation point.

If in the second case $\lambda_1 > -2\Gamma_{\min}$, where λ_1 is as in Lemma 4.4, then $\mu_-(\lambda)/\mu_+(\lambda) \to 1$ as $\lambda \to \lambda_1^+$, so that $\mu_-(\lambda)/\mu_+(\lambda)$, $\lambda > \lambda_1$, ranges through all real numbers greater than 1. Hence, there are bifurcation points of all modes (n_1, n_2), $n_1, n_2 \in \mathbb{Z}^+$, with $n_2 > n_1$. In particular, if $\gamma \leq 0$, this is the case. □

Remark. Theorem 4.1 shows that case 2a of Theorem 3.11 can occur, both with $n_1 \mid n_2$ and $n_1 \nmid n_2$. Case 1b occurs, for example, when (4.1) holds, while case 2b occurs if (4.2) holds and we take $k = \sqrt{|\mu_-(\lambda_0)|}/n$ for an arbitrary $n \in \mathbb{Z}^+$. Thus all the cases in Theorem 3.11 are possible.

5. On the limiting cases $g \to 0$ and $\sigma \to 0$.

5.1. The case $g \to 0$. We now investigate what happens if we consider g and σ as parameters. First we consider for a fixed $\sigma > 0$ what happens as $g \to 0$. Note that λ_0 is defined by (3.1), and that by letting $g \to 0$ we obtain

$$\int_{p_0}^{0} \frac{dp}{(\lambda_0 + 2\Gamma(p))^{3/2}} \to \infty$$

and thus $\lambda_0 \to -2\Gamma_{\min}$. Furthermore, it is easy to see that $Q(\lambda) \to \lambda$. Thus, at a formal level we recover the results in [40] on bifurcation for pure capillary waves, where the bifurcation parameter is Q and T is positive definite for every $\lambda > -2\Gamma_{\min}$. In the case of pure capillary waves there are only simple bifurcation points.

5.2. The case $\sigma \to 0$. If we instead fix $g > 0$ and let $\sigma \to 0$, we see that the value of λ_0 doesn't change. Note that for each $\mu \leq 0$, the function $W = W(p; \mu, \lambda)$ satisfying

$$\begin{cases} -(a^3 W')' = \mu a W, \quad p_0 < p < 0, \\ W(p_0) = 0, \qquad W'(p_0) = 1 \end{cases} \tag{5.1}$$

has the property $W(0) > 0$. By continuous dependence on parameters, it follows that for each compact subset K of I and $R > 0$, there is an $\varepsilon > 0$ and a $\delta > 0$ such that $W(0; \mu, \lambda) \neq 0$ and $\langle aW, W \rangle_{L^2} / |W(0)|^2 \geq \delta$ for $\mu \in [-R, 0] + i[-\varepsilon, \varepsilon]$ and $\lambda \in K$. In $[\lambda_0, \infty)$ we already know that the negative definite eigenvalues are real and satisfy

$\mu_-(\lambda) \le \mu_-(\lambda_0)$. Hence, for each $\varepsilon > 0$ and $R > 0$ there is a neighborhood U of $-R \le \operatorname{Re}\mu \le 0$, $\operatorname{Im}\mu = 0$ such that for $\lambda \ge -2\Gamma_{\min} + \varepsilon$ and σ sufficiently small there are no negative semidefinite eigenvalues in U.

For sufficiently small σ we are clearly in the second situation described in Lemma 4.4. We now prove that $\lambda_1 \to -2\Gamma_{\min}$ as $\sigma \to 0$. Indeed, if $L = \limsup_{\sigma\to 0}\lambda_1 > -2\Gamma_{\min}$, we obtain for each $k > 0$ the existence of a sequence $\lambda_j \to \lambda^* > -2\Gamma_{\min}$ and $\sigma_j \to 0$ such that $-k^2$ is a positive definite eigenvalue of $T(\lambda_j)$. The corresponding eigenfunctions solve

$$\begin{cases} (a_{\lambda_j}^3 W_j')' = k^2 a_{\lambda_j} W_j, \quad p_0 < p < 0, \\ a_{\lambda_j}^3(0) W_j'(0) = (k^2\sigma_j + g) W_j(0), \\ W_j(p_0) = 0. \end{cases}$$

Letting $j \to \infty$ we obtain a solution of the eigenvalue problem

$$\begin{cases} -(a_{\lambda^*}^3 W')' = \mu a_{\lambda^*} W, \quad p_0 < p < 0, \\ a_{\lambda^*}^3(0) W'(0) = g W(0), \\ W(p_0) = 0, \end{cases} \tag{5.2}$$

with $\mu = -k^2$. This would mean that for each $k > 0$ there is a $\lambda^* \in [L, \lambda_0]$ such that $-k^2$ is an eigenvalue of the above problem. This is clearly not possible since it is known that (5.2) has a unique nonpositive eigenvalue μ for each $-2\Gamma_{\min} < \lambda \le \lambda_0$ and that $\mu(\lambda)$ is a continuous function of λ in this interval.

To summarize, as $\sigma \to 0$ there is a family of intervals $I_\sigma \subset I$ and real numbers R_σ such that $I_\sigma \to I$ and $R_\sigma \to -\infty$ as $\sigma \to 0$ and for $\lambda \in I_\sigma$ with $\lambda \le \lambda_0$ the nonpositive part of the spectrum of T consists of the eigenvalues μ_- and μ_+ with $\mu_- \le R_\sigma \le \mu_+$, while for $\lambda \ge \lambda_0$ the nonpositive part of the spectrum just consists of $\mu_- \le R_\sigma$. Thus at a formal level we recover the results [7], where the bifurcation points are given by the $\lambda^* \in I$ such that $\mu = -(kn)^2$ is an eigenvalue of (5.2). The eigenvalue problem (5.2) has one negative eigenvalue for each $\lambda \in (-2\Gamma_{\min}, \lambda_0)$ (and none in $\lambda > \lambda_0$) and the negative eigenvalue is an increasing function of λ in this interval. This curve of eigenvalues is simply the limit of the smallest positive definite eigenvalue of T for $\lambda < \lambda_0$ as $\sigma \to 0$.

6. Examples.

6.1. Irrotational flow. We will now take a closer look at irrotational flows, $\gamma \equiv 0$. In this case $\Gamma_{\min} = 0$ and the eigenvalue problem is, for $\lambda > 0$,

$$\begin{cases} \lambda W'' = (kn)^2 W \quad \text{in} \quad p_0 < p < 0, \\ \lambda^{3/2} W'(0) = ((kn)^2\sigma + g) W(0), \\ W(p_0) = 0. \end{cases}$$

Letting W satisfy the boundary condition at p_0 and $W_p(p_0) \ne 0$, we get $W(p) = C\sinh\left(\frac{kn(p-p_0)}{\sqrt{\lambda}}\right)$. Thus the condition for $-(kn)^2$ to be an eigenvalue is that

$$kn\lambda\cosh\Big(\frac{kn|p_0|}{\sqrt{\lambda}}\Big) = ((kn)^2\sigma + g)\sinh\Big(\frac{kn|p_0|}{\sqrt{\lambda}}\Big),$$

that is,

$$f(kn;\lambda) := \frac{(kn)^2\sigma + g}{kn\lambda}\tanh\Big(\frac{kn|p_0|}{\sqrt{\lambda}}\Big) = 1.$$

The depth corresponding to uniform flow is $d = |p_0|/\sqrt{\lambda}$ and $\lambda_0 = (g|p_0|)^{2/3}$. Using the integral condition in Theorem 4.1 with $a_{\lambda_0}(p) = \lambda_0^{1/2}$, it is easy to see that $\sigma_0 = 1/3(g|p_0|^4)^{1/3}$ is the unique value of σ for which 0 is a neutral eigenvalue of T at λ_0. For $\sigma > \sigma_0$ the eigenvalue 0 at λ_0 is negative definite, while it is positive definite for $\sigma < \sigma_0$. Note that $f(x;\lambda) \to \infty$ as $x \to \infty$ for each $\lambda > 0$, while $f(x;\lambda)$ is a strictly decreasing function of λ with $f(x;\lambda) \to \infty$ uniformly as $\lambda \to 0$. It follows that if $\sigma \geq \sigma_0$, the operator $T(\lambda)$ has no negative eigenvalues for $\lambda < \lambda_0$, while if $\sigma < \sigma_0$, we are in the second situation described in Lemma 4.4 with $\lambda_1 > 0$, and $T(\lambda)$ has no negative eigenvalues for $\lambda < \lambda_1$. This could also be seen directly from Lemma 4.4, since $\gamma \leq 0$.

The uniform flow corresponding to the bifurcation point λ^* has the velocity components $(c - u^*, v^*) = (\sqrt{\lambda^*}, 0)$, since

$$H(p;\lambda^*) = \frac{p - p_0}{\sqrt{\lambda^*}}.$$

We thus obtain the dispersion relation [19, 28] for linearized small amplitude waves traveling on a uniform current of velocity $(u^*, 0)$:

$$c - u^* = \sqrt{\frac{(kn)^2\sigma + g}{kn}\tanh(knd)}.$$

This relation was obtained for a fixed p_0. However, for any real $d > 0$ we can choose $c - u^*$ according to the above dispersion relation. Setting $|p_0| = d(c - u^*)$, this will in principle yield a curve of solutions having minimal period $2\pi/(kn)$ bifurcating at $\lambda^* = c - u^*$. The problem is that we may be in case 2a of Theorem 3.11. The bifurcation point is simple or double according to whether or not there exists an $m \in \mathbb{Z}^+$ such that m gives the same intrinsic wave speed $c - u^*$ as n. If furthermore $n \mid m$, we cannot tell if there are solutions of minimal period $2\pi/(kn)$.

Let us now try to relate our results in the irrotational case with the work of Jones [20]. In [20] the bifurcation parameter is the speed c and d is held fixed, while $|p_0|$ is allowed to vary. Furthermore, the horizontal speed u of the trivial flow is taken to be zero. For every $n \in \mathbb{Z}^+$ a bifurcation point is obtained at

$$c_n^2 = f(kn) := \frac{(kn)^2\sigma + g}{kn}\tanh(knd),$$

and c_n is double or simple depending upon whether there is an $m \in \mathbb{Z}^+$ as above with $c_n = c_m$. If $\sigma/gd^2 \geq 1/3$, the function $f(x)$, $x \geq 0$, is strictly increasing with the minimum value gd at 0. In this case all bifurcation points are simple. If $\sigma/gd^2 < 1/3$, the function f has a single minimum at some point $x_0 > 0$, is strictly decreasing between 0 and x_0, and is strictly increasing in $x > x_0$. It follows that there is some $x_1 > x_0$ with $f(x_1) = f(0) = gd$. For every $x \in (0, x_0)$ there is an $x' \in (x_0, x_1)$, depending continuously on x, with $f(x') = f(x)$. Moreover, as $x \to 0$ we have $x' \to x_1$. Thus we can always choose x' such that $x'/x \in \mathbb{Q}$. In this way it is always possible to find a double bifurcation point (for some wave number) if $\sigma/gd^2 < 1/3$.

The condition on σ/gd^2 is useful also in our setting, but it needs some explaining. Replacing c by $\lambda = c - u$, we obtain precisely the same bifurcation points. However, since we let the depth vary, while holding p_0 fixed, the conditions $\sigma/gd^2 < 1/3$ and $\mu_-(\lambda_0) < 0$ are related, but the correspondence is not one-to-one. What is true is that $\mu_-(\lambda_0) < 0$ if and only if $\sigma/gd_0^2 < 1/3$, where d_0 is the depth corresponding to

the trivial solution at λ_0. This follows from expressing $|p_0|$ in terms of d_0 and g and substituting this in the formula for σ_0. If we start with any bifurcation point λ^* such that $\sigma/gd^2 \geq 1/3$, it follows that $\lambda_0 = (g|p_0|)^{2/3} = (gd\sqrt{\lambda^*})^{2/3} < \lambda^*$. Hence, the bifurcation point λ^* is in this case simple. Note, however, that it does not follow that $\mu_-(\lambda_0) = 0$ for the point λ_0 on our corresponding trivial solution curve (it is easy to construct a counterexample). If $\sigma/gd^2 < 1/3$ and we choose kn with $\lambda^* = f(kn) < gd$, then $\lambda^* < \lambda_0$. Hence, $\mu_-(\lambda_0) < 0$ in this case.

6.2. Constant vorticity. In the case of constant vorticity $\gamma \neq 0$, the substitution

$$W(p) = \frac{2\gamma}{\sqrt{\lambda + 2\gamma p}} W_0\Big(\frac{\sqrt{\lambda + 2\gamma p}}{\gamma}\Big)$$

transforms the differential equation (3.6) into $W_0'' = (kn)^2 W_0$. We infer that

$$W(p) = \frac{1}{\sqrt{\lambda + 2\gamma p}} \sinh\Big(\frac{kn(\sqrt{\lambda + 2\gamma p} - \sqrt{\lambda + 2\gamma p_0})}{\gamma}\Big).$$

The boundary condition at $p = 0$ is then $f(kn; \lambda) = 1$, where

$$f(x;\lambda) := \frac{x^2\sigma + g + \gamma\sqrt{\lambda}}{x\lambda} \tanh\Big(\frac{2x|p_0|}{\sqrt{\lambda} + \sqrt{\lambda + 2\gamma p_0}}\Big), \quad x \geq 0.$$

Note that $\lim_{x\to\infty} f(x;\lambda) = \infty$ for each $\lambda > -2\Gamma_{\min}$. As in the irrotational case, there is a σ_0 such that 0 is a neutral eigenvalue at λ_0 exactly for $\sigma = \sigma_0$. We discern between the following two cases.

Case 1. Suppose that $\gamma < 0$. In this case $\Gamma_{\min} = 0$ and for $x^2\sigma + g + \gamma\sqrt{\lambda} > 0$ the function f is strictly decreasing in λ. Moreover, we have $f(x;\lambda) \to \infty$ uniformly as $\lambda \to 0$. It follows that the situation is as in the case of $\gamma \equiv 0$, that is, if $\sigma \geq \sigma_0$, there are no negative eigenvalues for $\lambda < \lambda_0$, while if $\sigma < \sigma_0$, we have $\lambda_1 > 0$ and there are no negative eigenvalues for $\lambda < \lambda_1$.

Case 2. Suppose instead that $\gamma > 0$. In this case we have $\Gamma_{\min} = \gamma p_0$ so that f is defined for $\lambda > 2\gamma|p_0|$. The function $f(x;\lambda)$ is now strictly decreasing in λ for all $x \geq 0$ and the limit as $\lambda \to 2\gamma|p_0|$ is

$$f(x;0) = \frac{\sigma x^2 + g + \gamma\sqrt{2\gamma|p_0|}}{2x\gamma|p_0|} \tanh\left(\sqrt{\frac{2|p_0|}{\gamma}}x\right).$$

If $\sigma \geq \sigma_0$, the situation is exactly as before. Suppose instead that $\sigma < \sigma_0$. If the minimum value of $f(x;0)$ is greater than one, then as earlier $\lambda_1 > -2\Gamma_{\min}$ and to the left of λ_1 the operator $T(\lambda)$ has no negative eigenvalues. However, if $\min f(x;0) \leq 1$, then $\lambda_1 = -2\Gamma_{\min}$ so that for each $\lambda \in (-2\Gamma_{\min}, \lambda_0)$ the operator $T(\lambda)$ has two negative eigenvalues. Both alternatives can be achieved by varying σ while keeping the other variables fixed.

The trivial flow corresponding to λ^* has the velocity components $(c - u^*, v^*) = (\sqrt{\lambda^*} + \gamma y, 0)$ because $u_y = -\gamma$. Note that

$$H(p;\lambda^*) = \frac{\sqrt{\lambda^* + 2\gamma p} - \sqrt{\lambda^* + 2\gamma p_0}}{\gamma}.$$

From the definition of p_0 we infer that $\sqrt{\lambda^*}d - \frac{\gamma}{2}d^2 = |p_0|$ and thus

$$d = \frac{2|p_0|}{\sqrt{\lambda^*} + \sqrt{\lambda^* + 2\gamma p_0}}.$$

For $y = 0$ we obtain the dispersion relation

$$c - u_0^* = \frac{\gamma}{2kn}\tanh(knd) + \sqrt{\frac{(kn)^2\sigma + g}{kn}\tanh(knd) + \frac{\gamma^2}{4(kn)^2}\tanh^2(knd)},$$

where u_0^* is the speed of the trivial flow at the surface. This is the dispersion relation for linearized small-amplitude waves traveling on a linearly sheared current of constant vorticity γ.

Acknowledgment. The author is grateful to the referees for suggestions which significantly improved the paper.

REFERENCES

[1] J. Bognár, *Indefinite Inner Product Spaces*, Springer-Verlag, New York, Heidelberg, 1974.
[2] A. Constantin, *On the deep water wave motion*, J. Phys. A, 34 (2001), pp. 1405–1417.
[3] A. Constantin, *Edge waves along a sloping beach*, J. Phys. A, 34 (2001), pp. 9723–9731.
[4] A. Constantin and J. Escher, *Symmetry of steady periodic surface water waves with vorticity*, J. Fluid Mech., 498 (2004), pp. 171–181.
[5] A. Constantin and J. Escher, *Symmetry of steady deep-water waves with vorticity*, European J. Appl. Math., 15 (2004), pp. 755–768.
[6] A. Constantin and W. Strauss, *Exact periodic traveling water waves with vorticity*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 797–800.
[7] A. Constantin and W. Strauss, *Exact steady periodic water waves with vorticity*, Comm. Pure Appl. Math., 57 (2004), pp. 481–527.
[8] W. Craig and D. P. Nicholls, *Traveling two and three dimensional capillary gravity water waves*, SIAM J. Math. Anal., 32 (2000), pp. 323–359.
[9] M. Crandall and P. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.
[10] G. D. Crapper, *An exact solution for progressive capillary waves of arbitrary amplitude*, J. Fluid Mech., 2 (1957), pp. 532–540.
[11] G. D. Crapper, *Introduction to Water Waves*, Ellis Horwood, Chichester, Halsted Press, New York, 1984.
[12] F. Dias and C. Kharif, *Nonlinear gravity and capillary-gravity waves*, in Annual Review of Fluid Mechanics, Vol. 31, Annual Reviews, Palo Alto, CA, 1999, pp. 301–346.
[13] M.-L. Dubreil-Jacotin, *Sur la détermination rigoureuse des ondes permanentes périodiques d'ampleur finie*, J. Math. Pures Appl., 13 (1934), pp. 217–291.
[14] M. Ehrnström, *Uniqueness of steady symmetric deep-water waves with vorticity*, J. Nonlinear Math. Phys., 12 (2005), pp. 27–30.
[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001.
[16] M. D. Groves, *Steady water waves*, J. Nonlinear Math. Phys., 11 (2004), pp. 435–460.
[17] L. Hörmander, *Fourier integral operators.* I, Acta Math., 127 (1971), pp. 79–183.
[18] I. S. Iohvidov, M. G. Krein, and H. Langer, *Introduction to the Spectral Theory of Operators in Spaces with an Indefinite Metric*, Akademie-Verlag, Berlin, 1982.
[19] R. S. Johnson, *A Modern Introduction to the Mathematical Theory of Water Waves*, Cambridge University Press, Cambridge, UK, 1997.
[20] M. Jones, *Small amplitude capillary-gravity waves in a channel of finite depth*, Glasgow Math. J., 31 (1989), pp. 141–160.
[21] M. Jones and J. Toland, *The bifurcation and secondary bifurcation of capillary-gravity waves*, Proc. Roy. Soc. London Ser. A, 399 (1985), pp. 391–417.
[22] M. Jones and J. Toland, *Symmetry and the bifurcation of capillary-gravity waves*, Arch. Rational Mech. Anal., 96 (1986), pp. 29–53.

[23] H. Kalisch, *Periodic traveling water waves with isobaric streamlines*, J. Nonlinear Math. Phys., 11 (2004), pp. 461–471.
[24] H. Kalisch, *A uniqueness result for periodic traveling waves in water of finite depth*, Nonlinear Anal., 58 (2004), pp. 779–785.
[25] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1976.
[26] W. Kinnersley, *Exact large amplitude capillary waves on sheets of fluid*, J. Fluid Mech., 77 (1976), pp. 229–241.
[27] H. Lamb, *Hydrodynamics*, Cambridge University Press, London, 1924.
[28] J. Lighthill, *Waves in Fluids*, Cambridge University Press, Cambridge, UK, 1978.
[29] Y. Luo and N. Trudinger, *Linear second order elliptic equations with Venttsel boundary conditions*, Proc. Roy. Soc. Edinburgh Sect. A, 118 (1991), pp. 193–207.
[30] L. Nirenberg, *Topics in Nonlinear Functional Analysis*, AMS, Providence, RI, 2001.
[31] D. P. Nicholls and F. Reitich, *On analyticity of traveling water waves*, Proc. Roy. Soc. London Ser. A, 461 (2005), pp. 1283–1309.
[32] H. Okamoto and M. Shoji, *The Mathematical Theory of Permanent Progressive Water-Waves*, World Scientific, River Edge, NJ, 2001.
[33] J. Reeder and M. Shinbrot, *On Wilton ripples.* I. *Formal derivation of the phenomenon*, Wave Motion, 3 (1981), pp. 115–135.
[34] J. Reeder and M. Shinbrot, *On Wilton ripples.* II. *Rigorous results*, Arch. Rational Mech. Anal., 77 (1981), pp. 321–347.
[35] T. da Silva and D. H. Peregrine, *Steep, steady surface waves on water of finite depth with constant vorticity*, J. Fluid Mech., 195 (1988), pp. 281–302.
[36] C. Swan, I. Cummins, and R. James, *An experimental study of two-dimensional surface water waves propagating on depth-varying currents*, J. Fluid Mech., 428 (2001), pp. 273–304.
[37] G. Thomas and G. Klopman, *Wave-current interactions in the nearshore region*, in Gravity Waves in Water of Finite Depth, Adv. Fluid Mech. 10, Computational Mechanics Publications, Southampton, UK, 1997, pp. 215–319.
[38] J. F. Toland, *Stokes waves*, Topol. Methods Nonlinear Anal., 7 (1996), pp. 1–48.
[39] E. Wahlén, *Uniqueness for autonomous planar differential equations and the Lagrangian formulation of water flows with vorticity*, J. Nonlinear Math. Phys., 11 (2004), pp. 549–555.
[40] E. Wahlén, *A note on steady gravity waves with vorticity*, Int. Math. Res. Not., 7 (2005), pp. 389–396.
[41] E. Wahlén, *Steady periodic capillary waves with vorticity*, Ark. Mat., to appear.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 944–976

MULTISCALE ANALYSIS BY Γ-CONVERGENCE OF A ONE-DIMENSIONAL NONLOCAL FUNCTIONAL RELATED TO A SHELL-MEMBRANE TRANSITION*

NADIA ANSINI†, ANDREA BRAIDES‡, AND VANDA VALENTE§

Abstract. We study the asymptotic behavior of one-dimensional functionals associated with the energy of a thin nonlinear elastic spherical shell in the limit of vanishing thickness (proportional to a small parameter) ε and under the assumption of radial deformations. The functionals are characterized by the presence of a nonlocal potential term and defined on suitable weighted functional spaces. The shell-membrane transition is studied at three different relevant scales. For each we give a compactness result and compute the Γ-limit. In particular, we show that if the energies on a sequence of configurations scale as $\varepsilon^{3/2}$, then the limit configuration describes a (locally) finite number of transitions between the undeformed and the everted configurations of the shell. We also highlight a kind of "Gibbs phenomenon" by showing that nontrivial optimal sequences restricted between the undeformed and the everted configurations must have energy scaling of at least $\varepsilon^{4/3}$.

Key words. Γ-convergence, multiscale analysis, shell theory, membrane theory, relaxation, energy minimization, Gibbs phenomenon

AMS subject classifications. 49J45, 74K25, 74K15, 74B20, 74G65

DOI. 10.1137/050630829

1. Introduction.

It is well known that a spherical shell under zero loads can assume at least two configurations: the trivial reference configuration and the everted configuration, as they are both stable solutions of the equilibrium problem. The existence of everted shapes was first proved by Antman [1] for thick spherical shells. Later, the analysis concerning the eversion of thin shells was carried out from the theoretical and numerical points of view by Podio-Guidugli et al. [20] and Geymonat, Rosati, and Valente [15, 16].

The energy functional corresponding to the axially symmetric deformations of a spherical cap, without applied loads, may be written as

$$F_\varepsilon(u) = \varepsilon^2 \int_0^1 \theta^3 (u'(\theta))^2 \, d\theta + \int_0^1 \frac{1}{\theta^3} \left(\int_0^\theta \varphi^3 (u^2(\varphi) - 1) \, d\varphi \right)^2 d\theta \,.$$

The unknown u, which is a function of the normalized polar angle θ, is related to the slope of the deformed middle surface of the cap with respect to the initial spherical shape, and ε is the thinness parameter of the shell (for the derivation of the model

*Received by the editors May 6, 2005; accepted for publication (in revised form) March 30, 2006; published electronically October 12, 2006. This work has been partially supported by the European Community's Human Potential Programme under contract HPRN-CT-2002-00284 (SMART-SYSTEMS) and by the CNR/MIUR project "Materiali compositi per applicazioni strutturali di rilevante interesse industriale."

http://www.siam.org/journals/sima/38-3/63082.html

†Dipartimento di Matematica, Università di Roma "La Sapienza," piazzale A. Moro, 2, 00185 Rome, Italy (ansini@mat.uniroma1.it).

‡Dipartimento di Matematica, Università di Roma "Tor Vergata," via della Ricerca Scientifica, 00133 Rome, Italy (braides@mat.uniroma2.it). This author's work is part of the Marie Curie Research Training Network MRTN-CT-2004-505226 "Multi-scale modelling and characterization for phase transformations in advanced materials" (MULTIMAT).

§Istituto per le Applicazioni del Calcolo "M. Picone," viale del Policlinico, 137, 00161 Rome, Italy (valente@iac.rm.cnr.it).

and physical interpretation of the variable, see the pioneering papers by Reiss [21], Bauer, Reiss, and Keller [2], and Podio-Guidugli et al. [20]; see also the functional approach introduced by Berger [4] for the mathematical study of the thin elastic plate buckling).

In [20, 15, 16], under suitable boundary conditions, it has been proved that, beside the trivial stable solution $u = 1$, for ε small enough there exists a second stable solution, namely, the everted stressed solution. The sequence of everted configurations tends toward an unstressed configuration ($u = -1$) that can be described as the reflection of the cap reference configuration. More recently, the problem has been further investigated. The existence of infinitely many stable solutions for the limit problem has been predicted, and several numerical experiments have been proposed by Geymonat and Leger [14].

Although the problem has been carefully studied in [20, 15, 16] for ε small, the asymptotic analysis of F_ε has remained an open problem, and it is the object of the present work, analyzed using the notation and techniques of variational calculus.

To describe the asymptotic behavior of F_ε we first focus on sequences (u_ε) such that $F_\varepsilon(u_\varepsilon) = O(1)$. In that case we prove that (u_ε) is locally weakly compact in $L^1(0,1)$, and sequences giving the optimal lower bound may oscillate between the values -1 and 1. This behavior is described by the Γ-limit F^0 (see the next section for the precise form of the limit) that not only captures these oscillations but also shows that the nonlocal character of the functional is maintained in the limit. Minimizing sequences are responsible for folding effects, which are also observed for flat membranes. The analytical reconstruction of the shell surface texture could allow us to both understand the material elastic properties and study the interactions between the two surfaces.

It must be noted that the Γ-limit F^0 coincides with the lower-semicontinuous envelope of the functional

$$G(u) = \int_0^1 \theta^{-3} \left(\int_0^\theta (u^2 - 1)\varphi^3 \, d\varphi \right)^2 d\theta,$$

with respect to the local weak L^1-convergence, and that minimizers of this functionals are all functions with $|u| = 1$ a.e. In terms of recovery sequences, we note that they may develop oscillations, but the occurrence of these is due to a nonlocal effect (see the example in Remark 3).

The minimum value for the Γ-limit F^0 is 0 and is achieved exactly on all functions u with $|u| \leq 1$. This large class of minimizers justifies the analysis at finer scales. We show that the next meaningful scale occurs when $F_\varepsilon(u_\varepsilon) = O(\varepsilon^{3/2})$. If this is the case, then we show that such a (u_ε) is strongly precompact in $L^1(0,1)$, its limits u are locally piecewise constant in $(0,1)$, and $|u(\theta)| = 1$ a.e. We describe this behavior by showing that the Γ-limit of the scaled energies $\varepsilon^{-3/2} F_\varepsilon$ on those functions takes the form

$$F^{3/2}(u) = c_0 \sum_{\theta \in S(u)} \theta^3,$$

where we denote by $S(u)$ the set of discontinuities of u (see Theorem 3.3).

The mechanical interpretation of this energetic asymptotic description is that minimal states of the energies F_ε subjected to boundary conditions will be approximately described for ε small by a mixture of undeformed and everted configurations,

the transitions between which are such as to minimize the energy $F^{3/2}$. A similar analysis can be performed for suitably scaled forcing terms (in which case the corresponding forcing term must be added to $F^{3/2}$).

The formal analogies with the corresponding functionals of the gradient theory of phase transitions

$$H_\varepsilon(u) = \varepsilon^2 \int_0^1 \theta^3 (u'(\theta))^2 \, d\theta + \int_0^1 \theta^3 (u^2(\theta) - 1)^2 \, d\theta$$

must be noted. Upon a normalization factor, the functionals $\varepsilon^{-1} H_\varepsilon$ also Γ-converge to $F^{3/2}$. Apart from the different scaling ε^{-1}, this analogy does not extend to the details of the proof. First, it must be noted that the compactness properties for functions with $F_\varepsilon(u_\varepsilon) = O(\varepsilon^{3/2})$ are much more difficult to prove by the cancellations that may occur in the integral $\int_0^\theta \varphi^3 (u_\varepsilon^2 - 1) \, d\varphi$ owing to the fact that $u_\varepsilon^2 - 1$ may change sign. Second, the way the constant c_0 is computed involves some optimal transitions that exhibit a sort of *Gibbs phenomenon*: even though their limit takes only the value ± 1, these transitions must take values external to the interval $[-1, 1]$.

In the final section, we show that this Gibbs phenomenon is substantial: if we impose the constraint $|u_\varepsilon| \le 1$ onto a sequence (u_ε) converging to u, then the values $\varepsilon^{-3/2} F_\varepsilon(u_\varepsilon)$ cannot converge to the value $F^{3/2}(u)$, and they must even diverge. We show that with this additional constraint the correct scaling is $\varepsilon^{-4/3}$. The scaled energies still converge to a phase-transition functional, but this time in a nonlocal form (see Theorem 4.2).

We believe that the techniques developed here can be adapted to other transition problems in nonlinear elasticity, leading to the study of functionals with similar nonlocal terms. Moreover, our analysis can be generalized to different weighted spaces. For these reasons we report our results with reference to a more general weighted functional (details follow in the next section).

Finally, we point out that functionals F_ε are derived from the scaled energy

$$\frac{1}{\varepsilon} \int_{\mathcal{C}_\varepsilon} \left(\mu |\mathbf{D}|^2 + \frac{\lambda}{2} (\operatorname{trace} \mathbf{D})^2 \right) dx,$$

where $\mathcal{C}_\varepsilon$ parameterizes a thin spherical shell of thickness ε, $\mathbf{D} = \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T) + \frac{1}{2}(\nabla \mathbf{u} \nabla \mathbf{u}^T)$ is the *nonlinear deformation tensor* related to the deformation $\mathbf{u}$ of the shell, and λ, μ are the Lamé constants (we refer, e.g., to [20] for the precise derivation). In this way, our paper may be partly related to recent works on dimension reduction for thin structures by the use of Γ-convergence (see, e.g., Le Dret and Raoult [17] for the limit analysis of thin shells, Friesecke, James, and Müller [13] and Conti and Maggi [10] for the analysis under various scaling, Ben Belgacem et al. [3] for complex patterns in recovery sequences, and Braides, Fonseca, and Francfort [9] for an example of application of the localization methods of Γ-convergence to thin structures).

1.1. Analytical description of the results. We conclude this introduction with a brief analytical description of the results in a more general setting, where we replace the weight θ^3 with a more general weight $\rho(\theta)$. We use the techniques of Γ-convergence for the asymptotic analysis of our functionals; for the definition, notation, and comprehensive study of Γ-convergence we refer to [6] and [11] (see also [7] and [8, Part 2]).

Let $\rho : [0, 1] \mapsto \mathbf{R}$ be a nondecreasing, continuous function, strictly positive on

$(0,1]$. For all $\varepsilon > 0$ and $\alpha \geq 0$ we define

$$F_\varepsilon^\alpha(u) = \varepsilon^{2-\alpha} \int_0^1 \rho\,(u')^2\,d\theta + \varepsilon^{-\alpha} \int_0^1 \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho\,(u^2-1)\,d\varphi \right)^2 d\theta$$

for $u \in H^1(0,1)$. When needed, the functionals are understood to take the value $+\infty$ when not otherwise defined. We will isolate particular values of α for which the Γ-limit is nontrivial.

We will consider separately the following cases.

Case 1 (section 2). $\alpha = 0$. In this case, minimizing sequences are weakly precompact in $L^2_{\mathrm{loc}}(0,1)$; hence, we compute the Γ-limit F^0 of F_ε^0 with respect to that convergence, and for every $u \in L^2_{\mathrm{loc}}(0,1)$ we get

$$F^0(u) = \min \left\{ \int_0^1 \frac{1}{\rho(\theta)} \Big(\mu([0,\theta)) - \int_0^\theta \rho\; d\varphi \Big)^2 d\theta \;:\; \mu \geq \rho\, u^2\, d\varphi \right\},$$

where the minimum is taken over all nonnegative measures μ. The set of the minimum points of F^0 is $\{|u| \leq 1\}$.

Case 2 (section 3). $\alpha = 3/2$. In this case, we scale F_ε^0 further and study the limit of $F_\varepsilon^{3/2} = \varepsilon^{-3/2} F_\varepsilon^0$. We prove that minimizing sequences are precompact with respect to the strong $L^1_{\mathrm{loc}}(0,1)$-convergence, and their limits u belong to $BV_{\mathrm{loc}}((0,1);\{-1,1\})$; i.e., u is locally piecewise constant on $(0,1)$ and takes only the values 1 and -1. We compute the Γ-limit $F^{3/2}$ of $F_\varepsilon^{3/2}$ with respect to that convergence and we get

$$F^{3/2}(u) = c_0 \sum_{\theta \in S(u)} \rho(\theta)$$

for every $u \in BV_{\mathrm{loc}}((0,1);\{-1,1\})$, where $S(u)$ is the set of points where u jumps between the points 1 and -1, and

$$c_0 = \inf_{T>0} \inf \left\{ \int_{-T}^T (v')^2\,ds + \int_{-T}^T \left(\int_{-T}^\sigma (v^2-1)\,ds \right)^2 d\sigma \;: \right.$$
$$\left. v \in H^1(-T,T), \quad v(\pm T) = \pm 1, \quad \int_{-T}^T (v^2-1)\,ds = 0 \right\}.$$

With this choice of the scaling we get a result of "Modica–Mortola" type with a different characterization of the constant c_0 (see [18, 19], or [5]).

Case 3 (section 4). In this case, we show that the characteristic scale changes if we impose the restriction that $u \in H^1((0,1);[-1,1])$, and there the correct scaling power is $\alpha = 4/3$. We treat only the case $\rho = 1$ for the sake of simplicity. By Case 1 above, we have that the Γ-limit of the restriction of F_ε^0 to $H^1((0,1);[-1,1])$, G_ε^0, is zero. We then rescale G_ε^0 to get a nontrivial limit problem, considering the family of functionals $G_\varepsilon^{4/3} = \varepsilon^{-4/3} G_\varepsilon^0$, and we prove that the minimizing sequences are compact

with respect to the strong L^1-convergence and its Γ-limit is nonlocal,

$$G^{4/3}(u) = \inf_{T>0} \inf \left\{ \sum_{i\in I} \left(\int_{-T}^{T} (v_i')^2 \, ds + (\theta_{i+1} - \theta_i) \left(\sum_{j=1}^{i} \int_{-T}^{T} (v_j^2 - 1) \, ds \right)^2 \right) : \right.$$

$$\left. v_i \in H^1((-T,T);[-1,1]), \quad v_i(\pm T) = \pm 1 \right\},$$

for every $u \in BV_{\rm loc}((0,1);\{-1,1\})$, where we have labeled the points in $S(u)$ by a set of indices $I \subset \mathbf{N}$ in such a way that $\theta_i < \theta_{i+1}$.

2. The case $\alpha = 0$: Oscillations. We consider the case $\alpha = 0$ first; i.e.,

$$F_\varepsilon^0(u) = \varepsilon^2 \int_0^1 \rho \, (u')^2 \, d\theta + \int_0^1 \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho \, (u^2 - 1) \, d\varphi \right)^2 d\theta$$

for $u \in H^1_{\rm loc}(0,1)$. In order to choose the topology in which to frame our limit problem we have to examine the compactness properties of sequences with bounded energy. Note that the presence of ε only in the first term of $F_\varepsilon^0(u)$ suggests the use of the weak L^2-convergence.

THEOREM 2.1 (compactness). *Let (u_ε) be a sequence such that $\sup_\varepsilon F_\varepsilon^0(u_\varepsilon) < +\infty$; then, up to subsequences, the measures $\rho u_\varepsilon^2 \, d\varphi$ converge weakly* in $(-\infty, 1)$ and, in particular, (u_ε) converges weakly in $L^2_{\rm loc}(0,1)$.*

Proof. By assumption, $\sup_\varepsilon \int_0^1 (1/\rho(\theta)) (\int_0^\theta \rho \, (u_\varepsilon^2 - 1) \, d\varphi)^2 \, d\theta < +\infty$, which implies that $\sup_\varepsilon \int_0^1 (\int_0^\theta \rho \, u_\varepsilon^2 \, d\varphi)^2 \, d\theta < +\infty$; hence, by Hölder's inequality, there exists a constant c independent of ε such that

$$\int_0^1 \left(\int_0^\theta \rho \, u_\varepsilon^2 \, d\varphi \right) d\theta \le c \tag{2.1}$$

for every $\varepsilon > 0$. By the monotonicity of $\theta \mapsto \int_0^\theta \rho \, u_\varepsilon^2 \, d\varphi$ we have that for a fixed $\theta_0 \in (0,1)$,

$$\int_0^{\theta_0} \rho \, u_\varepsilon^2 \, d\varphi \le \int_0^\theta \rho \, u_\varepsilon^2 \, d\varphi$$

for every $\theta \ge \theta_0$. Hence we get that

$$\int_{\theta_0}^1 \left(\int_0^{\theta_0} \rho \, u_\varepsilon^2 \, d\varphi \right) d\theta \le \int_{\theta_0}^1 \left(\int_0^\theta \rho \, u_\varepsilon^2 \, d\varphi \right) d\theta \le \int_0^1 \left(\int_0^\theta \rho \, u_\varepsilon^2 \, d\varphi \right) d\theta,$$

and by (2.1) we can conclude that for every $\theta_0 \in (0,1)$ there exists a constant $c(\theta_0)$ depending only on θ_0 such that $\int_0^{\theta_0} \rho \, u_\varepsilon^2 \, d\varphi \le c(\theta_0)$ for every $\varepsilon > 0$, which gives the desired compactness properties. ∎

Remark 1. The following example shows that we cannot expect weak compactness, but only local weak compactness, in $L^2(0,1)$ for a sequence (u_ε) with $\sup_\varepsilon F_\varepsilon^0(u_\varepsilon) < +\infty$. In fact, consider

$$u_\varepsilon(\theta) = \varepsilon^{-7/5} (\theta - \theta_\varepsilon)^+$$

with $\theta_\varepsilon = 1 - \varepsilon^{4/5}$; then $F^0_\varepsilon(u_\varepsilon) \le c$ for every $\varepsilon > 0$, but

$$\int_0^1 u_\varepsilon^2 \, d\theta = \frac{1}{3}\varepsilon^{-2/5} .$$

LEMMA 2.2. *Let*

$$G(u) = \int_0^1 \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho \, (u^2 - 1) \, d\varphi \right)^2 d\theta .$$

The lower-semicontinuous envelope of G with respect to the weak $L^2_{\rm loc}$ topology is

$$\overline{G}(u) := \min \left\{ \int_0^1 \frac{1}{\rho(\theta)} \left(\mu([0,\theta)) - \int_0^\theta \rho \, d\varphi \right)^2 d\theta \, : \, \mu \ge \rho u^2 \, d\varphi \right\} ,$$

where the minimum is taken in the space $\mathcal{M}^+([0,1))$ of locally finite positive measures on $[0,1)$.

Proof. Let (u_N) be a sequence weakly converging to u in $L^2_{\rm loc}$ such that

$$\lim_{N \to +\infty} G(u_N) < +\infty$$

and the sequence of positive measures $(\rho u_N^2 \, d\varphi)$ converges weakly* to μ in $\mathcal{M}^+([0,1))$. Then, by the lower semicontinuity of the L^2-norm and the convergence of the measures of almost all intervals, we have that

$$\frac{\mu(\theta - \delta, \theta + \delta)}{\delta} = \lim_{N \to +\infty} \fint_{\theta-\delta}^{\theta+\delta} \rho u_N^2 \, d\varphi \ge \fint_{\theta-\delta}^{\theta+\delta} \rho u^2 \, d\varphi$$

for almost all (a.a.) $\delta \in (0, \theta)$; by the Besicovitch derivation theorem, we conclude that for a.e. θ,

$$\frac{d\mu}{d\varphi}(\theta) \ge \rho(\theta) u^2(\theta) .$$

Since μ has at most countably many atoms, by the weak*-convergence in $\mathcal{M}^+([0,1))$, we have that

$$\lim_{N \to +\infty} \int_0^\theta \rho u_N^2 \, d\varphi = \mu([0,\theta)) \quad \text{for a.e. } \theta \in (0,1) . \tag{2.2}$$

It follows that

$$\frac{1}{\sqrt{\rho(\theta)}} \left(\int_0^\theta \rho \, (u_N^2 - 1) \, d\varphi \right) \quad \overset{L^2}{\rightharpoonup} \quad \frac{1}{\sqrt{\rho(\theta)}} \left(\mu([0,\theta)) - \int_0^\theta \rho \, d\varphi \right) ,$$

and by the weak lower semicontinuity of the L^2-norm, we get

$$\lim_{N \to +\infty} G(u_N) \ge \overline{G}(u) .$$

Note that the functional

$$\mu \mapsto \int_0^1 \frac{1}{\rho(\theta)} \left(\mu([0,\theta)) - \int_0^\theta \rho \, d\varphi \right)^2 d\theta$$

is weakly lower semicontinuous and coercive in $\mathcal{M}^+([0,1))$. Moreover, the set $\{\mu \geq \rho u^2\, d\varphi\}$ is convex; hence, the minimum is attained.

We now check the lim sup inequality for every $u \in L^2_{\rm loc}(0,1)$ such that $\overline{G}(u) < +\infty$. Let $\mu \in \mathcal{M}^+([0,1))$ be such that

$$\overline{G}(u) = \int_0^1 \frac{1}{\rho(\theta)} \left(\mu([0,\theta)) - \int_0^\theta \rho\, d\varphi \right)^2 d\theta\,.$$

For $0 < a < 1$ we define

$$u^a(\theta) = \begin{cases} u(\theta), & a \leq \theta \leq 1-a, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mu^a_N([0,\theta)) = \begin{cases} 0 & \text{if } \theta < a - 1/N, \\ \mu([0,a)) & \text{if } a - 1/N \leq \theta < a, \\ \mu([0,\theta)) & \text{if } a \leq \theta < 1-a, \\ \mu([0,1-a)) & \text{if } 1-a \leq \theta < 1 \end{cases}$$

for $N \in \mathbf{N}$ and $N > 1$. Then, $u^a \in L^2(0,1)$, $u^a \to u$, in $L^2_{\rm loc}(0,1)$ as $a \to 0^+$, μ^a_N converges to the measure μ^a defined by

$$\mu^a([0,\theta)) = \begin{cases} 0 & \text{if } \theta < a, \\ \mu([0,\theta)) & \text{if } a \leq \theta < 1-a, \\ \mu([0,1-a)) & \text{if } 1-a \leq \theta < 1 \end{cases}$$

(this can be checked, e.g., by using Theorem 1.16 in [12]), and

$$\overline{G}(u^a) \leq \int_a^1 \frac{1}{\rho(\theta)} \left(\mu^a([0,\theta)) - \int_0^\theta \rho\, d\varphi \right)^2 d\theta + o(1) \leq \overline{G}(u) + o(1) \tag{2.3}$$

as $a \to 0^+$. Indeed, since by assumption $\overline{G}(u) < +\infty$, then we have

$$\begin{aligned} &\int_a^1 \frac{1}{\rho(\theta)} \left(\mu^a([0,\theta)) - \int_0^\theta \rho\, d\varphi \right)^2 d\theta \\ &\quad = \int_a^{1-a} \frac{1}{\rho(\theta)} \left(\mu([0,\theta)) - \int_0^\theta \rho\, d\varphi \right)^2 d\theta \\ &\qquad + \int_{1-a}^1 \frac{1}{\rho(\theta)} \left(\mu([0,1-a)) - \int_0^\theta \rho\, d\varphi \right)^2 d\theta \\ &\quad \leq \overline{G}(u) + 2 \int_{1-a}^1 \frac{\mu([1-a,\theta))^2}{\rho(\theta)}\, d\theta \\ &\qquad + 2 \int_{1-a}^1 \frac{1}{\rho(\theta)} \left(\mu([0,\theta)) - \int_0^\theta \rho\, d\varphi \right)^2 d\theta \\ &\quad = \overline{G}(u) + o(1)\,, \end{aligned}$$

as $a \to 0^+$.

We denote $I_N = \{0, \dots, N-1\}$ and

$$\begin{cases} \overline{u}_N(\theta) = \fint_{i/N}^{(i+1)/N} u^a \, d\varphi \,, & i/N \le \theta \le (i+1)/N \,, \quad i \in I_N, \\ v_N(\theta) = N \, \mu_N^a([i/N, (i+1)/N)) \,, & i/N \le \theta \le (i+1)/N \,, \quad i \in I_N \,. \end{cases}$$

Note that, in particular,

$$\overline{u}_N(\theta) = 0 \,, \qquad 0 \le \theta \le [Na]/N,$$

and

$$v_N(\theta) = \begin{cases} 0, & 0 \le \theta \le ([Na]-1)/N, \\ N\mu([0,a)), & ([Na]-1)/N \le \theta \le [Na]/N \,. \end{cases}$$

Hence, we define

$$u_N(\theta) = \begin{cases} 0, & 0 \le \theta \le ([Na]-1)/N, \\ \sqrt{N\mu([0,a))/\rho(\theta)}, & ([Na]-1)/N \le \theta \le [Na]/N, \\ \overline{u}_N + \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2}, & i/N \le \theta \le i/N + 1/2N, \\ & \quad i = [Na], \dots, N-1, \\ c_N\Big(\overline{u}_N - \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2}\Big), & i/N + 1/2N \le \theta \le (i+1)/N, \\ & \quad i = [Na], \dots, N-1, \end{cases} \tag{2.4}$$

where

$$c_N(\theta) = \rho(\theta - 1/2N)/\rho(\theta) \,, \quad [Na]/N + 1/2N \le \theta \le 1$$

([t] denotes the *integer part* of t). Finally, it remains to fix ρ_N^i for $i = [Na], \dots, N-1$ such that

$$v_N(\theta) \ge \rho_N^i \, (\overline{u}_N(\theta))^2 \,, \quad i/N \le \theta \le (i+1)/N \,, \quad i = [Na], \dots, N-1 \,.$$

Since $\mu_N^a \ge \rho \, (u^a)^2 \, d\varphi$, for every $i = [Na], \dots, N-1$, there exists $\eta_N^i \in [i/N, (i+1)/N)$ such that

$$\begin{aligned} \mu_N^a[i/N, (i+1)/N) &\ge \int_{i/N}^{(i+1)/N} \rho \, (u^a)^2 \, d\varphi = \rho(\eta_N^i) \int_{i/N}^{(i+1)/N} (u^a)^2 \, d\varphi \\ &\ge \rho(\eta_N^i) N \left(\int_{i/N}^{(i+1)/N} u^a \, d\varphi \right)^2 = \rho(\eta_N^i) \frac{1}{N} (\overline{u}_N)^2 \,. \end{aligned}$$

Hence, we choose $\rho_N^i = \rho(\eta_N^i)$ in (2.4).

By definition, (u_N) is bounded in L^2; hence, up to subsequence, it converges weakly in L^2. To identify the weak limit function with u^a it is sufficient to check that $\lim_{N \to +\infty} \int_b^d u_N = \int_b^d u^a$ for every $(b, d) \subseteq (0, 1)$. In fact,

$$\int_0^{[Na]/N} u_N \, d\theta = \sqrt{N\mu([0,a))} \int_{([Na]-1)/N}^{[Na]/N} \frac{d\theta}{\sqrt{\rho(\theta)}} \le \frac{1}{\sqrt{N}} \sqrt{\frac{\mu([0,a))}{\rho(([Na]-1)/N)}} \,, \tag{2.5}$$

while, for $i = [Na], \dots, N-1$, we have that

$$\begin{aligned}
\int_{i/N}^{(i+1)/N} u_N \, d\theta &= \int_{i/N}^{i/N+1/2N} \overline{u}_N + \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2}\, d\theta \\
&\quad + \int_{i/N+1/2N}^{(i+1)/N} \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \left(\overline{u}_N - \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2} \right) d\theta \\
&= \int_{i/N}^{(i+1)/N} \overline{u}_N \, d\theta + \int_{i/N+1/2N}^{(i+1)/N} \overline{u}_N \left(\frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} - 1 \right) d\theta \\
&\quad + \int_{i/N+1/2N}^{(i+1)/N} \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2} \left(1 - \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \right) d\theta \\
&= \left(\int_{i/N}^{(i+1)/N} u^a \, d\theta \right) + \int_{i/N+1/2N}^{(i+1)/N} \overline{u}_N \left(\frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} - 1 \right) d\theta \\
&\quad + \int_{i/N+1/2N}^{(i+1)/N} \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2} \left(1 - \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \right) d\theta \,. \qquad (2.6)
\end{aligned}$$

If we sum up on i, by the Hölder inequality, we get that

$$\begin{aligned}
&\left| \sum_{i \geq [Na]} \int_{i/N+1/2N}^{(i+1)/N} \overline{u}_N \left(\frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} - 1 \right) d\theta \right| \qquad (2.7) \\
&\qquad \leq \left(\int_0^1 (\overline{u}_N)^2 \, d\theta \right)^{1/2} \left(\int_{a/2}^1 \left(1 - \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \right)^2 d\theta \right)^{1/2}
\end{aligned}$$

and

$$\begin{aligned}
&\left| \sum_{i \geq [Na]} \int_{i/N+1/2N}^{(i+1)/N} \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2} \left(1 - \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \right) d\theta \right| \qquad (2.8) \\
&\quad \leq \left(\sum_{i \geq [Na]} \int_{i/N+1/2N}^{(i+1)/N} (v_N/\rho_N^i) + (\overline{u}_N)^2 \, d\theta \right)^{1/2} \left(\int_{a/2}^1 \left(1 - \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \right)^2 d\theta \right)^{1/2} \\
&\quad \leq \left(\frac{\mu([0, 1-a))}{\rho(a/2)} + \int_0^1 (\overline{u}_N)^2 \, d\theta \right)^{1/2} \left(\int_{a/2}^1 \left(1 - \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \right)^2 d\theta \right)^{1/2} .
\end{aligned}$$

Note that the sequence $(\overline{u}_N)$ is bounded in the L^2-norm since it converges to u^a strongly; moreover,

$$\lim_{N \to +\infty} \int_{a/2}^1 \left(1 - \frac{\rho(\theta - \frac{1}{2N})}{\rho(\theta)} \right)^2 d\theta = 0 \,.$$

Hence, by (2.5), (2.6), (2.7), and (2.8), we can easily conclude that $\lim_{N \to +\infty} \int_b^d u_N = \int_b^d u^a$ for every $(b, d) \subseteq (0, 1)$ and, therefore, we have the weak convergence of (u_N) to u^a in $L^2(0, 1)$.

We now examine

$$G(u_N) = \int_0^a \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho\, u_N^2 - \int_0^\theta \rho\ d\varphi \right)^2 d\theta + \int_a^1 \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho\, u_N^2 - \int_0^\theta \rho\ d\varphi \right)^2 d\theta\,; \tag{2.9}$$

in particular, we have that

$$\begin{aligned} &\int_0^{[Na]/N} \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho\, u_N^2\, d\varphi \right)^2 d\theta \\ &\quad = N^2 \mu([0,a))^2 \int_{([Na]-1)/N}^{[Na]/N} \frac{1}{\rho(\theta)} \left(\theta - \frac{[Na]-1}{N} \right)^2 d\theta \\ &\quad \le \frac{1}{3N} \frac{\mu([0,a))^2}{\rho(([Na]-1)/N)}\,. \end{aligned} \tag{2.10}$$

Then, passing to the limit as $N \to +\infty$ in (2.10), we get that

$$\limsup_{N\to+\infty} \int_0^a \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho\, u_N^2 - \int_0^\theta \rho\ d\varphi \right)^2 d\theta = o(1)\,, \tag{2.11}$$

as $a \to 0^+$. It remains to study the second term in (2.10). For $i = [Na], \dots, N-1$, we have that

$$\begin{aligned} &\int_{i/N}^{(i+1)/N} \rho\, u_N^2\, d\theta \\ &\quad = \int_{i/N}^{i/N+1/2N} \rho \left(\frac{v_N}{\rho_N^i} + 2\overline{u}_N \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2} \right) d\theta \\ &\qquad + \int_{i/N+1/2N}^{(i+1)/N} \rho \left(\theta - \frac{1}{2N} \right) \left(\frac{v_N}{\rho_N^i} - 2\overline{u}_N \sqrt{(v_N/\rho_N^i) - (\overline{u}_N)^2} \right) d\theta \\ &\quad = 2 \int_{i/N}^{i/N+1/2N} \frac{\rho(\theta)}{\rho(\eta_N^i)}\, v_N(\theta)\, d\theta\,, \end{aligned} \tag{2.12}$$

where $\eta_N^i \in [i/N, (i+1)/N)$, $i = [Na], \dots, N-1$. We recall that $\rho u_N^2 = 0$ in $[0, ([Na]-1)/N) \cup [([N(1-a)]+1)/N, 1]$, while $\rho u_N^2 = N\mu([0,a))$ in $[([Na]-1)/N, [Na]/N)$; hence, by (2.12), we have that

$$\lim_{N\to+\infty} \int_0^\theta \rho\, u_N^2\, d\theta = \mu^a([0,\theta)), \quad \text{a.e. } \theta \in (a,1)\,.$$

Moreover, as already observed, (u_N) is bounded in L^2; hence,

$$\frac{1}{\rho(\theta)} \left(\int_0^\theta \rho\, u_N^2\, d\varphi - \int_0^\theta \rho\ d\varphi \right)^2 \le c$$

for every $\theta \in (a,1)$ and for every N. Therefore, we can apply Lebesgue's theorem in the second term of (2.10), and by (2.11), (2.3) we have

$$\limsup_{N\to+\infty} G(u_N) \le \int_a^1 \frac{1}{\rho(\theta)} \left(\mu^a([0,\theta)) - \int_0^\theta \rho \, d\varphi \right)^2 d\theta + o(1)$$

$$\le \overline{G}(u) + o(1) \quad \text{as} \quad a \to 0^+ . \tag{2.13}$$

Since $u^a \to u$ in $L^2_{\rm loc}(0,1)$ as $a \to 0^+$, we can conclude that, passing to a further subsequence, $u_N \rightharpoonup u$ in $L^2_{\rm loc}(0,1)$ as $N \to +\infty$, and

$$\limsup_{N\to+\infty} G(u_N) \le \overline{G}(u),$$

as desired. □

Remark 2. Note that the set of minimizers for $\overline{G}(u)$ is $\{|u| \le 1\}$.

Remark 3. The functional $\overline{G}$ can be estimated from above and from below as

$$G^-(u) \le \overline{G}(u) \le G^+(u), \tag{2.14}$$

where

$$G^+(u) = \int_0^1 \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho \, (u^2-1)^+ \, d\varphi \right)^2 d\theta$$

and

$$G^-(u) = \int_0^1 \frac{1}{\rho(\theta)} \left(\left(\int_0^\theta \rho \, (u^2-1) \, d\varphi \right)^+ \right)^2 d\theta \, .$$

The second inequality in (2.14) easily follows by testing the definition of $\overline{G}(u)$ with the measure $\mu = \rho(u^2 \vee 1)\, d\varphi$. To check the first inequality from below, we first note that $G(u) \ge G^-(u)$, and by Fatou's lemma and the weak $L^2_{\rm loc}$-lower semicontinuity of the functional $\int_0^\theta \rho \, u^2 \, d\varphi$ for every $\theta \in (0,1)$, we get that G^- is weakly $L^2_{\rm loc}$-lower semicontinuous. Hence, by Lemma 2.2 and the lower semicontinuity of G^-, we have that for every $u \in L^2_{\rm loc}(0,1)$ there exists a sequence u_N weakly $L^2_{\rm loc}$ converging to u such that

$$\overline{G}(u) = \lim_{N\to+\infty} G(u_N) \ge \liminf_{N\to+\infty} G^-(u_N) \ge G^-(u) \, .$$

We show now an example of function u such that $\overline{G}(u) = G^-(u)$. Let us consider for simplicity $\rho = 1$. Let

$$u(\varphi) = \begin{cases} \sqrt{2}, & \varphi \in (0,1/4), \\ 0, & \varphi \in (1/4,1) \, . \end{cases} \tag{2.15}$$

Note that

$$\int_0^\theta (u^2-1) \, d\varphi < 0 \quad \text{for all } \theta > 1/2 \, ,$$

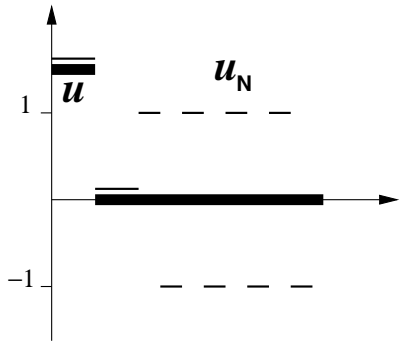


FIG. 1. *Nonlocal oscillations.*

so that

$$G^-(u) = \int_0^{1/4} \theta^2 \, d\theta + \int_{1/4}^{1/2} \left(\frac{1}{2} - \theta\right)^2 d\theta = \int_0^1 (\mu([0,\theta)) - \theta)^2 d\theta,$$

where the measure $\mu = v \, d\varphi$ is defined by

$$v(\varphi) = \begin{cases} 2, & \varphi \in (0, 1/4), \\ 0, & \varphi \in (1/4, 1/2), \\ 1, & (1/2, 1)\,. \end{cases}$$

Since $\mu \geq u^2 \, d\varphi$, we can test the definition of $\overline{G}(u)$ with μ getting $G^-(u) \geq \overline{G}(u)$, and then we have that $G^-(u) = \overline{G}(u)$ by (2.14). A recovery sequence (corresponding to (2.4) with $a = 0$ and $\rho = 1$) for $\overline{G}(u)$ is shown in Figure 1; it highlights the nonlocal nature of the oscillations that start at $\varphi = 1/2$, while the target function is 0 on the whole $(1/4, 1)$.

Note that u as in (2.15) is also an example of a function such that $\overline{G}(u) < G^+(u)$. Finally, we note that also the inequality $G^- \leq \overline{G}$ is sharp: if we consider

$$u(\varphi) = \begin{cases} 0, & \varphi \in (0, 1/2), \\ \sqrt{2}, & \varphi \in (1/2, 1), \end{cases}$$

we have that $G^-(u) = 0$ while $\overline{G}(u) > 0$ by Remark 2.

THEOREM 2.3 (Γ-convergence result). *We have*

$$\Gamma(\text{w-}L^2_{\text{loc}})\text{-}\lim_{\varepsilon \to 0} F^0_\varepsilon(u) = \overline{G}(u)$$

for every $u \in L^2_{\text{loc}}(0,1)$.

Proof. Let $u \in L^2_{\text{loc}}(0,1)$. By the definition of $\overline{G}$ for every sequence (u_ε) weakly L^2_{loc} converging to u, we have

$$F^0_\varepsilon(u_\varepsilon) \geq G(u_\varepsilon) \geq \overline{G}(u_\varepsilon)\,;$$

hence by the weak lower semicontinuity of $\overline{G}$ we get the lim inf inequality.

Conversely, let $u \in L^2_{\text{loc}}(0,1)$, and let

$$u^a(\theta) = \begin{cases} u(\theta) & \text{if } a \leq \theta \leq 1 - a, \\ 0 & \text{otherwise} \end{cases}$$

with $0 < a < 1$; hence, $u^a \in L^2(0,1)$ and $u^a \to u$ in $L^2_{\rm loc}(0,1)$ as $a \to 0^+$. By Lemma 2.2, there exists a sequence $(u_N) \in L^2(\mathbf{R})$ weakly converging to u^a in $L^2(0,1)$ such that

$$\overline{G}(u^a) = \lim_{N\to+\infty} G(u_N)\,.$$

Let $\eta : \mathbf{R} \mapsto [0,+\infty)$ be a mollifier; we define $\eta_\varepsilon(\theta) = \frac{1}{\sqrt{\varepsilon}}\,\eta(\frac{\theta}{\sqrt{\varepsilon}})$, then we have $u_\varepsilon^N = u_N * \eta_\varepsilon \in C_c^\infty(\mathbf{R})$ and $u_\varepsilon^N \to u_N$ in $L^2(0,1)$ as $\varepsilon \to 0$ for every N. Hence,

$$\begin{aligned}\lim_{\varepsilon\to0} F_\varepsilon^0(u_\varepsilon^N) &= \lim_{\varepsilon\to0}\Big(\varepsilon^2 \int_0^1 \rho\,(u_N * \eta_\varepsilon')^2\,d\theta + G(u_\varepsilon^N)\Big)\\ &= \lim_{\varepsilon\to0} G(u_\varepsilon^N) = G(u_N),\end{aligned}$$

and by the lower semicontinuity of the Γ-lim sup and (2.3), we have that

$$\begin{aligned}\Gamma\text{-}\limsup_{\varepsilon\to0} F_\varepsilon^0(u^a) &\le \liminf_{N\to+\infty}\Big(\Gamma\text{-}\limsup_{\varepsilon\to0} F_\varepsilon^0(u_N)\Big)\\ &\le \liminf_{N\to+\infty} G(u_N) = \overline{G}(u^a) \le \overline{G}(u) + o(1)\end{aligned}$$

as $a \to 0^+$. We again use the lower semicontinuity of the Γ-lim sup to get, as $a \to 0^+$, that

$$\Gamma\text{-}\limsup_{\varepsilon\to0} F_\varepsilon^0(u) \le \overline{G}(u),$$

which concludes the proof of the lim sup inequality (see [6, Remark 1.29]). ∎

3. The case $\alpha = 3/2$: Phase transitions. In section 2 we showed that the set of the minimum points of the Γ-limit F^0 is $\{|u| \le 1\}$ and showed that $\min F^0 = 0$. To reduce the choice in the minimizers of the limit problem we may further rescale F_ε^0; the next meaningful scaling is $\alpha = 3/2$. We then consider the family of functionals

$$F_\varepsilon^{3/2}(u) = \sqrt{\varepsilon}\int_0^1 \rho\,(u')^2\,d\theta + \varepsilon^{-3/2}\int_0^1 \frac{1}{\rho(\theta)}\left(\int_0^\theta \rho\,(u^2-1)\,d\varphi\right)^2 d\theta$$

for $u \in H^1(0,1)$.

THEOREM 3.1 (compactness). *Let (u_ε) be a sequence of equibounded energy, i.e., $\sup_\varepsilon F_\varepsilon^{3/2}(u_\varepsilon) < +\infty$; then (u_ε) is equibounded in $L^\infty_{\rm loc}(0,1)$ and, up to subsequences, (u_ε) converges to $u \in BV_{\rm loc}((0,1);\{-1,1\})$ strongly in $L^1_{\rm loc}$.*

Proof. Let $\eta_\varepsilon^\pm \neq \pm1$ such that $-1 < \eta_\varepsilon^+, \eta_\varepsilon^- < 1$ or $\eta_\varepsilon^+, \eta_\varepsilon^- > 1$ or $\eta_\varepsilon^+, \eta_\varepsilon^- < -1$. We denote by $(\delta_\varepsilon^-, \delta_\varepsilon^+)$ an interval such that $u_\varepsilon(\delta_\varepsilon^-) = \eta_\varepsilon^-$, $u_\varepsilon(\delta_\varepsilon^+) = \eta_\varepsilon^+$ and u_ε takes values between η_ε^- and η_ε^+. In what follows, we use $\eta^\pm = \eta_\varepsilon^\pm$, $\delta^\pm = \delta_\varepsilon^\pm$ so that we do not overburden notation. By assumption,

$$\sqrt{\varepsilon}\int_{\delta^-}^{\delta^+} \rho\,(u_\varepsilon')^2\,d\theta + \varepsilon^{-3/2}\int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)}\left(\int_0^\theta \rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta \le c\,. \tag{3.1}$$

For any fixed $0 < a < 1$, we assume $\delta^- \ge a$.

The crucial argument in the proof will be to estimate the length $\delta^+ - \delta^-$ of any such interval. We will show that

$$\sqrt{\varepsilon}\int_{\delta^-}^{\delta^+}\rho\,(u_\varepsilon')^2\,d\theta\;+\varepsilon^{-3/2}\int_{\delta^-}^{\delta^+}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta$$
$$\geq \rho(a)|\eta^+-\eta^-|^2\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+c(\lambda)\left(\frac{\delta}{\sqrt{\varepsilon}}\right)^3,$$

where λ is the minimum between $|\eta_-^2-1|$ and $|\eta_+^2-1|$, and $c(\lambda)$ is a suitable positive constant depending only on λ. From this estimate we will deduce that if $|\eta^+-\eta^-|\geq\zeta>0$, then $\delta^+-\delta^-$ is of the order $\sqrt{\varepsilon}$, and that

$$\sqrt{\varepsilon}\int_{\delta^-}^{\delta^+}\rho\,(u_\varepsilon')^2\,d\theta\;+\varepsilon^{-3/2}\int_{\delta^-}^{\delta^+}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\geq c\zeta^{3/2}.$$

From this, we will finally obtain that frequent large oscillations are forbidden, and in particular that the number of transitions from values close to 1 to values close to -1, and conversely, is equibounded, and then deduce the $L^1_{\rm loc}$-convergence in $(0,1)$.

Step 1. We define the set

$$A_\varepsilon=\left\{\theta\in(a,1]:\quad\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2\leq\varepsilon\right\},$$

and we denote by A_ε^c its complementary set. Since $\sup_\varepsilon F_\varepsilon^{3/2}(u_\varepsilon)<+\infty$, we have that

$$|A_\varepsilon^c|\varepsilon<\int_{A_\varepsilon^c}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2\leq c\varepsilon^{3/2},$$

which implies that there exists $\overline{c}(\varepsilon)\leq c$ such that

$$|A_\varepsilon^c|=\overline{c}(\varepsilon)\,\sqrt{\varepsilon}\,. \tag{3.2}$$

Let $\theta_\varepsilon\in A_\varepsilon$ be such that $\theta_\varepsilon=\max\{\theta\in A_\varepsilon:\,\theta\leq\delta^-\}$; by definition,

$$\varepsilon\geq\frac{1}{\rho(\theta_\varepsilon)}\left(\int_0^{\theta_\varepsilon}\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2\geq\frac{1}{\rho(1)}\left(\int_0^{\theta_\varepsilon}\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2.$$

It follows that there exists $c_1(\varepsilon)$ such that $|c_1(\varepsilon)|\leq\sqrt{\rho(1)}$ and

$$\int_0^{\theta_\varepsilon}\rho\,(u_\varepsilon^2-1)\,d\varphi=c_1(\varepsilon)\,\sqrt{\varepsilon}\,. \tag{3.3}$$

Moreover, for every $\theta\in(\theta_\varepsilon,\delta^-)\subset A_\varepsilon^c$, we have that

$$|u_\varepsilon(\theta)-\eta^-|=|u_\varepsilon(\theta)-u_\varepsilon(\delta^-)|\leq\left(|A_\varepsilon^c|\int_{\theta_\varepsilon}^{\delta^-}(u_\varepsilon')^2\,d\theta\right)^{1/2}$$
$$\leq\left(\frac{|A_\varepsilon^c|}{\rho(a)}\int_{\theta_\varepsilon}^{\delta^-}\rho\,(u_\varepsilon')^2\,d\theta\right)^{1/2}.$$

Since $\sup_\varepsilon F_\varepsilon^{3/2}(u_\varepsilon) < +\infty$, by (3.2) we have that

$$|u_\varepsilon(\theta)| \le |\eta^-| + \tilde{c} \qquad \text{for all } \theta \in [\theta_\varepsilon, \delta^-], \tag{3.4}$$

which in turn implies

$$\left| \int_{\theta_\varepsilon}^{\delta^-} \rho\,(u_\varepsilon^2 - 1)\,d\theta \right| \le c(\eta^-)\,\sqrt{\varepsilon},$$

where $c(\eta^-) = c\,\rho(1)(1+(|\eta^-|+\tilde{c})^2)$. Hence, there exists $c_2(\varepsilon,\eta^-)$ such that $|c_2(\varepsilon,\eta^-)| \le c(\eta^-)$ and

$$\int_{\theta_\varepsilon}^{\delta^-} \rho\,(u_\varepsilon^2 - 1)\,d\theta = c_2(\varepsilon,\eta^-)\,\sqrt{\varepsilon}\,. \tag{3.5}$$

By (3.3) and (3.5) we get that

(3.6)

$$\begin{aligned}
&\int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho\,(u_\varepsilon^2 - 1)\,d\varphi \right)^2 d\theta \\
&\quad = \int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)} \left(\int_0^{\theta_\varepsilon} \rho\,(u_\varepsilon^2 - 1)\,d\varphi + \int_{\theta_\varepsilon}^{\delta^-} \rho\,(u_\varepsilon^2 - 1)\,d\varphi + \int_{\delta^-}^{\theta} \rho\,(u_\varepsilon^2 - 1)\,d\varphi \right)^2 d\theta \\
&\quad = c_5(c_1(\varepsilon) + c_2(\varepsilon,\eta^-))^2\,\varepsilon\,\delta + \int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)} \left(\int_{\delta^-}^{\theta} \rho\,(u_\varepsilon^2 - 1)\,d\varphi \right)^2 d\theta \\
&\qquad + 2(c_1(\varepsilon) + c_2(\varepsilon,\eta^-))\,\sqrt{\varepsilon} \int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)} \left(\int_{\delta^-}^{\theta} \rho\,(u_\varepsilon^2 - 1)\,d\varphi \right) d\theta\,,
\end{aligned}$$

where $c_5 = \int_{\delta^-}^{\delta^+} (1/\rho)\,d\theta$ and $\delta = \delta^+ - \delta^-$.

By assumption, u_ε takes values between η^- and η^+ for every $\theta \in [\delta^-, \delta^+]$ (note that either $(u_\varepsilon^2 - 1) > 0$ or $(u_\varepsilon^2 - 1) < 0$ for every $\theta \in [\delta^-, \delta^+]$); hence,

$$|u_\varepsilon^2 - 1| \ge \lambda := ||\eta^-|^2 - 1| \wedge ||\eta^+|^2 - 1|,$$

where we use the notation λ without an explicit dependence on $\eta^\pm$ since we want to emphasize that λ does not depend on the values of $\eta^\pm$ but on the minimum distance of $|\eta^\pm|^2$ from 1.

Moreover, on $[\delta^-, \delta^+]$,

$$|u_\varepsilon^2 - 1| \le c(\eta^\pm) := ||\eta^-|^2 - 1| \vee ||\eta^+|^2 - 1|;$$

then there exist $c_3(\varepsilon,\eta^\pm)$ and $c_4(\varepsilon,\lambda)$ such that

$$0 < |c_3(\varepsilon,\eta^\pm)| \le \frac{c(\eta^\pm)}{2}\,, \qquad c_4(\varepsilon,\lambda) \ge \frac{\lambda^2}{3}\,\frac{\rho(a)^2}{\rho(1)} \tag{3.7}$$

and

$$\int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)} \left(\int_{\delta^-}^{\theta} \rho\,(u_\varepsilon^2 - 1)\,d\varphi \right) d\theta = c_3(\varepsilon,\eta^\pm)\,\delta^2\,, \tag{3.8}$$

$$\int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)} \left(\int_{\delta^-}^{\theta} \rho\,(u_\varepsilon^2 - 1)\,d\varphi \right)^2 d\theta = c_4(\varepsilon,\lambda)\,\delta^3\,. \tag{3.9}$$

By (3.6), (3.8), and (3.9) we get that

$$\begin{aligned}
&\int_{\delta^-}^{\delta^+} \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho \,(u_\varepsilon^2 - 1)\, d\varphi \right)^2 d\theta \\
&\qquad = \delta^3 \Bigg\{ c_4(\varepsilon,\lambda) + c_5 (c_1(\varepsilon) + c_2(\varepsilon,\eta^-))^2 \left(\frac{\sqrt{\varepsilon}}{\delta} \right)^2 \\
&\qquad\qquad + \; 2c_3(\varepsilon,\eta^\pm)(c_1(\varepsilon) + c_2(\varepsilon,\eta^-)) \left(\frac{\sqrt{\varepsilon}}{\delta} \right) \Bigg\}. \qquad (3.10)
\end{aligned}$$

Note that c_4 and c_5 are always strictly positive while $c_3 \neq 0$ with $\operatorname{sign}(c_3) = \operatorname{sign}(u_\varepsilon^2 - 1)$. Moreover, we claim that

$$c_4 - (c_3^2/c_5) \geq c(\lambda), \qquad (3.11)$$

where $c(\lambda) := \lambda^2\, c(\delta^\pm)\rho(a)^2/\rho(1)$ with $1/12 \leq c(\delta^\pm) \leq 1/3$ for every $\delta^\pm$. Also in this case we prefer to use the notation $c(\lambda)$, omitting the dependence on $\delta^\pm$ because of the bound on $c(\delta^\pm)$. We check now (3.11). Let us denote

$$f(\theta) = \int_{\delta^-}^\theta \rho \,|u_\varepsilon^2 - 1|\, d\varphi\,;$$

hence, $f'(\theta) = \rho(\theta)|u_\varepsilon^2(\theta) - 1| \geq \lambda\rho(a)$. Then, there exists $\delta_0 \in (\delta^-, \delta^+)$ such that

$$|c_3| = \frac{1}{\delta^2} \int_{\delta^-}^{\delta^+} \frac{f(\theta)}{\rho(\theta)}\, d\theta = \frac{f(\delta_0)}{\delta^2} \int_{\delta^-}^{\delta^+} \frac{d\theta}{\rho(\theta)} = \frac{f(\delta_0)}{\delta} c_5 \qquad (3.12)$$

and

$$|f(\theta) - f(\delta_0)| \geq \lambda\rho(a)\,|\theta - \delta_0|\,.$$

It follows that

$$\begin{aligned}
\int_{\delta^-}^{\delta^+} \frac{(f(\theta) - f(\delta_0))^2}{\rho(\theta)}\, d\theta &\geq (\lambda\rho(a))^2 \int_{\delta^-}^{\delta^+} \frac{(\theta - \delta_0)^2}{\rho(\theta)}\, d\theta \\
&\geq (\lambda\rho(a))^2 \, \frac{(\delta^+ - \delta_0)^3 - (\delta^- - \delta_0)^3}{3\rho(1)} \\
&= \delta^3 \left(\lambda^2\, c(\delta^\pm)\, \frac{\rho(a)^2}{\rho(1)} \right), \qquad (3.13)
\end{aligned}$$

where $1/12 \leq c(\delta^\pm) \leq 1/3$ for every $\delta^\pm$. On the other hand, by (3.12) we have that

$$\begin{aligned}
&\int_{\delta^-}^{\delta^+} \frac{(f(\theta) - f(\delta_0))^2}{\rho(\theta)}\, d\theta \\
&\quad = \int_{\delta^-}^{\delta^+} \frac{f(\theta)^2}{\rho(\theta)}\, d\theta + f(\delta_0)^2 \int_{\delta^-}^{\delta^+} \frac{d\theta}{\rho(\theta)} - 2f(\delta_0) \int_{\delta^-}^{\delta^+} \frac{f(\theta)}{\rho(\theta)}\, d\theta \\
&\quad = \delta^3 \left(c_4 - c_5 \left(\frac{f(\delta_0)}{\delta} \right)^2 \right) \\
&\quad = \delta^3 \left(c_4 - \frac{c_3^2}{c_5} \right); \qquad (3.14)
\end{aligned}$$

hence, by (3.13) we get (3.11).

We now estimate the term with the derivative in (3.1); by Hölder's inequality we get that

$$\sqrt{\varepsilon}\int_{\delta^-}^{\delta^+}\rho\,(u_\varepsilon')^2\,d\theta \geq \rho(a)\,|\eta^+-\eta^-|^2\left(\frac{\sqrt{\varepsilon}}{\delta}\right). \tag{3.15}$$

By (3.10) and (3.15) we have then

$$\begin{aligned}&\sqrt{\varepsilon}\int_{\delta^-}^{\delta^+}\rho\,(u_\varepsilon')^2\,d\theta\;+\varepsilon^{-3/2}\int_{\delta^-}^{\delta^+}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\\ &\geq \rho(a)\,|\eta^+-\eta^-|^2\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+\left(\frac{\delta}{\sqrt{\varepsilon}}\right)^3\left(\gamma\left(\frac{\sqrt{\varepsilon}}{\delta}\right)^2\pm\beta\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+\alpha\right),\end{aligned} \tag{3.16}$$

where

$$\alpha=c_4\,,\quad \gamma=c_5(c_1+c_2)^2\,,\quad \beta=2|c_3(c_1+c_2)|\,.$$

Note that, as already observed, $\alpha>0$, $\gamma\geq 0$, and $c_3(c_1+c_2)$ may be ≥ 0 or ≤ 0 ($\gamma=0$ if and only if $\beta=0$). By (3.11), if we minimize $\gamma(\sqrt{\varepsilon}/\delta)^2\pm\beta(\sqrt{\varepsilon}/\delta)+\alpha$ in $(\sqrt{\varepsilon}/\delta)$, we have

$$\begin{aligned}&\rho(a)\,|\eta^+-\eta^-|^2\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+\left(\gamma\left(\frac{\sqrt{\varepsilon}}{\delta}\right)^2\pm\beta\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+\alpha\right)\left(\frac{\delta}{\sqrt{\varepsilon}}\right)^3\\ &\geq \rho(a)|\eta^+-\eta^-|^2\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+\left(\alpha-\frac{\beta^2}{4\gamma}\right)\left(\frac{\delta}{\sqrt{\varepsilon}}\right)^3\\ &= \rho(a)|\eta^+-\eta^-|^2\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+\left(c_4-\frac{c_3^2}{c_5}\right)\left(\frac{\delta}{\sqrt{\varepsilon}}\right)^3\\ &\geq \rho(a)|\eta^+-\eta^-|^2\left(\frac{\sqrt{\varepsilon}}{\delta}\right)+c(\lambda)\left(\frac{\delta}{\sqrt{\varepsilon}}\right)^3.\end{aligned} \tag{3.17}$$

Step 2. If

$$|\eta^+-\eta^-|\geq\zeta>0\,,$$

with ζ independent of ε, studying the function $x\mapsto\rho(a)|\eta^+-\eta^-|^2x+c(\lambda)/x^3$ for $x>0$, by (3.1), (3.16), and (3.17) we have that $\delta/\sqrt{\varepsilon}$ is bounded; i.e., there exist two positive constants α_1,α_2 such that $\alpha_1\sqrt{\varepsilon}\leq\delta\leq\alpha_2\sqrt{\varepsilon}$.

Step 3. The minimum point of $x\mapsto\rho(a)|\eta^+-\eta^-|^2x+c(\lambda)/x^3$ for $x>0$ is $x_m=c_m/|\eta^+-\eta^-|^{1/2}$; then by (3.16) and (3.17) we have that

$$\begin{aligned}&\sqrt{\varepsilon}\int_{\delta^-}^{\delta^+}\rho\,(u_\varepsilon')^2\,d\theta\;+\varepsilon^{-3/2}\int_{\delta^-}^{\delta^+}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\\ &\geq\tilde{c}(\lambda)\,|\eta^+-\eta^-|^{3/2}\geq\tilde{c}(\lambda)\,\zeta^{3/2}\,,\end{aligned} \tag{3.18}$$

where $\tilde{c}(\lambda)=(\rho(a)c_m+c(\lambda)/c_m^3)$. We recall that λ is the minimum distance of $|\eta^\pm|^2$ from 1. Since (u_ε) is a sequence with bounded energy, and estimate (3.18) depends on λ only, we deduce that the number of transitions of u_ε from η^- to η^+ is equibounded independently of ε.

Conclusions. By (3.1) and (3.18) we conclude that (u_ε) is equibounded in $L^\infty_{\rm loc}(0,1)$. Moreover, by Steps 2 and 3, we have that for every fixed $0<a<1$, (u_ε) converges in measure in $(a,1)$ to $u\in BV_{\rm loc}((0,1);\{-1,1\})$; hence, up to subsequences, $u_\varepsilon\to u$ a.e. $\theta\in(a,1)$. Since (u_ε) is equibounded in $L^\infty_{\rm loc}(0,1)$, we can conclude that, up to subsequences, (u_ε) converges strongly in $L^1_{\rm loc}(0,1)$ to $u\in BV_{\rm loc}((0,1);\{-1,1\})$. ∎

The following proposition gives an estimate of the measure of the set, where a sequence of bounded energy is not close to ±1.

PROPOSITION 3.2. *Let (u_ε) be a sequence converging to u in $L^1_{\rm loc}(0,1)$ such that $\sup_\varepsilon F_\varepsilon^{3/2}(u_\varepsilon)<+\infty$. Then for every fixed $0<a<1$ and $\eta>0$, there exists $\overline{c}=\overline{c}(\eta)$ such that*

$$|\{\theta\in[a,1-a]:\ |u_\varepsilon^2(\theta)-1|>\eta\}|\le\overline{c}\,\sqrt{\varepsilon}\,. \tag{3.19}$$

Proof. Let $0<a<1$ and $\eta>0$ be fixed. We define $\delta_i^\pm$ such that

$$\{\theta\in[a,1-a]:\ |u_\varepsilon^2(\theta)-1|>\eta\}\subseteq\bigcup_i[\delta_i^-,\delta_i^+]\,,$$

with $\delta_i^-<\delta_i^+$, $\delta_i^+\le\delta_j^-$ for every $i<j$ and

$$\begin{cases}|(u_\varepsilon(\delta_i^\pm))^2-1|=\eta/2 & \text{except if } \delta_i^+=1,\\ |u_\varepsilon^2(\theta)-1|>\eta/2 & \text{if } \theta\in(\delta_i^-,\delta_i^+),\\ \exists\delta_i\in(\delta_i^-,\delta_i^+) & \text{such that } |(u_\varepsilon(\delta_i))^2-1|=\eta\,.\end{cases}$$

Hence, we may have two cases:

$$u_\varepsilon(\delta_i^-)\ne u_\varepsilon(\delta_i^+),\quad\text{and hence}\quad u_\varepsilon(\delta_i^\pm)\in\left\{\pm\sqrt{1-\eta/2}\right\}, \tag{3.20}$$

or

$$u_\varepsilon(\delta_i^-)=u_\varepsilon(\delta_i^+)\in\left\{\pm\sqrt{1-\eta/2}\,,\,\pm\sqrt{1+\eta/2}\right\}. \tag{3.21}$$

In case (3.20) we may apply Steps 1–3 from the proof of Theorem 3.1 with $\eta^\pm\in\{\pm\sqrt{1-\eta/2}\}$, $\lambda=\eta/2$, and $\zeta=2\sqrt{1-\eta/2}$. If instead we are in case (3.21), we consider

$$\eta^-\in\left\{\pm\sqrt{1-\eta/2}\,,\,\pm\sqrt{1+\eta/2}\right\},\quad \eta^+\in\left\{\max_{[\delta_i^-,\delta_i^+]}u_\varepsilon\,,\ \min_{[\delta_i^-,\delta_i^+]}u_\varepsilon\right\},$$

or the converse. For example, if $u_\varepsilon(\delta_i^-)=u_\varepsilon(\delta_i^+)=\sqrt{1+\eta/2}$, then we apply Steps 1–3 two times: the first to

$$\eta^-=\sqrt{1+\eta/2}\,,\quad \eta^+=\max_{[\delta_i^-,\delta_i^+]}u_\varepsilon,$$

and then to

$$\eta^-=\max_{[\delta_i^-,\delta_i^+]}u_\varepsilon\,,\quad \eta^+=\sqrt{1+\eta/2}\,.$$

Hence, also in this case we may apply Steps 1–3 with $\zeta>0$.

We can then conclude that the number of intervals $[\delta_i^-, \delta_i^+]$ is finite (and independent of ε) and that there exist $\alpha_i^1, \alpha_i^2 > 0$ such that $\alpha_i^1\sqrt{\varepsilon} \le (\delta_i^+ - \delta_i^-) \le \alpha_i^2\sqrt{\varepsilon}$, which proves (3.19). □

THEOREM 3.3 (Γ-convergence result). *We have*

$$\Gamma(L^1_{\rm loc})\text{-}\lim_{\varepsilon\to 0} F_\varepsilon^{3/2}(u) = c_0 \sum_{\theta\in S(u)} \rho(\theta)$$

for every $u \in BV_{\rm loc}((0,1);\{-1,1\})$, *where*

$$c_0 = \inf_{T>0} \inf \left\{ \int_{-T}^{T} (v')^2\, ds + \int_{-T}^{T} \left(\int_{-T}^{\sigma} (v^2-1)\, ds \right)^2 d\sigma : \right.$$
$$\left. v \in H^1(-T,T), \quad v(\pm T) = \pm 1, \quad \int_{-T}^{T} (v^2-1)\, ds = 0 \right\}.$$

Proof (lim inf inequality). Let $u \in BV_{\rm loc}((0,1);\{-1,1\})$ and let (u_ε) be a sequence converging to u in $L^1_{\rm loc}(0,1)$ such that $\sup_\varepsilon F_\varepsilon^{3/2}(u_\varepsilon) < +\infty$.

Step 1. We fix $0 < a < 1$ and consider $\theta_i \in S(u)$ such that $a < \theta_i < 1-a$. Without loss of generality, we may assume that $u(\theta_i\pm) = \mp 1$. Let $\theta_\varepsilon \to \theta_i$, as $\varepsilon \to 0$, and let $M > 0$ be such that

$$u_\varepsilon > \frac{1}{2} \quad \text{on } I_\varepsilon^- := (\theta_\varepsilon - 2M\sqrt{\varepsilon}, \theta_\varepsilon - M\sqrt{\varepsilon})$$

and

$$u_\varepsilon < \frac{1}{2} \quad \text{on } I_\varepsilon^+ := (\theta_\varepsilon + M\sqrt{\varepsilon}, \theta_\varepsilon + 2M\sqrt{\varepsilon}).$$

We claim that for every fixed $\eta > 0$, there exist a constant $c > 0$ and $\theta_\varepsilon^\pm \in I_\varepsilon^\pm$ of the form $\theta_\varepsilon^\pm = \theta_\varepsilon \pm M_\varepsilon\sqrt{\varepsilon}$, with $M \le M_\varepsilon \le 2M$, such that

$$\left| \int_0^{\theta_\varepsilon^\pm} \rho\,(u_\varepsilon^2 - 1)\, d\varphi \right| \le \frac{c}{\sqrt{M}}\sqrt{\varepsilon} \tag{3.22}$$

and

$$|u_\varepsilon^2(\theta_\varepsilon^\pm) - 1| \le \eta\,. \tag{3.23}$$

In fact, reasoning by contradiction, assume that for every constant $c > 0$ we cannot find two points $\theta_\varepsilon^\pm \in I_\varepsilon^\pm$ such that (3.22) and (3.23) are satisfied at the same time. If we denote

$$B_c = \left\{ \theta : M\sqrt{\varepsilon} \le |\theta - \theta_\varepsilon| \le 2M\sqrt{\varepsilon},\ \left| \int_0^{\theta} \rho\,(u_\varepsilon^2 - 1)\, d\varphi \right| > \frac{c}{\sqrt{M}}\sqrt{\varepsilon} \right\},$$

then by (3.19) we have $|B_c| \ge \sqrt{\varepsilon}(2M - \overline{c})$, and hence

$$F_\varepsilon^{3/2}(u_\varepsilon) \ge \varepsilon^{-3/2} \int_{B_c} \frac{1}{\rho(\theta)} \left(\int_0^{\theta} \rho\,(u_\varepsilon^2 - 1)\, d\varphi \right)^2 d\theta$$
$$\ge \varepsilon^{-1/2} \frac{c^2}{\rho(1)} \frac{|B_c|}{M} \ge \frac{c^2}{\rho(1)} \left(2 - \frac{\overline{c}}{M} \right).$$

Note that we can choose M large enough such that $\overline{c} < 2M$. Since $\sup_\varepsilon F_\varepsilon^{3/2}(u_\varepsilon) < +\infty$, we get a contradiction by the arbitrariness of $c > 0$.

Step 2. We give an estimate on the contribution between θ_ε^- and θ_ε^+. By (3.22), there exists a sequence (c_-^ε), bounded independently of ε, such that

(3.24)

$$\begin{aligned}
&\sqrt{\varepsilon}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\rho\,(u_\varepsilon')^2\,d\theta+\varepsilon^{-3/2}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\\
&\quad=\sqrt{\varepsilon}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\rho\,(u_\varepsilon')^2\,d\theta+\varepsilon^{-3/2}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\frac{1}{\rho(\theta)}\left(\frac{c_-^\varepsilon}{\sqrt{M}}\sqrt{\varepsilon}+\int_{\theta_\varepsilon^-}^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\\
&\quad=\int_{-M_\varepsilon}^{M_\varepsilon}\rho_\varepsilon\,(v')^2\,ds+\int_{-M_\varepsilon}^{M_\varepsilon}\frac{1}{\rho_\varepsilon(\sigma)}\left(\frac{c_-^\varepsilon}{\sqrt{M}}+\int_{-M_\varepsilon}^\sigma\rho_\varepsilon\,(v^2-1)\,ds\right)^2 d\sigma,
\end{aligned}$$

where $s=(\varphi-\theta_\varepsilon)/\sqrt{\varepsilon}$, $\sigma=(\theta-\theta_\varepsilon)/\sqrt{\varepsilon}$, $\rho_\varepsilon(t)=\rho(\theta_\varepsilon+t\sqrt{\varepsilon})$ and $v(s)=u_\varepsilon(\theta_\varepsilon+\sqrt{\varepsilon}s)$. By (3.22) there exists also a sequence (c_+^ε), bounded independently of ε, such that

$$\int_0^{\theta_\varepsilon^+}\rho\,(u_\varepsilon^2-1)\,d\varphi=\frac{c_+^\varepsilon}{\sqrt{M}}\sqrt{\varepsilon}\,;$$

hence,

$$\frac{1}{\sqrt{\varepsilon}}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\rho\,(u_\varepsilon-1)^2\,d\varphi=\int_{-M_\varepsilon}^{M_\varepsilon}\rho_\varepsilon\,(v^2-1)\,ds=\frac{c_+^\varepsilon-c_-^\varepsilon}{\sqrt{M}}\,. \tag{3.25}$$

By (3.23) we get that $|v^2(\pm M_\varepsilon)-1|\le\eta$. We consider $v(s)=u_\varepsilon(\theta_\varepsilon+\sqrt{\varepsilon}s)$ for $s\in[-M_\varepsilon,M_\varepsilon]$ and we extend it to $[-3M,3M]$ in such a way that

$$v(\pm 3M)=\mp 1,\quad \int_{-3M}^{3M}\rho_\varepsilon\,(v^2-1)\,ds=0, \tag{3.26}$$

$$\int_{-3M}^{-M_\varepsilon}\rho_\varepsilon\,(v^2-1)\,ds=\frac{c_-^\varepsilon}{\sqrt{M}}\,.$$

Note that by (3.25) and (3.26) we have that

$$\int_{M_\varepsilon}^{3M}\rho_\varepsilon\,(v^2-1)\,ds=-\frac{c_+^\varepsilon}{\sqrt{M}}\,.$$

We explicitly construct v on $[-3M,-M_\varepsilon]$, the construction on $[M_\varepsilon,3M]$ being analogous. We also suppose that $c_-^\varepsilon\le 0$ and $v(-M_\varepsilon)\ge 1$ (the construction being the same or simpler in the other cases). We define v on $[-3M,-M_\varepsilon]$ as

$$v(s)=\begin{cases}1 & \text{if } s\in[-3M,-h_\varepsilon),\\ -s+\sqrt{1-\eta}-k_\varepsilon^- & \text{if } s\in[-h_\varepsilon,-k_\varepsilon^-),\\ \sqrt{1-\eta} & \text{if } s\in[-k_\varepsilon^-,-k_\varepsilon^+),\\ s+\sqrt{1-\eta}+k_\varepsilon^+ & \text{if } s\in[-k_\varepsilon^+,-M_\varepsilon],\end{cases} \tag{3.27}$$

where

$$-k_\varepsilon^- = -h_\varepsilon + 1 - \sqrt{1-\eta}, \qquad -k_\varepsilon^+ = -M_\varepsilon - v(-M_\varepsilon) + \sqrt{1-\eta}\,.$$

Note that k_ε^+ is fixed by Step 1 while k_ε^- is fixed by (3.26). In fact, since $|v^2-1| \le \eta$ on $[-h_\varepsilon, -k_\varepsilon^-)$ and $[-k_\varepsilon^+, -M_\varepsilon)$, we have that

$$\left| \int_{-h_\varepsilon}^{-k_\varepsilon^-} \rho_\varepsilon\,(v^2-1)\,ds + \int_{-k_\varepsilon^+}^{-M_\varepsilon} \rho_\varepsilon\,(v^2-1)\,ds \right| \le \eta\,\rho(1)\left(1+\sqrt{1+\eta}-2\sqrt{1-\eta}\right);$$

hence, there exists $r(\eta,\varepsilon)$ such that $|r(\eta,\varepsilon)| \le \rho(1)(1+\sqrt{1+\eta}-2\sqrt{1-\eta})$ uniformly in ε and

$$\int_{-3M}^{-M_\varepsilon} \rho_\varepsilon\,(v^2-1)\,ds = \eta\,(-k_\varepsilon + r(\eta,\varepsilon)) = \frac{c_-^\varepsilon}{\sqrt{M}}\,,$$

where $k_\varepsilon := -k_\varepsilon^+ + k_\varepsilon^-$. It follows that

$$k_\varepsilon = \frac{|c_-^\varepsilon|}{\eta\sqrt{M}} + r(\eta,\varepsilon)\,, \tag{3.28}$$

with $\lim_{\eta\to 0^+} |r(\eta,\varepsilon)| = 0$. Reasoning as above, by (3.28), we can also observe that

$$\begin{aligned}
&\int_{-3M}^{-M_\varepsilon} \frac{1}{\rho_\varepsilon(\sigma)} \left(\int_{-3M}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds \right)^2 d\sigma \\
&\quad = \int_{-h_\varepsilon}^{-M_\varepsilon} \frac{1}{\rho_\varepsilon(\sigma)} \left(\int_{-h_\varepsilon}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds \right)^2 d\sigma \\
&\quad \le \eta^2\,\frac{\rho(1)}{3}\Big(k_\varepsilon + 1 + \sqrt{1+\eta} - 2\sqrt{1-\eta}\Big)^3 \\
&\quad = \eta^2\,\frac{\rho(1)}{3}\left(\frac{|c_-^\varepsilon|}{\eta\sqrt{M}} + r(\eta,\varepsilon) + 1 + \sqrt{1+\eta} - 2\sqrt{1-\eta}\right)^3 .
\end{aligned}$$

We can prove a similar estimate also for the contribution corresponding to the interval $[M_\varepsilon, 3M]$; hence, we can conclude that there exists $R(M,\eta,\varepsilon) > 0$, bounded independently of ε, such that

$$\begin{aligned}
R(M,\eta,\varepsilon) = &\int_{-3M}^{-M_\varepsilon} \frac{1}{\rho_\varepsilon(\sigma)} \left(\int_{-3M}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds \right)^2 d\sigma \\
&+ \int_{M_\varepsilon}^{3M} \frac{1}{\rho_\varepsilon(\sigma)} \left(\frac{c_+^\varepsilon}{\sqrt{M}} + \int_{M_\varepsilon}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds \right)^2 d\sigma
\end{aligned} \tag{3.29}$$

and

$$\limsup_{\eta\to 0^+} \limsup_{M\to\infty} \limsup_{\varepsilon\to 0} R(M,\eta,\varepsilon) = 0\,. \tag{3.30}$$

By (3.29) we have that

$$\begin{aligned}
&\int_{-M_\varepsilon}^{M_\varepsilon} \frac{1}{\rho_\varepsilon(\sigma)} \left(\frac{c_-^\varepsilon}{\sqrt{M}} + \int_{-M_\varepsilon}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds \right)^2 d\sigma \\
&\quad = \int_{-3M}^{3M} \frac{1}{\rho_\varepsilon(\sigma)} \left(\int_{-3M}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds \right)^2 d\sigma - R(M,\eta,\varepsilon)\,.
\end{aligned} \tag{3.31}$$

The function v, constructed as in (3.27), gives the following contribution of the term with the derivative:

$$\int_{-3M}^{-M_\varepsilon} \rho_\varepsilon\,(v')^2\,ds \le \rho(1)\left(1+\sqrt{1+\eta}-2\sqrt{1-\eta}\right);$$

hence, reasoning similarly on $[M_\varepsilon, 3M]$, there exists $R_1(\eta,\varepsilon) > 0$, bounded independently of ε, such that

$$\int_{-M_\varepsilon}^{M_\varepsilon} \rho_\varepsilon\,(v')^2\,ds = \int_{-3M}^{3M} \rho_\varepsilon\,(v')^2\,ds - R_1(\eta,\varepsilon) \tag{3.32}$$

and

$$\limsup_{\eta\to 0^+}\limsup_{\varepsilon\to 0} R_1(\eta,\varepsilon) = 0\,. \tag{3.33}$$

By (3.24), (3.32), and (3.31) we get that

$$\begin{aligned}
&\sqrt{\varepsilon}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+} \rho(\theta)(u_\varepsilon')^2\,d\theta + \varepsilon^{-3/2}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+} \frac{1}{\rho(\theta)}\left(\int_0^\theta \rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\\
&\quad= \int_{-3M}^{3M} \rho_\varepsilon\,(v')^2\,ds + \int_{-3M}^{3M} \frac{1}{\rho_\varepsilon(\sigma)}\left(\int_{-3M}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds\right)^2 d\sigma\\
&\qquad - R_1(\eta,\varepsilon) - R(M,\eta,\varepsilon)\\
&\quad\ge \inf\left\{\int_{-T}^{T} \rho_\varepsilon\,(v')^2\,ds + \int_{-T}^{T} \frac{1}{\rho_\varepsilon(\sigma)}\left(\int_{-T}^{\sigma} \rho_\varepsilon\,(v^2-1)\,ds\right)^2 d\sigma\ :\right.\\
&\qquad\qquad \left. v\in H^1(-T,T),\quad v(\pm T)=\pm 1,\quad \int_{-T}^{T} \rho_\varepsilon\,(v^2-1)\,ds = 0\right\}\\
&\qquad + \delta(T,\eta,\varepsilon),
\end{aligned} \tag{3.34}$$

where $T = 3M$ and $\delta(T,\eta,\varepsilon) = -R_1(\eta,\varepsilon) - \tilde{R}(T,\eta,\varepsilon)$ with $\tilde{R}(T,\eta,\varepsilon) = R(M,\eta,\varepsilon)$. Note that in the last infimum problem we can take the boundary values indifferently as $v(\pm T) = \pm 1$ or $v(\pm T) = \mp 1$, by the symmetry of the problem, so that both types of transitions are taken into account.

By (3.30) and (3.33), we have that

$$\limsup_{\eta\to 0^+}\limsup_{T\to\infty}\limsup_{\varepsilon\to 0}|\delta(T,\eta,\varepsilon)|=0\,;$$

hence, it remains to study the behavior of the minimum problems as ε tends to 0. By the uniform convergence of ρ_ε to $\rho(\theta_i)$, as ε tends to 0, we have immediately the Γ-convergence of the functionals with respect to the strong L^2 convergence; i.e.,

$$\begin{aligned}\Gamma(L^2)\text{-}\lim_{\varepsilon\to 0}&\left(\int_{-T}^{T}\rho_\varepsilon\,(v')^2\,ds\,+\,\int_{-T}^{T}\frac{1}{\rho_\varepsilon(\sigma)}\left(\int_{-T}^{\sigma}\rho_\varepsilon\,(v^2-1)\,ds\right)^2\,d\sigma\right)\\&=\rho(\theta_i)\left(\int_{-T}^{T}(v')^2\,ds\,+\,\int_{-T}^{T}\left(\int_{-T}^{\sigma}(v^2-1)\,ds\right)^2\,d\sigma\right).\end{aligned}$$

This Γ-convergence result is stable by adding the constraint; indeed, since the constraint $\int_{-T}^{T}\rho_\varepsilon\,(v^2-1)\,ds=0$ is close for the strong L^2-convergence, the lim inf inequality is trivial. To check the lim sup inequality, let $v\in H^1(-T,T)$ such that $v(\pm T)=\pm 1$ and $\int_{-T}^{T}(v^2-1)\,ds=0$. To obtain a recovery sequence we consider

$$v_\varepsilon(s)=v(s)+t_\varepsilon\phi(s),\qquad s\in[-T,T],\tag{3.35}$$

where

$$\phi(s)=\begin{cases}(s+T)/T, & s\in[-T,0),\\[2mm](T-s)/T, & s\in[0,T]\end{cases}\tag{3.36}$$

(note that $v_\varepsilon\in H^1(-T,T)$ and $v_\varepsilon(\pm T)=\pm 1$) and $t_\varepsilon\in\mathbf{R}$ is chosen such that (v_ε) satisfies the constraint and converges to v in $L^2(-T,T)$; i.e.,

$$\int_{-T}^{T}\rho_\varepsilon\,((v+t_\varepsilon\phi)^2-1)\,ds=0\quad\text{and}\quad t_\varepsilon\to 0\quad\text{as}\quad\varepsilon\to 0\,.\tag{3.37}$$

More precisely, (v_ε) satisfies the constraint in (3.37) if $t_\varepsilon\in\mathbf{R}$ is a solution of the second order equation

$$\left(\int_{-T}^{T}\rho_\varepsilon\,\phi^2\,ds\right)t_\varepsilon^2+2\left(\int_{-T}^{T}\rho_\varepsilon\,v\phi\,ds\right)t_\varepsilon+\left(\int_{-T}^{T}\rho_\varepsilon\,(v^2-1)\,ds\right)=0\,.\tag{3.38}$$

Since $\lim_{\varepsilon\to 0}\int_{-T}^{T}\rho_\varepsilon(v^2-1)\,ds=0$, for ε small enough, (3.38) has two solutions, real and distinct, such that one of the two tends to 0 as ε tends to 0.

To conclude the proof of the lim sup inequality, we have to note that

$$\begin{aligned}\lim_{\varepsilon\to 0}&\left(\int_{-T}^{T}\rho_\varepsilon\,(v_\varepsilon')^2\,ds\,+\,\int_{-T}^{T}\frac{1}{\rho_\varepsilon(\sigma)}\left(\int_{-T}^{\sigma}\rho_\varepsilon\,(v_\varepsilon^2-1)\,ds\right)^2\,d\sigma\right)\\&=\rho(\theta_i)\left(\int_{-T}^{T}(v')^2\,ds\,+\,\int_{-T}^{T}\left(\int_{-T}^{\sigma}(v^2-1)\,ds\right)^2\,d\sigma\right).\end{aligned}\tag{3.39}$$

We go back now to (3.34); by the property of convergence of minima (see [6, Thm. 1.21]) we have that

$$\begin{aligned}
&\liminf_{\varepsilon\to 0}\left(\sqrt{\varepsilon}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\rho\,(u_\varepsilon')^2\,d\theta+\varepsilon^{-3/2}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\right)\\
&\quad\geq\rho(\theta_i)\inf\left\{\int_{-T}^{T}(v')^2\,ds+\int_{-T}^{T}\left(\int_{-T}^{\sigma}(v^2-1)\,ds\right)^2 d\sigma\,:\,v\in H^1(-T,T),\right.\\
&\qquad\qquad\left.v(\pm T)=\pm1,\quad\int_{-T}^{T}(v^2-1)\,ds=0\right\}\\
&\qquad-\limsup_{\varepsilon\to 0}|\delta(T,\eta,\varepsilon)|\\
&\quad\geq\rho(\theta_i)\inf_{T>0}\inf\left\{\int_{-T}^{T}(v')^2\,ds+\int_{-T}^{T}\left(\int_{-T}^{\sigma}(v^2-1)\,ds\right)^2 d\sigma\,:\right.\\
&\qquad\qquad\left.v\in H^1(-T,T),\quad v(\pm T)=\pm1,\quad\int_{-T}^{T}(v^2-1)\,ds=0\right\}\\
&\qquad-\limsup_{T\to+\infty}\limsup_{\varepsilon\to 0}|\delta(T,\eta,\varepsilon)|\\
&\quad=\rho(\theta_i)\,c_0-\limsup_{T\to+\infty}\limsup_{\varepsilon\to 0}|\delta(T,\eta,\varepsilon)|\,,
\end{aligned}$$

where

$$\begin{aligned}
c_0:=\inf_{T>0}\inf\left\{\int_{-T}^{T}(v')^2\,ds+\int_{-T}^{T}\left(\int_{-T}^{\sigma}(v^2-1)\,ds\right)^2 d\sigma\,:\,v\in H^1(-T,T),\right.\\
\left.v(\pm T)=\pm1,\quad\text{and}\quad\int_{-T}^{T}(v^2-1)\,ds=0\right\}.
\end{aligned}$$

Passing to the limit as $\eta\to 0^+$, we get

$$\begin{aligned}
&\liminf_{\varepsilon\to 0}\left(\sqrt{\varepsilon}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\rho\,(u_\varepsilon')^2\,d\theta+\varepsilon^{-3/2}\int_{\theta_\varepsilon^-}^{\theta_\varepsilon^+}\frac{1}{\rho(\theta)}\left(\int_0^\theta\rho\,(u_\varepsilon^2-1)\,d\varphi\right)^2 d\theta\right)\\
&\qquad\geq\rho(\theta_i)\,c_0\,.
\end{aligned}$$

Step 3. If we repeat Steps 1–2 for every $\theta\in S(u)\cap(a,1-a)$, we immediately get that

$$\liminf_{\varepsilon\to 0}F_\varepsilon^{3/2}(u_\varepsilon)\geq c_0\sum_{\theta\in S(u)\cap(a,1-a)}\rho(\theta)$$

and then the lim inf inequality taking the supremum in a.

Lim sup *inequality.* Let $u\in BV((0,1);\{-1,1\})$. We denote $S(u)=\{\theta_1,\cdots,\theta_N\}$ with $\theta_i<\theta_{i+1}$. With fixed $\eta>0$ there exist $T>0$ and $v\in H^1(-T,T)$ such that $v(\pm T)=\pm1$, $\int_{-T}^{T}(v^2-1)\,ds=0$, and

$$\int_{-T}^{T}(v')^2\,ds+\int_{-T}^{T}\left(\int_{-T}^{\sigma}(v^2-1)\,ds\right)^2 d\sigma\leq c_0+\eta\,. \tag{3.40}$$

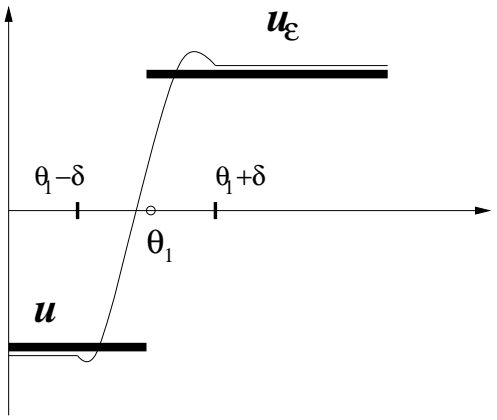


FIG. 2. *Recovery sequence with a Gibbs phenomenon.*

We denote $\delta = T\sqrt{\varepsilon}$, $K_i = [\theta_i + \delta, \theta_{i+1} - \delta]$ for $i = 1, \dots, N-1$, $K_N = [\theta_N + \delta, 1]$. We construct a sequence u_ε by setting

$$u_\varepsilon(\theta) = \begin{cases} v^i_\varepsilon(\pm\varepsilon^{-1/2}(\theta - \theta_i)) & \text{if } \theta \in [\theta_i - \delta, \theta_i + \delta], \quad i = 1, \dots, N, \\ u(\theta) & \text{if } \theta \in (0, \theta_1 - \delta) \cup_{i=1}^N K_i, \end{cases} \tag{3.41}$$

where $v^i_\varepsilon = v + t^i_\varepsilon \phi$ is defined as in Step 2 with ϕ given by (3.36) and $t^i_\varepsilon \in \mathbf{R}$ such that

$$\int_{-T}^{T} \rho^i_\varepsilon \left((v + t^i_\varepsilon \phi)^2 - 1\right) ds = 0 \quad \text{and} \quad t^i_\varepsilon \to 0 \quad \text{as} \quad \varepsilon \to 0 \tag{3.42}$$

for every $i = 1, \dots, N$, where $\rho^i_\varepsilon(s) = \rho(\theta_i + s\sqrt{\varepsilon})$. Note that the choice between the plus and minus sign, in (3.41), is made in such a way that the resulting function is continuous. The construction of u_ε is illustrated in Figure 2.

Note that, reasoning as in Step 2, we get

$$\begin{aligned} \lim_{\varepsilon \to 0} &\left(\int_{-T}^{T} \rho^i_\varepsilon \left(v^i_\varepsilon\right)'^2 ds + \int_{-T}^{T} \frac{1}{\rho^i_\varepsilon(\sigma)} \left(\int_{-T}^{\sigma} \rho^i_\varepsilon \left((v^i_\varepsilon)^2 - 1\right) ds \right)^2 d\sigma \right) \\ &= \rho(\theta_i) \left(\int_{-T}^{T} (v')^2 ds + \int_{-T}^{T} \left(\int_{-T}^{\sigma} (v^2 - 1) ds \right)^2 d\sigma \right). \end{aligned} \tag{3.43}$$

Hence,

$$\begin{aligned} F^{3/2}_\varepsilon(u_\varepsilon) = \sqrt{\varepsilon} \sum_{i=1}^{N} \int_{\theta_i - \delta}^{\theta_i + \delta} \rho \left(u'_\varepsilon\right)^2 d\theta \\ + \varepsilon^{-3/2} \sum_{i=1}^{N} \left(\int_{K_i} \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho \left(u^2_\varepsilon - 1\right) d\varphi \right)^2 d\theta \right. \\ \left. + \int_{\theta_i - \delta}^{\theta_i + \delta} \frac{1}{\rho(\theta)} \left(\int_0^\theta \rho \left(u^2_\varepsilon - 1\right) d\varphi \right)^2 d\theta \right). \end{aligned}$$

By the changes of variable $s = \varepsilon^{-1/2}(\varphi - \theta_i)$, $\sigma = \varepsilon^{-1/2}(\theta - \theta_i)$ and (3.42), we get that

$$\begin{aligned} F_\varepsilon^{3/2}(u_\varepsilon) &= \sum_{i=1}^N \int_{-T}^T \rho_\varepsilon^i \left(v_\varepsilon^i\right)'^2 ds \\ &\quad + \varepsilon^{-1/2} \sum_{i=1}^N \left(\int_{K_i} \frac{d\theta}{\rho(\theta)}\right) \left(\sum_{j=1}^i \int_{-T}^T \rho_\varepsilon^j \left((v_\varepsilon^j)^2 - 1\right) ds\right)^2 \\ &\quad + \varepsilon^{-1/2} \sum_{i=1}^N \int_{\theta_i-\delta}^{\theta_i+\delta} \frac{1}{\rho(\theta)} \left(\sum_{j=1}^{i-1} \int_{-T}^T \rho_\varepsilon^j \left((v_\varepsilon^j)^2 - 1\right) ds \right. \\ &\qquad \left. + \int_{-T}^{\varepsilon^{-1/2}(\theta-\theta_i)} \rho_\varepsilon^i \left((v_\varepsilon^i)^2 - 1\right) ds\right)^2 d\theta \\ &= \sum_{i=1}^N \int_{-T}^T \rho_\varepsilon^i \left(v_\varepsilon^i\right)'^2 ds + \int_{-T}^T \frac{1}{\rho_\varepsilon^i(\sigma)} \left(\int_{-T}^\sigma \rho_\varepsilon^i \left((v_\varepsilon^i)^2 - 1\right) ds\right)^2 d\sigma\,. \end{aligned}$$

By (3.40) and (3.43), we have

$$\begin{aligned} \limsup_{\varepsilon\to 0} F_\varepsilon^{3/2}(u_\varepsilon) &= \left(\sum_{i=1}^N \rho(\theta_i)\right) \left(\int_{-T}^T (v')^2 ds + \int_{-T}^T \left(\int_{-T}^\sigma (v^2 - 1) ds\right)^2 d\sigma\right) \\ &\le \left(\sum_{i=1}^N \rho(\theta_i)\right) (c_0 + \eta)\,. \end{aligned}$$

By the arbitrariness of η we get the lim sup inequality for every $u \in BV((0,1);\{-1,1\})$.

We now consider $u \in BV_{\rm loc}((0,1);\{-1,1\})$; let $u^a \in BV((0,1);\{-1,1\})$ be such that u^a converges to u strongly in $L^1(0,1)$ as $a \to 0^+$; e.g.,

$$u^a(\theta) = \begin{cases} u(a), & \theta \in [0,a), \\ u(\theta), & \theta \in [a, 1-a], \\ u(1-a), & \theta \in (1-a, 1], \end{cases}$$

with $0 < a < 1$ and $a, 1-a \notin S(u)$. By the lower semicontinuity of the Γ-lim sup, we have that

$$\begin{aligned} \Gamma\text{-}\limsup_{\varepsilon\to 0} F_\varepsilon^{3/2}(u) &\le \liminf_{a\to 0^+} \Big(\Gamma\text{-}\limsup_{\varepsilon\to 0} F_\varepsilon^{3/2}(u^a)\Big) \\ &\le c_0 \liminf_{a\to 0^+} \sum_{\theta\in S(u^a)} \rho(\theta) \le c_0 \sum_{\theta\in S(u)} \rho(\theta) \end{aligned}$$

for every $u \in BV_{\rm loc}((0,1);\{-1,1\})$ (see [6, Remark 1.29]). □

4. The case $\alpha = 4/3$ with constrained phase transitions. If we define G_ε^0 as the restriction of F_ε^0 to the space of functions $u : (0,1) \to [-1,1]$, then by Remark 2 the Γ-limit of G_ε^0 is identically 0. Still, one may find another scaling, $\alpha = 4/3$, such that the Γ-limit of $G_\varepsilon^{4/3} = \varepsilon^{-4/3} G_\varepsilon^0$ is not trivial. We consider for simplicity $\rho \equiv 1$ so that

$$G_\varepsilon^{4/3}(u) = \varepsilon^{2/3} \int_0^1 (u')^2 \, d\theta + \varepsilon^{-4/3} \int_0^1 \left(\int_0^\theta (u^2 - 1)\, d\varphi\right)^2 d\theta$$

for every $u \in H^1((0,1);[-1,1])$. Note that since $\varepsilon^{-4/3}F_\varepsilon^0 = \varepsilon^{1/6}F_\varepsilon^{3/2}$, by section 3 the Γ-limit of $\varepsilon^{-4/3}F_\varepsilon^0$ with respect to the strong L^1-convergence is zero. Hence, the constraint $|u| \le 1$ completely changes the characteristic scaling of the energy.

THEOREM 4.1 (compactness). *Let $(u_\varepsilon) \in H^1((0,1);[-1,1])$ be a sequence such that $\sup_\varepsilon G_\varepsilon^{4/3}(u_\varepsilon) < +\infty$; then up to subsequences, (u_ε) converges strongly in $L^1(0,1)$ to $u \in BV_{\rm loc}((0,1);\{-1,1\})$.*

Proof. Let $\eta_\varepsilon^\pm \neq \pm 1$ such that $-1 < \eta_\varepsilon^+, \eta_\varepsilon^- < 1$. We denote by $(\delta_\varepsilon^-, \delta_\varepsilon^+)$ an interval such that $u_\varepsilon(\delta_\varepsilon^-) = \eta_\varepsilon^-$, $u_\varepsilon(\delta_\varepsilon^+) = \eta_\varepsilon^+$ and u_ε takes values between η_ε^- and η_ε^+. In what follows, we use the notation $\eta^\pm = \eta_\varepsilon^\pm$, $\delta^\pm = \delta_\varepsilon^\pm$ to not overburden notation. By the constraint, $(u_\varepsilon^2 - 1)$ never changes sign; hence, we can select in the energy $G_\varepsilon^{4/3}(u_\varepsilon)$ the most significant contribution which permits us to easily estimate $(\delta^+ - \delta^-)$. More precisely, for every fixed $0 < a < 1$, we consider $\delta^+ \le 1 - a$; then,

$$\begin{aligned} G_\varepsilon^{4/3}(u_\varepsilon) &\ge \varepsilon^{2/3} \int_{\delta^-}^{\delta^+} (u')^2 \, d\theta + \varepsilon^{-4/3} \int_{\delta^+}^{1} \left(\int_{\delta^-}^{\delta^+} (u_\varepsilon^2 - 1) \, d\varphi \right)^2 d\theta \\ &\ge |\eta^+ - \eta^-|^2 \left(\frac{\varepsilon^{2/3}}{\delta} \right) + a\, \lambda^2 \left(\frac{\delta}{\varepsilon^{2/3}} \right)^2, \end{aligned}$$

where $\lambda := ||\eta^-|^2 - 1| \wedge ||\eta^+|^2 - 1|$, i.e., the minimum distance of $|\eta^\pm|^2$ from 1. Hence, if $|\eta^+ - \eta^-| \ge \zeta > 0$, with ζ independent of ε, we have that $\delta/\varepsilon^{2/3}$ is bounded; i.e., there exist two positive constants α_1, α_2 such that $\alpha_1 \varepsilon^{2/3} \le \delta \le \alpha_2 \varepsilon^{2/3}$. Moreover, the number of intervals (δ^-, δ^+) is finite in $[0, 1-a]$ for every $0 < a < 1$. Then (u_ε) converges in measure to $u \in BV_{\rm loc}((0,1);\{-1,1\})$, and since $|u_\varepsilon| \le 1$ we can conclude that, up to subsequences, (u_ε) converges strongly to u in L^1. ∎

Remark 4. Note that in general we cannot expect $u \in BV((0,1);\{-1,1\})$. To show this we construct a sequence (u_ε) with $\sup_\varepsilon G_\varepsilon^{4/3}(u_\varepsilon) < +\infty$ and strongly converging in L^1 to $u \in BV_{\rm loc}((0,1);\{-1,1\})$ with infinitely many jump points. To this end, consider a strictly increasing sequence $(\theta_i) \in (0,1)$ such that $\sup_{i \in \mathbf{N}} \theta_i = 1$, and let $v_i(s) = (s + T_i/T_i) - 1$ for $s \in [-T_i, T_i]$. With fixed $k \in \mathbf{N}$ and ε small enough, we define the sequence (u_ε^k) as

$$u_\varepsilon^k(\theta) = \begin{cases} v_i(\pm \varepsilon^{-2/3}(\theta - \theta_i)), & \theta \in [\theta_i - T_i\varepsilon^{2/3}, \theta_i + T_i\varepsilon^{2/3}], \quad i = 1, \dots, k, \\ \pm 1 & \text{otherwise in } [0,1], \end{cases}$$

where the choice between the plus and minus sign is made in such a way that the resulting function u_ε^k is continuous. For every $k \in \mathbf{N}$, we have that

$$\begin{aligned} G_\varepsilon^{4/3}(u_\varepsilon^k) &= \sum_{i=1}^{k} \left(\frac{2}{T_i} + (\theta_{i+1} - \theta_i) \left(c \sum_{j \le i} T_j \right)^2 \right) + O(\varepsilon^{2/3}) \\ &\le \sum_{i \in \mathbf{N}} \left(\frac{2}{T_i} + (\theta_{i+1} - \theta_i) \left(c \sum_{j \le i} T_j \right)^2 \right) + O(\varepsilon^{2/3}) \end{aligned}$$

(see a similar computation in Theorem 4.2 for the proof of the lim sup inequality). If we fix $T_i = i^\beta$, with $\beta > 1$, and (θ_i) such that $(\theta_{i+1} - \theta_i) = \gamma\, i^{(-3\beta-2)}$, with γ satisfying the condition $\gamma \sum_{i \in \mathbf{N}} i^{(-3\beta-2)} = (1 - \theta_1)$, then $\sup_\varepsilon G_\varepsilon^{4/3}(u_\varepsilon^k) \le c$, with c independent on k. Therefore, if $(\theta_i)_{i \in \mathbf{N}}$ is an increasing sequence of points distributed in $(0,1)$

as above, for every fixed $k \in \mathbf{N}$, we can construct a suitable sequence (u_ε^k) strongly converging in L^1 to $u^k \in BV((0,1);\{-1,1\})$, as $\varepsilon \to 0$, with $\sup_\varepsilon G_\varepsilon^{4/3}(u_\varepsilon^k) \le c$ and $S(u^k) = \{\theta_1, \dots, \theta_k\}$. We now consider $u \in BV_{\rm loc}((0,1);\{-1,1\})$ such that $S(u) = (\theta_i)_{i\in\mathbf{N}}$ and $u = u^k$ in $[0, \theta_{k+1})$; then u^k converges strongly in L^1 to u as k tends to $+\infty$. By a diagonal procedure we may extract from (u_ε^k) a subsequence with bounded energy and strongly converging to u in L^1.

THEOREM 4.2 (nonlocal Γ-limit). *We have*

$$\begin{aligned}\Gamma(L^1)\text{-}\lim_{\varepsilon\to 0} G_\varepsilon^{4/3}(u)\\ = \inf_{T>0} \inf\left\{ \sum_{i\in I}\left(\int_{-T}^{T} (v_i')^2\, ds + (\theta_{i+1}-\theta_i)\left(\sum_{j=1}^{i}\int_{-T}^{T}(v_j^2-1)\, ds\right)^2\right) :\right.\\ \left. v_i \in H^1((-T,T);[-1,1]),\quad v_i(\pm T) = \pm 1\right\}\end{aligned}$$

for every $u \in BV_{\rm loc}((0,1);\{-1,1\})$, where $I = \{i \in \mathbf{N} : \theta_i \in S(u)\,, \theta_i < \theta_{i+1}\}$.

Proof. Let $u \in BV_{\rm loc}((0,1);\{-1,1\})$, and let $(u_\varepsilon) \in H^1((0,1);[-1,1])$ be a sequence strongly converging to u in L^1 such that

$$\liminf_{\varepsilon\to 0} G_\varepsilon^{4/3}(u_\varepsilon) < +\infty\,.$$

For every fixed $0 < a < 1$, by Theorem 4.1 the limit function u has a finite number of discontinuity points in the interval $(0, 1-a]$; i.e., $S(u) \cap (0, 1-a] = \{\theta_1, \dots, \theta_{N(a)}\}$ with $\theta_i < \theta_{i+1}$. Up to subsequences, $u_\varepsilon \to u$ for a.e. $\theta \in (0,1)$, as ε tends to 0; hence, with fixed $\eta \in (0,1)$, we consider δ_i^1, δ_i^2 such that

$$u_\varepsilon(\theta_i - \delta_i^1) = -(1-\eta)\,, \qquad u_\varepsilon(\theta_i + \delta_i^2) = 1-\eta$$

or

$$u_\varepsilon(\theta_i - \delta_i^1) = 1-\eta\,, \qquad u_\varepsilon(\theta_i + \delta_i^2) = -(1-\eta)$$

for $i = 1, \dots, N(a)$. The following estimate is obtained by eliminating all the contributions of u_ε on the intervals where the sequence takes values "close" to $\{-1,1\}$; this choice is justified by the construction of the optimal sequence, in the lim sup inequality, that will be equal to $\{-1,1\}$ on such intervals (see (4.2)). We have then

$$\begin{aligned}G_\varepsilon^{4/3}(u_\varepsilon) \ge \varepsilon^{2/3}\sum_{i=1}^{N(a)}\int_{I_i}(u_\varepsilon')^2 d\theta + \varepsilon^{-4/3}\left(\sum_{i=1}^{N(a)}\int_{K_i}\left(\sum_{j=1}^{i}\int_{I_j}(u_\varepsilon^2-1)\, d\varphi\right)^2 d\theta\right.\\ \left. + \sum_{i=1}^{N(a)}\int_{I_i}\left(\sum_{j=1}^{i-1}\int_{I_j}(u_\varepsilon^2-1)\, d\varphi + \int_{\theta_i-\delta_i^1}^{\theta}(u_\varepsilon^2-1)\, d\varphi\right)^2 d\theta\right),\end{aligned}$$

where $I_i = [\theta_i - \delta_i^1, \theta_i + \delta_i^2]$ for $i = 1, \dots, N(a)$ and $K_i = (\theta_i + \delta_i^2, \theta_{i+1} - \delta_{i+1}^1)$ for $i = 1, \dots, N(a)-1$, $K_{N(a)} = (\theta_{N(a)} + \delta_{N(a)}^2, 1-a)$. We make the change of variable

$$w_j(s) = u_\varepsilon\left(\varepsilon^{2/3}s + \theta_j + \frac{\delta_j^2 - \delta_j^1}{2}\right);$$

hence, setting $T_j = \varepsilon^{-2/3}\delta_j$ with $\delta_j = \left(\frac{\delta_j^2+\delta_j^1}{2}\right)$ and

$$T_i(\theta) = \varepsilon^{-2/3}\left(\theta - \theta_i - \frac{(\delta_i^2 - \delta_i^1)}{2}\right),$$

we get

$$G_\varepsilon^{4/3}(u_\varepsilon) \geq \sum_{i=1}^{N(a)} \left(\int_{-T_i}^{T_i} (w_i')^2\, ds + \int_{K_i} \left(\sum_{j=1}^{i} \int_{-T_j}^{T_j} (w_j^2 - 1)\, ds \right)^2 d\theta \right.$$
$$\left. + \int_{I_i} \left(\sum_{j=1}^{i-1} \int_{-T_j}^{T_j} (w_j^2 - 1)\, ds + \int_{-T_i}^{T_i(\theta)} (w_i^2 - 1)\, ds \right)^2 d\theta \right).$$

We denote now

$$B_i = \int_{-T_i}^{T_i} (w_i')^2\, ds\,, \qquad A_j = \int_{-T_j}^{T_j} (w_j^2 - 1)\, ds\,;$$

by the change of variable $\sigma = T_i(\theta)$, we get

$$G_\varepsilon^{4/3}(u_\varepsilon) \geq \sum_{i=1}^{N(a)} B_i + \sum_{i=1}^{N(a)} \left((2\delta_i + \theta_{i+1} - \theta_i - \delta_{i+1}^1 - \delta_i^2) \left(\sum_{j=1}^{i} A_j \right)^2 \right)$$
$$+ \varepsilon^{2/3} \sum_{i=1}^{N(a)} \left(\int_{-T_i}^{T_i} \left(\int_\sigma^{T_i} (w_i^2 - 1)\, ds \right)^2 d\sigma \right.$$
$$\left. - 2 \left(\sum_{j=1}^{i} A_j \right) \int_{-T_i}^{T_i} \left(\int_\sigma^{T_i} (w_i^2 - 1)\, ds \right) d\sigma \right)$$
$$= \sum_{i=1}^{N(a)} \left(B_i + (\theta_{i+1} - \theta_i) \left(\sum_{j=1}^{i} A_j \right)^2 \right) + \sum_{i=1}^{N(a)} (\delta_i^1 - \delta_{i+1}^1) \left(\sum_{j=1}^{i} A_j \right)^2$$
$$+ \varepsilon^{2/3} \left(\sum_{i=1}^{N(a)} \int_{-T_i}^{T_i} \left(\int_\sigma^{T_i} (w_i^2 - 1)\, ds \right)^2 d\sigma \right.$$
$$\left. - 2 \left(\sum_{j=1}^{i} A_j \right) \int_{-T_i}^{T_i} \left(\int_\sigma^{T_i} (w_i^2 - 1)\, ds \right) d\sigma \right),$$

where $\theta_{N(a)+1} = 1 - a$ and $\delta_{N(a)+1}^1 = 0$. Hence,

$$G_\varepsilon^{4/3}(u_\varepsilon) \geq \sum_{i=1}^{N(a)} \left(\int_{-T_i}^{T_i} (w_i')^2\, ds + (\theta_{i+1} - \theta_i) \left(\sum_{j=1}^{i} \int_{-T_j}^{T_j} (w_j^2 - 1)\, ds \right)^2 \right) + O(\varepsilon^{2/3}),$$

where $w_i(\pm T_i)$ is equal to $\pm(1-\eta)$ or $\mp(1-\eta)$, and

$$\begin{aligned}&\liminf_{\varepsilon\to 0} G_\varepsilon^{4/3}(u_\varepsilon)\\ &\quad\geq \inf_{T>0}\ \inf\left\{\sum_{i=1}^{N(a)}\left(\int_{-T}^{T}(v_i')^2\,ds+(\theta_{i+1}-\theta_i)\left(\sum_{j=1}^{i}\int_{-T}^{T}(v_j^2-1)\,ds\right)^2\right):\right.\\ &\qquad\left. v_i\in H^1((-T,T);[-1,1]),\quad v_i(\pm T)=\pm(1-\eta)\right\},\end{aligned}$$

where, by symmetry, we may fix the boundary conditions as $v_i(\pm T)=\pm(1-\eta)$. We may now first pass to the limit as $\eta\to 0$ and then take the supremum on a; i.e.,

$$\begin{aligned}&\liminf_{\varepsilon\to 0} G_\varepsilon^{4/3}(u_\varepsilon)\\ &\geq \sup_{0<a<1}\inf_{T>0}\ \inf\left\{\sum_{i=1}^{N(a)}\left(\int_{-T}^{T}(v_i')^2\,ds+(\theta_{i+1}-\theta_i)\left(\sum_{j=1}^{i}\int_{-T}^{T}(v_j^2-1)\,ds\right)^2\right):\right.\\ &\qquad\left. v_i\in H^1((-T,T);[-1,1]),\quad v_i(\pm T)=\pm 1\right\}\\ &=\inf_{T>0}\ \inf\left\{\sum_{i\in I}\left(\int_{-T}^{T}(v_i')^2\,ds+(\theta_{i+1}-\theta_i)\left(\sum_{j=1}^{i}\int_{-T}^{T}(v_j^2-1)\,ds\right)^2\right):\right.\\ &\qquad\left. v_i\in H^1((-T,T);[-1,1]),\quad v_i(\pm T)=\pm 1\right\},\end{aligned}$$

where we have labeled the points in $S(u)$ by a set of indices $I\subset\mathbf{N}$ in such a way that $\theta_i<\theta_{i+1}$.

We now check the lim sup inequality. Let $u\in BV((0,1);\{-1,1\})$ with $S(u)=\{\theta_1,\dots,\theta_N\}$ and $\theta_i<\theta_{i+1}$. We denote

$$\begin{aligned}G^{4/3}(u)=\inf_{T>0}\ \inf\left\{\sum_{i=1}^{N}\left(\int_{-T}^{T}(v_i')^2\,ds+(\theta_{i+1}-\theta_i)\left(\sum_{j=1}^{i}\int_{-T}^{T}(v_j^2-1)\,ds\right)^2\right):\right.\\ \left. v_i\in H^1((-T,T);[-1,1]),\quad v_i(\pm T)=\pm 1\right\}.\end{aligned}$$

With fixed $\eta>0$, there exist $T>0$ and $(v_1,\dots,v_N)\in H^1((-T,T);[-1,1])$ such that $v_i(\pm T)=\pm 1$ and

$$\sum_{i=1}^{N}\left(\int_{-T}^{T}(v_i')^2\,ds+(\theta_{i+1}-\theta_i)\left(\sum_{j=1}^{i}\int_{-T}^{T}(v_j^2-1)\,ds\right)^2\right)\leq G^{4/3}(u)+\eta\,. \tag{4.1}$$

We denote $\delta = T\varepsilon^{2/3}$; $I_i = [\theta_i - \delta, \theta_i + \delta]$ for $i = 1, \dots, N$; $K_i = (\theta_i + \delta, \theta_{i+1} - \delta)$ for $i = 1, \dots, N-1$; and $K_N = (\theta_N + \delta, 1]$. We construct a sequence u_ε by setting

$$u_\varepsilon(\theta) = \begin{cases} v_i(\pm\varepsilon^{-2/3}(\theta - \theta_i)) & \text{if } \theta \in I_i, \quad i = 1, \dots, N, \\ u(\theta) & \text{if } \theta \in (0, \theta_1 - \delta) \cup \left(\cup_{i=1}^N K_i\right), \end{cases} \tag{4.2}$$

where the choice between the plus and minus sign is made in such a way that the resulting function is continuous. Hence, by the change of variables $s = \varepsilon^{-2/3}(\theta - \theta_i)$, we get

$$\begin{aligned} G_\varepsilon^{4/3}(u_\varepsilon) &= \varepsilon^{2/3} \sum_{i=1}^N \int_{I_i} (u_\varepsilon')^2 \, d\theta + \varepsilon^{-4/3} \sum_{i=1}^N \left(\int_{K_i} \left(\int_0^\theta (u_\varepsilon^2 - 1) \, d\varphi \right)^2 d\theta \right) \\ &\quad + \varepsilon^{-4/3} \sum_{i=1}^N \left(\int_{I_i} \left(\int_0^\theta (u_\varepsilon^2 - 1) \, d\varphi \right)^2 d\theta \right) \\ &= \sum_{i=1}^N \left(\int_{-T}^T (v_i')^2 \, ds + \int_{K_i} \left(\sum_{j=1}^i \int_{-T}^T (v_j^2 - 1) \, ds \right)^2 d\theta \right. \\ &\quad \left. + \int_{I_i} \left(\sum_{j=1}^{i-1} \int_{-T}^T (v_j^2 - 1) \, ds + \int_{-T}^{T_i(\theta)} (v_i^2 - 1) \, ds \right)^2 d\theta \right), \end{aligned}$$

where $T_i(\theta) = \varepsilon^{-2/3}(\theta - \theta_i)$. Setting

$$B_i = \int_{-T}^T (v_i')^2 \, ds \,, \qquad A_j = \int_{-T}^T (v_j^2 - 1) \, ds,$$

we then have

$$\begin{aligned} G_\varepsilon^{4/3}(u_\varepsilon) &= \sum_{i=1}^N \left(B_i + (\theta_{i+1} - \theta_i - 2\delta) \left(\sum_{j=1}^i A_j \right)^2 \right) \\ &\quad + \sum_{i=1}^N \int_{I_i} \left(\sum_{j=1}^i A_j - \int_{T_i(\theta)}^T (v_i^2 - 1) \, ds \right)^2 d\theta \\ &= \sum_{i=1}^N \left(B_i + (\theta_{i+1} - \theta_i) \left(\sum_{j=1}^i A_j \right)^2 \right) + \sum_{i=1}^N \int_{I_i} \left(\int_{T_i(\theta)}^T (v_i^2 - 1) \, ds \right)^2 d\theta \\ &\quad - 2 \sum_{i=1}^N \left(\sum_{j=1}^i A_j \right) \int_{I_i} \left(\int_{T_i(\theta)}^T (v_i^2 - 1) \, ds \right) d\theta \,. \end{aligned}$$

By (4.1) and the change of variable $\sigma = T_i(\theta)$, we get that

$$G_\varepsilon^{4/3}(u_\varepsilon) = \sum_{i=1}^{N} \left(B_i + (\theta_{i+1} - \theta_i) \left(\sum_{j=1}^{i} A_j \right)^2 \right) + O(\varepsilon^{2/3})$$

$$\leq G^{4/3}(u) + \eta + O(\varepsilon^{2/3})\,.$$

Passing to the limit as ε tends to 0, by the arbitrariness of η, we get the lim sup inequality for every $u \in BV((0,1);\{-1,1\})$.

We now consider $u \in BV_{\mathrm{loc}}((0,1);\{-1,1\})$. There exists $u^a \in BV((0,1);\{-1,1\})$ such that u^a converges to u strongly in $L^1(0,1)$ as $a \to 0^+$; i.e.,

$$u^a(\theta) = \begin{cases} u(a), & \theta \in [0,a), \\ u(\theta), & \theta \in [a, 1-a], \\ u(1-a), & \theta \in (1-a, 1], \end{cases}$$

with $0 < a < 1$ and $a, 1-a \notin S(u)$. Hence, by the lower semicontinuity of the Γ-lim sup, we have that

$$\begin{aligned}
&\Gamma\text{-}\limsup_{\varepsilon\to 0} G_\varepsilon^{4/3}(u) \\
&\quad \leq \liminf_{a\to 0^+} \left(\Gamma\text{-}\limsup_{\varepsilon\to 0} G_\varepsilon^{4/3}(u^a) \right) \\
&\quad \leq \liminf_{a\to 0^+} \inf_{T>0} \inf \left\{ \sum_{i\in I(a)} \left(\int_{-T}^{T} (v_i')^2\, ds + (\theta_{i+1} - \theta_i) \left(\sum_{j=1}^{i} \int_{-T}^{T} (v_j^2 - 1)\, ds \right)^2 \right) : \right. \\
&\qquad\qquad \left. v_i \in H^1((-T,T);[-1,1]), \quad v_i(\pm T) = \pm 1 \right\} \\
&\quad \leq \inf_{T>0} \inf \left\{ \sum_{i\in I} \left(\int_{-T}^{T} (v_i')^2\, ds + (\theta_{i+1} - \theta_i) \left(\sum_{j=1}^{i} \int_{-T}^{T} (v_j^2 - 1)\, ds \right)^2 \right) : \right. \\
&\qquad\qquad \left. v_i \in H^1((-T,T);[-1,1]), \quad v_i(\pm T) = \pm 1 \right\},
\end{aligned}$$

where $I(a) = \{i \in \mathbf{N} : \theta_i \in S(u^a)\,,\, \theta_i < \theta_{i+1}\}$ and $I = \{i \in \mathbf{N} : \theta_i \in S(u)\,,\, \theta_i < \theta_{i+1}\}$. □

Acknowledgments. We gratefully acknowledge the careful reading of the manuscript by the referees.

REFERENCES

[1] S. S. Antman, *The eversion of thick spherical shells*, Arch. Rational Mech. Anal., 70 (1979), pp. 113–123.

[2] L. Bauer, E. L. Reiss, and H. B. Keller, *Axisymmetric buckling of hollow spheres and hemispheres*, Comm. Pure Appl. Math., 23 (1970), pp. 529–568.

[3] H. Ben Belgacem, S. Conti, A. DeSimone, and S. Müller, *Energy scaling of compressed elastic films—three-dimensional elasticity and reduced theories*, Arch. Ration. Mech. Anal., 164 (2002), pp. 1–37.
[4] M. S. Berger, *On von Karman's equations and the buckling of a thin elastic plate* I: *The clamped plate*, Comm. Pure Appl. Math., 20 (1967), pp. 687–719.
[5] A. Braides, *Approximation of Free-Discontinuity Problems*, Springer-Verlag, Berlin, 1998.
[6] A. Braides, Γ-*Convergence for Beginners*, Oxford University Press, Oxford, UK, 2002.
[7] A. Braides, *A handbook of* Γ-*convergence*, in Handbook of Differential Equations: Stationary Partial Differential Equations, Vol. 3, M. Chipot and P. Quittner, eds., Elsevier, Amsterdam, 2006, pp. 101–213.
[8] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford University Press, Oxford, UK, 1998.
[9] A. Braides, I. Fonseca, and G. Francfort, $3D$–$2D$ *asymptotic analysis for inhomogeneous thin films*, Indiana Univ. Math. J., 49 (2000), pp. 1367–1404.
[10] S. Conti and F. Maggi, *Confining Thin Elastic Sheets and Folding*, preprint, Universität Duisburg-Essen, Duisburg, Germany, 2005.
[11] G. Dal Maso, *An Introduction to* Γ-*Convergence*, Birkhäuser Boston, Boston, MA, 1993.
[12] G. B. Folland, *Real Analysis. Modern Techniques and Their Applications*, Wiley, New York, 1984.
[13] G. Friesecke, R. D. James, and S. Müller, *A hierarchy of plate models derived from nonlinear elasticity by Gamma-convergence*, Arch. Ration. Mech. Anal., 180 (2006), pp. 183–236.
[14] G. Geymonat and A. Leger, *Nonlinear spherical caps and associated plate and membrane problems*, J. Elasticity, 57 (1999), pp. 171–200.
[15] G. Geymonat, M. Rosati, and V. Valente, *Numerical analysis for eversion in elastic spherical caps equilibrium*, Comput. Methods Appl. Mech. Engrg., 75 (1989), pp. 39–52.
[16] G. Geymonat, M. Rosati, and V. Valente, *The maximal thickness for everted equilibrium shapes of elastic spherical caps*, Calcolo, 27 (1990), pp. 103–125.
[17] H. Le Dret and A. Raoult, *The membrane shell model in nonlinear elasticity: A variational asymptotic derivation*, J. Nonlinear Sci., 6 (1996), pp. 59–84.
[18] L. Modica and S. Mortola, *Un esempio di* Γ-*convergenza*, Boll. Un. Mat. Ital. B, 14 (1977), pp. 285–299.
[19] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.
[20] P. Podio-Guidugli, M. Rosati, A. Schiaffino, and V. Valente, *Equilibrium of an elastic spherical cap pulled at the rim*, SIAM J. Math. Anal., 20 (1989), pp. 643–663.
[21] E. L. Reiss, *Bifurcation buckling of spherical caps*, Comm. Pure Appl. Math., 18 (1965), pp. 65–82.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 977–984

POISSON KERNELS AS EXPANSIONS IN q-RACAH POLYNOMIALS*

CLAUDIO ALBANESE† AND STEPHAN LAWI‡

Abstract. This paper concerns stochastic processes on chains of arbitrary length whose Poisson kernel can be expressed in terms of the q-Racah polynomials, the most general q-deformed orthogonal polynomials in the discrete series of the Askey scheme. We give a new interpretation of this kernel as the probability transition density for a subordinated Markov process with only nearest neighbor hops. As an application, we give an elementary proof and extend a positivity result for a class of Poisson kernels which Gasper and Rahman established with direct methods.

Key words. orthogonal polynomials, basic hypergeometric polynomials, q-Racah polynomials, Markov chains, subordinators, positivity of Poisson kernels

AMS subject classifications. 33D45, 60J27, 60J35, 60J99

DOI. 10.1137/S0036141003435035

1. Introduction.

In this paper, we give general conditions that allow one to construct Markov processes whose infinitesimal generator admits the q-Racah polynomials as eigenstates.

DEFINITION 1.1. *For $0 < q < 1$, the q-Racah polynomials*

$$R_n(\mu(x)) := R_n(\mu(x); \alpha, \beta, \gamma, \delta | q)$$

are defined using the q-deformed hypergeometric function ${}_4\phi_3$ as follows:

$$R_n(\mu(x)) = {}_4\phi_3 \left(\begin{matrix} q^{-n},\ \alpha\beta q^{n+1}, q^{-x}, \gamma\delta q^{x+1} \\ \alpha q, \beta\delta q, \gamma q \end{matrix} \middle| \, q; q \right) \tag{1.1}$$

for $n = 0, 1, \ldots, N$ and where $\mu(x) = q^{-x} + \gamma\delta q^{x+1}$. Moreover, they must belong to one of the three families defined by the additional condition that either $\alpha = q^{-N-1}$ or $\beta\delta = q^{-N-1}$ or $\gamma = q^{-N-1}$.

DEFINITION 1.2. *Let Λ_N denote the set $\{0, 1, \ldots, N\}$ and let $l^2(\Lambda_N)$ be the Hilbert space of the real valued functions on Λ_N.*

The objective of this paper is to construct a class of Markov processes supported on the set Λ_N whose transitional probability density can be expressed as a series in the q-Racah polynomials. Whether this can be done is a natural question, as the q-Racah polynomials satisfy an orthogonality relation with respect to a discrete weight supported on the set Λ_N (see [1], [10], [7], and the references therein for the properties of orthogonal polynomials in the Askey scheme). Namely,

$$\sum_{x \in \Lambda_N} R_m(\mu(x)) R_n(\mu(x)) w(x) = h_n \delta_{nm}, \tag{1.2}$$

*Received by the editors September 18, 2003; accepted for publication (in revised form) July 6, 2005; published electronically October 16, 2006.
http://www.siam.org/journals/sima/38-3/43503.html
†Department of Mathematics, Imperial College, London SW7 2AZ, UK (claudio.albanese@imperial.ac.uk).
‡Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599, Université Paris VI & Université Paris VII, 4 Place Jussieu, 75252 Paris Cedex 05, France (lawi@proba.jussieu.fr).

where the weight is

$$w(x) = \frac{(\alpha q, \beta\delta q, \gamma q, \gamma\delta q; q)_x}{(q, \alpha^{-1}\gamma\delta q, \beta^{-1}\gamma q, \delta q; q)_x} \frac{1 - \gamma\delta q^{2x+1}}{(\alpha\beta q)^x (1 - \gamma\delta q)} \tag{1.3}$$

and the normalization factor is

$$h_n = \frac{(\alpha^{-1}\beta^{-1}\gamma, \alpha^{-1}\delta, \beta^{-1}, \gamma\delta q^2; q)_\infty}{(\alpha^{-1}\beta^{-1}q^{-1}, \alpha^{-1}\gamma\delta q, \beta^{-1}\gamma q, \delta q; q)_\infty} \frac{(1 - \alpha\beta q)(\gamma\delta q)^n}{1 - \alpha\beta q^{2n+1}} \frac{(q, \alpha\beta\gamma^{-1}q, \alpha\delta^{-1}q, \beta q; q)_n}{(\alpha q, \alpha\beta q, \beta\delta q, \gamma q; q)_n}. \tag{1.4}$$

Following [7], we denote $(a; q)_k$ for the q-analogue of the Pochhammer symbols:

$$(a; q)_0 = 1 \quad \text{and} \quad (a; q)_k = \prod_{i=0}^{k-1} (1 - a\ q^i) \tag{1.5}$$

for $k > 0$, with $(a_1, \dots, a_r; q)_k = (a_1; q)_k \cdots (a_r; q)_k$.

Recall now the following definitions of the Markov semigroup and the Bernstein function.

DEFINITION 1.3. *The one-parameter family of operators $(\mathcal{K}_t)_{t\geq 0}$ on $l^2(\Lambda_N)$ is said to be a Markov semigroup if its kernel $K_t(x, y)$ satisfies the following conditions for all $t, t' \geq 0$, $x, y \in \Lambda_N$:*

(i) *nonnegativity: $K_t(x, y) \geq 0$;*
(ii) *initial condition: $K_0(x, y) = \delta_{x,y}$;*
(iii) *contraction property: $\sum_{y\in\Lambda_N} K_t(x, y) \leq 1$;*
(iv) *semigroup property: $K_t(x, y)K_{t'}(x, y) = K_{t+t'}(x, y)$.*

DEFINITION 1.4. *Let $\Phi \in C^\infty([0, \infty))$ with $\Phi \geq 0$. Φ is a Bernstein function if $(-1)^n \Phi^{(n)} \leq 0$ for all integer $n > 0$.*

Our main result can be stated as follows.

THEOREM 1.5. *Assume that the parameters $q, \alpha, \beta, \gamma, \delta$ satisfy the restrictions of Definition 1.1 for the q-Racah polynomials. Assume also that the following two functions are negative or zero:*

$$B(x) = \frac{(1 - \alpha q^{x+1})(1 - \beta\delta q^{x+1})(1 - \gamma q^{x+1})(1 - \gamma\delta q^{x+1})}{(1 - \gamma\delta q^{2x+1})(1 - \gamma\delta q^{2x+2})},$$

$$D(x) = \frac{q(1 - q^x)(1 - \delta q^x)(\beta - \gamma q^x)(\alpha - \gamma\delta q^x)}{(1 - \gamma\delta q^{2x})(1 - \gamma\delta q^{2x+1})}.$$

Then for all Bernstein functions Φ the following series represents the kernel of a Markov semigroup:

$$K_t^\Phi(x, y) = \sum_{n=0}^{N} \frac{e^{-t\Phi(-\lambda_n)}}{h_n} R_n(\mu(x)) R_n(\mu(y))\ w(y), \tag{1.6}$$

where

$$\lambda_n = (1 - q^{-n})(1 - \alpha\beta q^{n+1}). \tag{1.7}$$

We also prove that there exists a Bernstein function such that $\Phi(-\lambda_n) = n$ for all integers $n = 1, 2, \dots, N$ so that we obtain the following corollary.

COROLLARY 1.6. *If in addition to the hypothesis of the theorem we also have that* $\alpha\beta \geq 0$, *then the function*

$$\Phi(z) = \frac{1}{|\ln q|} \ln \left(\frac{1+\alpha\beta q+z}{2} + \sqrt{\left(\frac{1+\alpha\beta q+z}{2}\right)^2 - \alpha\beta q} \right) \tag{1.8}$$

is a Bernstein function such that $\Phi(-\lambda_n) = n$ *for all integers* $n = 1, 2, \ldots, N$, *and the following series represents the kernel of a Markov semigroup on* Λ_N:

$$K_t^{\Phi}(x,y) = \sum_{n=0}^{N} \frac{e^{-nt}}{h_n} R_n(\mu(x)) R_n(\mu(y))\ w(y) - \max\left(1 - (\alpha\beta q)^{-\frac{t}{|\ln q|}}, 0\right)\ w(y). \tag{1.9}$$

This corollary extends the result by Gasper and Rahman in [6], where positivity of the Poisson kernel, defined as the series

$$\sum_{n=0}^{N} \frac{e^{-nt}}{h_n} R_n(\mu(x)) R_n(\mu(y)), \tag{1.10}$$

was established for the third family of q-Racah polynomials, i.e., $\gamma = q^{-N-1}$, and under the additional conditions that $0 < \alpha q < 1$, $0 \leq \beta q < 1$, and $0 < \delta < \alpha q^N$. The proof in [6] is based on a term by term analysis of the series expansion and resummation formulas of independent interest. Our argument is instead more elementary and based on probabilistic considerations.

2. Background and proofs. As a preparation to the proof of our results, we recall the basics of the theory of Markov processes and Bochner subordination. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space equipped with the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ satisfying the usual conditions; i.e., the filtration is right continuous and $\mathcal{F}_0$ contains all events of P-measure 0 in $\mathcal{F}$. Let $X = (X_t)_{t\geq 0}$ be an $\mathcal{F}_t$-adapted stochastic process on $D \subseteq \mathbb{R}$.

DEFINITION 2.1. *The stochastic process* X *is called a Markov process if*

$$\mathrm{E}\left[\ f(X_t) \mid \mathcal{F}_s\ \right] = \mathrm{E}\left[\ f(X_t) \mid X_s\ \right] \tag{2.1}$$

for all bounded measurable function $f : D \to \mathbb{R}$ *and all* $t \geq s \geq 0$.

The family of operators $(\mathcal{K}_t)_{t\geq 0}$ from Definition 1.3 defines the transitional probability density of a Markov process on $l^2(\Lambda_N)$ as follows:

$$\mathcal{K}_{t-s} f(x) = \mathrm{E}\left[\ f(X_t) \mid X_s = x\ \right] = \sum_{y\in\Lambda_N} f(y) K_{t-s}(x,y). \tag{2.2}$$

Subordinators are also Markov processes, but on $\mathbb{R}_+$, and can be thought of as stochastic time changes. They must therefore satisfy a few additional conditions, such as, if $T = (T_t)_{t\geq 0}$ is a subordinator, then $T_t \geq 0$ for each $t > 0$ almost surely (a.s.) and $T_s \leq T_t$ whenever $s \leq t$. Moreover, for the purpose of the present demonstration, only subordinators with an infinite lifetime will be considered, i.e., $\inf\{t \geq 0 : T_t = \infty\} = \infty$ a.s. The formal definition, as given by Bertoin in [3], follows.

DEFINITION 2.2. *Let* $T = (T_t)_{t\geq 0}$ *be a right-continuous increasing* $\mathcal{F}_t$*-adapted process started from* 0 *with values on* $[0, \infty]$, *where* ∞ *is an absorbing state.* T *is called a subordinator if it has independent and homogeneous increments on* $[0, \infty)$.

That is, the increment $T_{t+s} - T_t$ is independent of $\mathcal{F}_t$ and has the same distribution as T_s.

A remarkable probabilistic application of subordinators is due to Bochner [4], who proved the following.

THEOREM 2.3. *Let $X = (X_t)_{t\geq 0}$ be a Markov process and $T = (T_t)_{t\geq 0}$ be a subordinator. Then the stochastic process $Y = (Y_t)_{t\geq 0} = (X_{T_t})_{t\geq 0}$ is again a Markov process. Moreover, if X is stationary, then Y is stationary.*

A way to characterize subordinators is through their Laplace transform. A consequence of the independence and homogeneity of the increments is the following property:

$$\mathrm{E}\left[e^{\lambda T_{t+s}}\right] = \mathrm{E}\left[e^{\lambda T_t}\right]\mathrm{E}\left[e^{\lambda T_s}\right] \tag{2.3}$$

for every $s, t \geq 0$ and $\lambda \leq 0$. The Laplace transform can thus be expressed as

$$\mathrm{E}\left[e^{\lambda T_t}\right] = e^{-t\Psi(-\lambda)}, \tag{2.4}$$

where the function $\Psi : [0,\infty) \to [0,\infty)$ is called the Laplace exponent of the subordinator. A theorem due to de Finetti, Lévy, and Khintchine (see Bertoin [3]) gives a necessary and sufficient condition for a function to be the Laplace exponent of a subordinator. Moreover, one can establish a one-to-one correspondence between Laplace exponents and Bernstein functions [2]. We summarize these results in the following theorem.

THEOREM 2.4. *Let $\Phi \in C^\infty([0,\infty))$. Then, the following arguments are equivalent:*

1. *Φ is a Bernstein function.*
2. *Φ is the Laplace exponent of a subordinator.*
3. *There exists a semigroup $\{\rho_t\}_{t\geq 0}$ of positive measures on $[0,\infty)$ for which, for all $z \geq 0$,*

$$e^{-t\Phi(z)} = \int_0^\infty e^{-sz}\rho_t(ds). \tag{2.5}$$

The next theorem comes as a corollary of the previous two.

THEOREM 2.5. *Let $(\mathcal{K}_t)_{t\geq 0}$ be a Markov semigroup on $l^2(\Lambda_N)$ with kernel $K_t(x,y)$, and let Φ be a Bernstein function. Then, the semigroup of the subordinated process, denoted $(\mathcal{K}_t^\Phi)_{t\geq 0}$, is Markov and there exists a semigroup $\{\rho_t\}_{t\geq 0}$ of positive measures on $[0,\infty)$ such that for any $f \in l^2(\Lambda_N)$,*

$$\mathcal{K}_t^\Phi f(x) = \int_0^\infty \mathcal{K}_s f(x)\ \rho_t(ds). \tag{2.6}$$

This property extends to the transitional probability densities as follows:

$$K_t^\Phi(x,y) = \int_0^\infty K_s(x,y)\rho_t(ds). \tag{2.7}$$

The following definition leads to the explicit relation between q-Racah polynomials and Markov semigroups.

DEFINITION 2.6. *The operator $\mathcal{L}$ is the infinitesimal generator of a Markov semigroup on $l^2(\Lambda_N)$ if its kernel, $L(x,y)$ with $x, y \in \Lambda_N$, satisfies the following two conditions:*

(i) *nonnegativity:* $L(x,y) \geq 0$ *for all* $x \neq y \in \Lambda_N$;
(ii) *probability conservation:* $\sum_{y\in\Lambda_N} L(x,y) = 0$ *for all* $x \in \Lambda_N$.

Let us recall the elementary result from probability theory according to which a Markov semigroup $(\mathcal{K}_t)_{t\geq 0}$ on $l^2(\Lambda_N)$ can always be represented as follows in terms of the infinitesimal generator $\mathcal{L}$:

$$\mathcal{K}_t = \exp(t\mathcal{L}). \tag{2.8}$$

DEFINITION 2.7. *Let* Δ *and* ∇ *denote the difference operators on* $l^2(\Lambda_N)$ *defined as follows:*

$$\Delta f(x) = f(x+1) - f(x), \quad \nabla f(x) = f(x) - f(x-1), \tag{2.9}$$

with $f(N+1) = f(-1) = 0$.

The q-Racah polynomials introduced in Definition 1.1 satisfy the q-difference eigenvalue equation

$$\mathcal{L}R_n(\mu(x)) = \lambda_n R_n(\mu(x)), \tag{2.10}$$

where the operator $\mathcal{L}$ is given by

$$\mathcal{L} = -B(x)\Delta + D(x)\nabla, \tag{2.11}$$

with $B(x), D(x)$, and λ_n defined in Theorem 1.5 (see [7] and the references therein for a proof).

PROPOSITION 2.8. *If* $B(x) \leq 0$ *and* $D(x) \leq 0$, *then* $\mathcal{L}$ *is the infinitesimal generator of a Markov semigroup.*

Proof. Since the off-diagonal entries of the kernel $L(x,y)$ of $\mathcal{L}$ contain only $-B(x), -D(x)$, or 0, the nonnegativity condition in Definition 2.6 is satisfied by assumption. The second condition is also satisfied since

$$\sum_{y\in\Lambda_N} L(x,y) = -D(x) + (B(x) + D(x)) - B(x) = 0 \ \forall x = 1, \ldots, N-1,$$
$$\sum_{y\in\Lambda_N} L(0,y) = (B(0) + D(0)) - B(0) = D(0) = 0,$$
$$\sum_{y\in\Lambda_N} L(N,y) = -D(N) + (B(N) + D(N)) = B(N) = 0. \qquad \square$$

Note that one property of the infinitesimal generator of a Markov process is that it must satisfy the positive maximum principle [9],

$$\sup\{f(x) : x \in \Lambda_N\} = f(x_0) \geq 0 \Rightarrow (\mathcal{L}f)(x_0) \leq 0. \tag{2.12}$$

In particular, $\sup\{R_n(\mu(x)) : x \in \Lambda_N\} = R_n(\mu(x_0)) \geq 0$ for some $x_0 \in \Lambda_N$. Hence

$$\mathcal{L}R_n(\mu(x_0)) = \lambda_n R_n(\mu(x_0)) \leq 0, \tag{2.13}$$

which implies that the eigenvalues of the infinitesimal generator are negative or zero.

PROPOSITION 2.9. *The kernel of* $\mathcal{K}_t = \exp(t\mathcal{L})$ *has the following representation in terms of the q-Racah polynomials:*

$$K_t(x,y) = \sum_{n=0}^{N} \frac{e^{t\lambda_n}}{h_n} R_n(\mu(x)) R_n(\mu(y))\ w(y), \tag{2.14}$$

where $w(y) > 0$ is the invariant measure density.

Proof. The q-Racah polynomials form an orthogonal basis of the Hilbert space $l^2(\Lambda_N)$, with invariant measure density $w(y)$. From (2.10), $\mathcal{K}_t$ satisfies the eigenvalue equation

$$\mathcal{K}_t R_n(\mu(x)) = e^{t\lambda_n} R_n(\mu(x)). \tag{2.15}$$

The kernel can thus be written as an expansion in the q-Racah polynomials, and we recover (2.14) in the proposition. The invariant measure density $w(y)$ can be expressed recurrently as

$$w(0) = 1 \quad \text{and} \quad w(y+1) = \frac{B(y)}{D(y+1)}\, w(y) \quad \text{for } y = 0, 1, \ldots, N-1, \tag{2.16}$$

which proves its positivity since $B(y) < 0$ and $D(y+1) < 0$ for $y = 0, 1, \ldots, N-1$. ∎

We now proceed to the proof of our main result stated in the introduction. By Theorem 2.4, there exists a semigroup $\{\rho_t\}_{t\geq 0}$ of positive measures on $[0, \infty)$ associated with the Bernstein function Φ of Theorem 1.5. Theorem 2.5, together with Propositions 2.8 and 2.9, implies that the subordinated kernel takes the form

$$K_t^{\Phi}(x, y) = \sum_{n=0}^{N} \int_0^{\infty} e^{s\lambda_n} \rho_t(ds)\ \frac{1}{h_n} R_n(\mu(x)) R_n(\mu(y))\ w(y), \tag{2.17}$$

which gives (1.6) by definition of the measure $\rho_t(ds)$. Furthermore, the associated semigroup is again a Markov semigroup, so we proved Theorem 1.5.

Corollary 1.6 follows from the fact that the function

$$\Phi(z) = \frac{1}{|\ln q|} \ln \left[\frac{1 + \alpha\beta q + z}{2} + \sqrt{\left(\frac{1 + \alpha\beta q + z}{2} \right)^2 - \alpha\beta q} \right]$$

is a Bernstein function. One can indeed find an explicit expression for the semigroup $\{\rho_t\}_{t\geq 0}$ of positive measures that describes the subordinator. That is, for $\alpha\beta > 0$, there exists a one-parameter family of positive measures

$$\rho_t(ds) = \frac{t}{s|\ln q|}\ (\alpha\beta q)^{-\frac{t}{2|\ln q|}}\ e^{-(1+\alpha\beta q)s}\ I_{\frac{t}{|\ln q|}}\left(2\sqrt{\alpha\beta q}\ s \right) ds \tag{2.18}$$

that entirely characterizes Φ by Theorem 2.4 (see [5] or formula (8), page 314 in [8]). Notice that Φ is defined such that $\Phi(-\lambda_n) = n$ for $n = 1, 2, \ldots, N$ and is clearly non-negative for all $x \geq 0$. The eigenvalues λ_n being negative or zero implies in particular that $\alpha\beta q^2 \leq 1$. $\Phi(\lambda_0) = \Phi(0)$ thus depends on the value of $\alpha\beta q$ as follows:

$$\Phi(0) = \max\left(\frac{\ln(\alpha\beta q)}{|\ln q|}, 0 \right), \tag{2.19}$$

which leads to (1.9). Positivity of the Poisson kernel in (1.10) comes from (1.9) being Markovian and hence positive, $w(y) > 0$, and $\max(1 - (\alpha\beta q)^{-\frac{t}{|\ln q|}}, 0) \geq 0$.

In the limiting case $\alpha\beta = 0$, while the q-Racah polynomials converge either to the dual q-Hahn polynomials for $\alpha = 0$ or $\beta = 0$, or to the dual q-Krawtchouk polynomials

for $\alpha = \beta = 0$, the semigroup of positive measures consists of the well-known Gamma measures

$$\rho_t(ds) = \frac{s^{\frac{t}{|\ln q|}-1} e^{-s}}{\Gamma\left(\frac{t}{|\ln q|}\right)}\, ds. \tag{2.20}$$

Therefore, Φ is a Bernstein function for $\alpha\beta \geq 0$, which completes the proof of Corollary 1.6.

Remark 2.10. Corollary 1.6 unravels an interesting connection with free probability theory. Let $\tau = \frac{t}{|\ln q|}$ be the renormalized time of the process; then the semigroup of measure $\{\rho_\tau\}_{\tau \geq 0}$ reads

$$\rho_\tau(ds) = \frac{\tau}{s}\, (\alpha\beta q)^{-\frac{\tau}{2}}\, e^{-(1+\alpha\beta q)s}\, I_\tau\left(2\sqrt{\alpha\beta q}\, s\right) ds. \tag{2.21}$$

In the limit $q \to 0$, one recovers the Gamma measures

$$\rho_\tau(ds) = \frac{s^{\tau-1} e^{-s}}{\Gamma(\tau)}\, ds, \tag{2.22}$$

suggesting that the semigroup of Gamma measures defines a subordinator in the free case.

We conclude with the proof that Corollary 1.6 extends the result of positivity of the Poisson kernel given by Gasper and Rahman in [6].

PROPOSITION 2.11. *The set of conditions*

$$0 < q < 1, \quad \gamma = q^{-N-1}, \quad 0 < \alpha q < 1, \quad 0 \leq \beta q < 1, \quad 0 < \delta < \alpha q^N \tag{2.23}$$

is strictly stronger than

$$0 < q < 1, \quad 0 \leq \alpha\beta, \quad B(x) \leq 0, \quad D(x) \leq 0 \tag{2.24}$$

for all $x \in \Lambda_N$ and where $B(x)$ and $D(x)$ are defined as in Theorem 1.5.

Proof. Notice first that $0 < \alpha q < 1$ and $0 \leq \beta q < 1$ imply in particular $\alpha\beta \geq 0$. We show next that $B(x) \leq 0$ for all $x \in \Lambda_N$. With conditions (2.23), $B(x)$ becomes

$$\begin{aligned}
B(0) &= \frac{(1-\alpha q)(1-\beta\delta q)(1-q^{-N})}{1-\delta q^{1-N}},\\
B(x) &= \frac{(1-\alpha q^{x+1})(1-\beta\delta q^{x+1})(1-q^{x-N})(1-\delta q^{x-N})}{(1-\delta q^{2x-N})(1-\delta q^{2x+1-N})}, \quad x \in \{1,\ldots,N-1\},\\
B(N) &= 0.
\end{aligned}$$

For $x \in \{0,1,\ldots,N-1\}$ and since $q < 1$, we have

$$\begin{aligned}
&1 > 1-\alpha q^{x+1} > 1-q^x \geq 0,\\
&1 \geq 1-\beta\delta q^{x+1} > 1-\delta q^x > 1-q^{N+x-1} > 0,\\
&1 > 1-\delta q^{2x+1-N} > 1-\alpha q^{2x+1} > 1-q^{2x} \geq 0.
\end{aligned}$$

For $x \in \{1,\ldots,N-1\}$, we have

$$\begin{aligned}
&1 > 1-\delta q^{x-N} > 1-\alpha q^x > 1-q^{x-1} \geq 0,\\
&1 > 1-\delta q^{2x-N} > 1-\alpha q^{2x} > 1-q^{2x-1} \geq 0.
\end{aligned}$$

Since the only negative factor in $B(x)$ is $1 - q^{x-N}$ for all $x \in \{0, 1, \ldots, N-1\}$, $B(x) \leq 0$ for all $x \in \Lambda_N$. With conditions (2.23), $D(x)$ can be expressed as

$$\begin{aligned} D(0) &= 0, \\ D(x) &= \frac{q(1-q^x)(1-\delta q^x)(\beta - q^{x-N-1})(\alpha - \delta q^{x-N-1})}{(1-\delta q^{2x-N-1})(1-\delta q^{2x-N})}, \quad x \in \{1, \ldots, N-1\}, \\ D(N) &= \frac{q(1-q^N)(\beta - q^{-1})(\alpha - \delta q^{-1})}{1-\delta q^{N-1}}. \end{aligned}$$

For $x \in \{1, \ldots, N\}$, we have

$$\begin{aligned} &0 < \delta q^{-N} < \alpha \;\Rightarrow\; \delta q^{x-N-1} < \alpha \;\Rightarrow\; \alpha - \delta q^{x-N-1} > 0, \\ &0 \leq \beta < q^{-1} < q^{x-N-1} \;\Rightarrow\; \beta - q^{x-N-1} < 0 \end{aligned}$$

and, similarly as before,

$$\begin{aligned} &1 > 1 - \delta q^x > 1 - \alpha q^{x+N} > 1 - q^{x+N-1} > 0, \\ &1 > 1 - \delta q^{2x-N} > 1 - \alpha q^{2x} > 1 - q^{2x-1} > 0, \\ &1 > 1 - \delta q^{2x-N-1} > 1 - \alpha q^{2x-1} > 1 - q^{2x-2} \geq 0. \end{aligned}$$

Hence $D(x) \leq 0$ since the only negative factor is $\beta - q^{x-N-1}$. The condition $\alpha q < 1$ implies furthermore that the denominators of $B(x)$ and $D(x)$, respectively, are never zero. Finally, let

$$\alpha = \beta = -\delta = 1, \quad \gamma = q^{-N-1}.$$

The conditions in (2.24) are clearly satisfied, whereas $\delta > 0$ in (2.23) fails, so the inclusion is strict. □

REFERENCES

[1] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or $6-j$ symbols*, SIAM J. Math. Anal., 10 (1979), pp. 1008–1016.

[2] C. Berg and G. Frost, *Potential Theory on Locally Compact Abelian Groups*, Springer-Verlag, Berlin, 1975.

[3] J. Bertoin, *Subordinators: Examples and applications*, in Lectures on Probability Theory and Statistics (Saint-Flour, 1997), Springer, Berlin, 1999, pp. 1–91.

[4] S. Bochner, *Harmonic Analysis and the Theory of Probability*, University of California Press, Berkeley, 1955.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed., Wiley, New York, 1971.

[6] G. Gasper and M. Rahman, *Product formulas of Watson, Bailey and Bateman types and positivity of the Poisson kernel for q-Racah polynomials*, SIAM J. Math. Anal., 15 (1984), pp. 768–789.

[7] R. Koekoek and R. Swarttouw, *The Askey Scheme of Hypergeometric Orthogonal Polynomials and Its q-analogue*, Tech. report 98-17, Department of Technical Mathematics and Informatics, Delft University of Technology, Delft, The Netherlands, 1998.

[8] A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series, Volume* 4: *Direct Laplace Transforms*, Gordon and Breach, New York, 1992.

[9] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, Springer-Verlag, New York, 1994.

[10] G. Szego, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Coll. Publ. 23, AMS, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 985–1011

STABILITY OF SOLITARY WAVES OF A GENERALIZED OSTROVSKY EQUATION*

STEVE LEVANDOSKY† AND YUE LIU‡

Abstract. Considered herein is the stability problem of solitary wave solutions of a generalized Ostrovsky equation, which is a modification of the Korteweg–de Vries equation widely used to describe the effect of rotation on surface and internal solitary waves or capillary waves.

Key words. solitary waves, stability, variational methods

AMS subject classifications. 35Q53, 76B25, 58E30, 58F10

DOI. 10.1137/050638722

1. Introduction.

The nonlinear dispersive equation

$$\left(u_t - \beta u_{xxx} + (u^2)_x\right)_x = \gamma u, \quad x \in \mathbf{R}, \tag{1.1}$$

was derived by Ostrovsky [20] in dimensionless space-time variables (x,t) as a model for the unidirectional propagation of weakly nonlinear long surface and internal waves of small amplitude in a rotating fluid. The liquid is assumed to be incompressible and inviscid. Here $u(t,x)$ represents the free surface of the liquid and the parameter $\gamma > 0$ measures the effect of rotation. The parameter β determines the type of dispersion, namely, $\beta < 0$ (negative dispersion) for surface and internal waves in the ocean or surface waves in a shallow channel with an uneven bottom and $\beta > 0$ (positive dispersion) for capillary waves on the surface of liquid or for oblique magneto-acoustic waves in plasma. See Benilov [2], Grimshaw [9], Galkin and Stepanyants [7], Gilman, Grimshaw, and Stepanyants [8], and Ostrovsky and Stepanyants [21].

Considered herein is the generalization of the Ostrovsky equation

$$(u_t - \beta u_{xxx} + f(u)_x)_x = \gamma u, \qquad x \in \mathbf{R}, \tag{1.2}$$

where f is a C^2 function which is homogeneous of degree $p \geq 2$, in the sense that it satisfies $sf'(s) = pf(s)$. This includes, for instance, nonlinearities of the form $f(u) = \pm|u|^p$ and $\pm|u|^{p-1}u$. Certain equations of this class have a direct relation to physical systems. In particular, when $p = 3$, (1.2) describes the propagation of internal waves of even modes, which possess a cubic nonlinearity, in the ocean. See Galkin and Stepanyants [7], Leonov [13], and Shrira [22, 23].

In this paper, we investigate the stability of solitary wave solutions of (1.2). Using variational methods, we prove the existence of solitary waves (Theorem 2.1). Solitary waves thus obtained are called *ground states*, and the set of all ground states is denoted by $G(\beta, c, \gamma)$. The variational characterization of the ground states permits us to consider the limiting behavior of the solitary waves as the rotation parameter γ

*Received by the editors August 23, 2005; accepted for publication (in revised form) May 22, 2006; published electronically November 3, 2006.
http://www.siam.org/journals/sima/38-3/63872.html
†Mathematics and Computer Science Department, College of the Holy Cross, Worcester, MA 01610 (spl@mathcs.holycross.edu).
‡Department of Mathematics, The University of Texas at Arlington, Arlington, TX 76019 (yliu@uta.edu).

vanishes, and we show that the ground state solitary waves converge to solitary waves of the Korteweg–de Vries (KdV) equation (Theorem 2.5).

The stability analysis makes use of the conserved quantities

$$E(u) = \int_{\mathbf{R}} \frac{\beta}{2} u_x^2 + \frac{\gamma}{2} |D_x^{-1} u|^2 + F(u)\, dx$$

and

$$V(u) = \frac{1}{2} \int_{\mathbf{R}} u^2\, dx,$$

where $F' = f$ and $F(0) = 0$, and the operator D_x^{-1} is defined via the Fourier transform as

$$\widehat{D_x^{-1} f} = (-i\xi)^{-1} \hat{f}(\xi).$$

It was shown by Liu and Varlamov [18] that the classical Ostrovsky equation (1.1) is well-posed in the space

$$X_s = \{ f \in H^s(\mathbf{R}) \mid D_x^{-1} f \in H^s(\mathbf{R}) \}$$

with norm

$$\|f\|_{X_s} = \|f\|_s + \|D_x^{-1} f\|_s$$

for $s > 3/2$. The methods therein also imply the same result for the generalized Ostrovsky equation (1.2). We therefore make the following definition.

Definition 1.1. *A set $S \subset X$ is X-stable with respect to* (1.2) *if for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $u_0 \in X \cap X_s$, $s > 3/2$, with*

$$\inf_{v \in S} \|u_0 - v\|_X < \delta, \tag{1.3}$$

the solution $u(t)$ of (1.2) *with initial value $u(0) = u_0$ can be extended to a solution in the space $C([0, \infty), X \cap X_s)$ and satisfies*

$$\inf_{v \in S} \|u(t) - v\|_X < \epsilon \tag{1.4}$$

for all $t \geq 0$. Otherwise we say that S is X-unstable.

Our main results apply to the set $G(\beta, c, \gamma)$, defined by (2.8). For each $y \in \mathbf{R}$, we define the translation operator by $\tau_y v = v(\cdot + y)$. Given a ground state φ in $G(\beta, c, \gamma)$, the *orbit* of φ is the set $\mathcal{O}_\varphi = \{\tau_y \varphi \mid y \in \mathbf{R}\}$. We show in Theorems 3.1 and 4.2 that the function d defined by (3.1) determines the stability or instability of the solitary waves in the sense that if $d''(c) > 0$, then $G(\beta, c, \gamma)$ is X_1-stable and if $d''(c) < 0$, then $\mathcal{O}_\varphi$ is X_1-unstable. Although these results are not quite complementary, the only difference is due to the possible nonuniqueness of ground states up to translation. That is, if ground states are unique up to translation, then $G(\beta, c, \gamma) = \mathcal{O}_\varphi$.

One difficulty in applying these results is the fact that an explicit formula for d is not available. It is also not known if $d(c)$ is twice differentiable. To remedy this, we also prove a second result, Theorem 4.3, which provides sufficient conditions for instability directly in terms of the parameters β, c, γ, and p. The result is based on the work of Gonçalves Ribeiro [10]. Another approach to dealing with the lack of information about $d(c)$ is to compute it numerically. We conclude the paper with some numerical calculations of d'' which approximately determine regions of stability and instability in terms of the parameters.

Notation. The norm in the classical Sobolev spaces $H^s(\mathbf{R})$ will be written $\|\cdot\|_s$. For $1 \leq q \leq \infty$, the norm in $L^q(\mathbf{R})$ will be written $|\cdot|_q$.

2. Solitary waves. Solitary-wave solutions of the form $u(x,t) = \varphi(x-ct)$ satisfy the stationary equation

$$\beta\varphi_{xx} + c\varphi + \gamma D_x^{-2}\varphi = f(\varphi). \tag{2.1}$$

We will prove existence of solitary waves in the space X_1 by considering the following variational problem. Define the functionals

$$I(u) = I(u;\beta,c,\gamma) = \int_{\mathbf{R}} \beta u_x^2 - cu^2 + \gamma(D_x^{-1}u)^2\,dx \tag{2.2}$$

and

$$K(u) = -(p+1)\int_{\mathbf{R}} F(u)\,dx, \tag{2.3}$$

where F satisfies $F' = f$ and $F(0) = 0$. Then if $\psi \in X_1$ achieves the minimum

$$M_\lambda = \inf\{I(u) \mid u \in X_1, K(u) = \lambda\} \tag{2.4}$$

for some $\lambda > 0$, then there exists a Lagrange multiplier μ such that

$$\beta\psi_{xx} + c\psi + \gamma D_x^{-2}\psi = \mu f(\psi). \tag{2.5}$$

Hence $\varphi = \mu^{\frac{1}{p-1}}\psi$ satisfies (2.1). We call such solutions *ground state* solutions and denote the set of all ground state solutions by $G(\beta,c,\gamma)$. By the homogeneity of I and K, ground states also achieve the minimum

$$m = m(\beta,c,\gamma) = \inf\left\{\frac{I(u)}{(K(u))^{\frac{2}{p+1}}} : u \in X_1, K(u) > 0\right\}, \tag{2.6}$$

and it follows that

$$M_\lambda = m\lambda^{\frac{2}{p+1}}. \tag{2.7}$$

We next note that the properties $sf'(s) = pf(s)$ and $F' = f$ imply that $sf(s) = (p+1)F(s)$, so that

$$K(u) = -\int_{\mathbf{R}} uf(u)\,dx,$$

and therefore multiplying (2.1) by φ and integrating yields $I(\varphi) = K(\varphi)$. Thus we may characterize the set of ground state solutions $G(\beta,c,\gamma)$ as

$$G(\beta,c,\gamma) = \left\{\varphi \in X_1 \mid K(\varphi) = I(\varphi;\beta,c,\gamma) = (m(\beta,c,\gamma))^{\frac{p+1}{p-1}}\right\}. \tag{2.8}$$

We now seek to prove that this set is nonempty. We say that a sequence ψ_k is a *minimizing sequence* if for some $\lambda > 0$,

$$\lim_{k\to\infty} K(\psi_k) = \lambda \quad \text{and} \quad \lim_{k\to\infty} I(\psi_k) = M_\lambda.$$

THEOREM 2.1. *Let $\beta > 0$, $\gamma > 0$, and $c < 2\sqrt{\beta\gamma}$. Let ψ_k be a minimizing sequence for some $\lambda > 0$. Then there exist a subsequence (renamed ψ_k) and scalars*

$y_k \in \mathbf{R}$ and $\psi \in X_1$ such that $\psi_k(\cdot + y_k) \to \psi$ in X_1. The function ψ achieves the minimum $I(\psi) = M_\lambda$ subject to the constraint $K(\psi) = \lambda$.

Proof. The result is an application of the concentration compactness lemma of Lions [16]. We outline the proof here. First, observe that by (2.7) the strict subadditivity condition

$$M_\alpha + M_{\lambda-\alpha} > M_\lambda \tag{2.9}$$

holds for any $\alpha \in (0, \lambda)$. Next, since $\beta > 0$, $\gamma > 0$, and $c < 2\sqrt{\beta\gamma}$, the functional I satisfies the coercivity condition

$$I(u) \geq A \int_{\mathbf{R}} u_x^2 + (D_x^{-1}u)^2 \, dx = A\|u\|_{X_1}^2,$$

where

$$A = \left\{ \begin{array}{ccc} \frac{4\beta\gamma - c^2}{2(\beta+\gamma+\sqrt{(\beta-\gamma)^2+c^2})} & \text{for} & 0 < c < 2\sqrt{\beta\gamma} \\ \min\{\beta, \gamma\} & \text{for} & c \leq 0 \end{array} \right\} > 0. \tag{2.10}$$

It is also clear that $I(u) \leq C\|u\|_{X_1}^2$ for some constant C, so $I(u)^{1/2}$ is equivalent to the norm on X_1. Now let ψ_k be a minimizing sequence. Then by coercivity of I, the sequence ψ_k is bounded in X_1, so if we define

$$\rho_k = |D_x\psi_k|^2 + |D_x^{-1}\psi_k|^2,$$

then after extracting a subsequence, we may assume

$$\lim_{k\to\infty} \int_{\mathbf{R}} \rho_k \, dx = L > 0.$$

We may assume further after normalizing that $\int_{\mathbf{R}} \rho_k \, dx = L$ for all k. By the concentration compactness lemma, a further subsequence ρ_k satisfies one of the following three conditions.

1. Vanishing: For every $R > 0$,

$$\lim_{k\to\infty} \sup_{y\in\mathbf{R}} \int_{B(y,R)} \rho_k \, dx = 0.$$

2. Dichotomy: There exists some $l \in (0, L)$ such that for any $\epsilon > 0$ there exist $R > 0$ and $R_k \to \infty$, $y_k \in \mathbf{R}$ and k_0 such that

$$\left| \int_{B(y_k,R)} \rho_k \, dx - l \right| < \epsilon \quad \text{and} \quad \left| \int_{R<|x-y_k|<R_k} \rho_k \, dx \right| < \epsilon$$

for $k \geq k_0$.

3. Compactness: There exists $y_k \in \mathbf{R}$ such that for any $\epsilon > 0$ there exists $R(\epsilon)$ such that

$$\int_{B(y_k,R(\epsilon))} \rho_k \, dx \geq \int_{\mathbf{R}} \rho_k \, dx - \epsilon$$

for all k.

In the same manner as in [14], it follows from the coercivity of I, the Sobolev inequality, and the subadditivity condition (2.9) that both vanishing and dichotomy may be ruled out, and therefore the sequence ρ_k is compact. Now set $\varphi_k(x) = \psi_k(x + y_k)$. Since φ_k is bounded in X_1, a subsequence φ_k converges weakly to some $\psi \in X_1$, and by the weak lower semicontinuity of I over X_1, we have

$$I(\psi) \le \lim_{k\to\infty} I(\varphi_k) = M_\lambda.$$

Furthermore, weak convergence in X_1, compactness of ρ_k, and the Sobolev inequality imply strong convergence of φ_k to ψ in L^{p+1}. Therefore

$$K(\psi) = \lim_{k\to\infty} K(\varphi_k) = \lambda,$$

so $I(\psi) \ge M_\lambda$. Together with the inequality above, this implies $I(\psi) = M_\lambda$, so ψ is a minimizer of I subject to the constraint $K(\varphi) = \lambda$. Finally, since I is equivalent to the norm on X_1, $\phi_k \rightharpoonup \psi$, and $I(\phi_k) \to I(\psi)$, it follows that ϕ_k converges strongly to ψ in X_1. □

At this time it is unknown whether or not the ground states are unique up to translation. Uniqueness would imply that if $\varphi \in G(\beta, c, \gamma)$ is any ground state, then $G(\beta, c, \gamma) = \mathcal{O}_\varphi$, in which case the stability and instability theorems (Theorems 3.1 and 4.2) are complementary. We suspect that the ground states are unique, at least in the case $c < -2\sqrt{\beta\gamma}$, when the ground states have nonoscillatory tails.

The function $m(\beta, c, \gamma)$ defined above plays an important role in our later results, so we now will investigate some of its properties. The first is a simple scaling identity.

LEMMA 2.2. *Let $\beta > 0$, $\gamma > 0$, and $c < 2\sqrt{\beta\gamma}$. For any $r > 0$ and $s > 0$ we have*

$$m(rs^2\beta, rc, rs^{-2}\gamma) = rs^{\frac{p-1}{p+1}} m(\beta, c, \gamma).$$

Proof. Let $u \in X_1$ with $K(u) \ne 0$. For any $r > 0$ we have

$$I(u; r\beta, rc, r\gamma) = rI(u; \beta, c, \gamma),$$

so $m(r\beta, rc, r\gamma) = rm(\beta, c, \gamma)$. Next let $v(x) = u(sx)$ for $s > 0$. Then

$$I(v; \beta, c, \gamma) = \frac{1}{s} I(u; s^2\beta, c, s^{-2}\gamma), \qquad K(v) = \frac{1}{s} K(u),$$

so

$$\frac{I(v; \beta, c, \gamma)}{K(v)^{\frac{2}{p+1}}} = s^{\frac{1-p}{p+1}} \frac{I(u; s^2\beta, c, s^{-2}\gamma)}{K(u)^{\frac{2}{p+1}}},$$

and consequently

$$m(s^2\beta, c, s^{-2}\gamma) = s^{\frac{p-1}{p+1}} m(\beta, c, \gamma). \qquad \square$$

Next, we show that m is continuous and monotone in each of its variables.

LEMMA 2.3. *The function m is continuous on the domain $\beta > 0$, $\gamma > 0$, $c < 2\sqrt{\gamma\beta}$. Furthermore, m is strictly increasing in γ and β and strictly decreasing in c.*

Proof. First, fix $\beta > 0$ and $\gamma > 0$ and consider $c_1 < c_2 < 2\sqrt{\beta\gamma}$. Let φ_{c_1} and φ_{c_2} be ground states with $c = c_1$ and $c = c_2$, respectively. Then

$$\begin{aligned} m(\beta, c_2, \gamma) &\leq \frac{I(\varphi_{c_1}; \beta, c_2, \gamma)}{K(\varphi_{c_1})^{\frac{2}{p+1}}} \\ &= \frac{I(\varphi_{c_1}; \beta, c_1, \gamma) + (c_1 - c_2)\int \varphi_{c_1}^2\, dx}{K(\varphi_{c_1})^{\frac{2}{p+1}}} \\ &= \frac{I(\varphi_{c_1}; \beta, c_1, \gamma)}{K(\varphi_{c_1})^{\frac{2}{p+1}}} + (c_1 - c_2)\frac{\int \varphi_{c_1}^2\, dx}{K(\varphi_{c_1})^{\frac{2}{p+1}}} \\ &= m(\beta, c_1, \gamma) + (c_1 - c_2)\frac{\int \varphi_{c_1}^2\, dx}{K(\varphi_{c_1})^{\frac{2}{p+1}}} \\ &< m(\beta, c_1, \gamma), \end{aligned}$$

so m is strictly decreasing in c. On the other hand,

$$\begin{aligned} m(\beta, c_1, \gamma) &\leq \frac{I(\varphi_{c_2}; \beta, c_1, \gamma)}{K(\varphi_{c_2})^{\frac{2}{p+1}}} \\ &= \frac{I(\varphi_{c_2}; \beta, c_2, \gamma) + (c_2 - c_1)\int \varphi_{c_2}^2\, dx}{K(\varphi_{c_2})^{\frac{2}{p+1}}} \\ &= m(\beta, c_2, \gamma) + (c_2 - c_1)\frac{\int \varphi_{c_2}^2\, dx}{K(\varphi_{c_2})^{\frac{2}{p+1}}}, \end{aligned}$$

so

$$0 \leq m(\beta, c_1, \gamma) - m(\beta, c_2, \gamma) \leq (c_2 - c_1)\frac{\int \varphi_{c_2}^2\, dx}{K(\varphi_{c_2})^{\frac{2}{p+1}}}.$$

Now since

$$I(\varphi_{c_2}; \beta, c_2, \gamma) \geq A\int \varphi_{c_2}^2\, dx,$$

where A is defined by (2.10), it follows that

$$|m(\beta, c_2, \gamma) - m(\beta, c_1, \gamma)| \leq A^{-1} m(\beta, c_2, \gamma)(c_2 - c_1),$$

so m is locally Lipschitz continuous in c. By similar reasoning it follows that m is increasing and locally Lipschitz in β and γ. ∎

We now consider the effect of letting the rotation parameter γ approach zero. Formally, this results in the generalized KdV equation

$$\beta\varphi_{xx} + c\varphi = f(\varphi). \tag{2.11}$$

For $c < 0$, ground state solutions of (2.11) achieve the minimum

$$m(\beta, c, 0) = \inf\left\{ \frac{I(u; \beta, c, 0)}{K(u)^{\frac{2}{p+1}}} : u \in H^1 K(u) > 0 \right\},$$

where

$$I(u; \beta, c, 0) = \int \beta u_x^2 - cu^2\, dx,$$

and K is defined as before by (2.3). Thus the set of ground states may be characterized as

$$G(\beta,c,\gamma)=\left\{\varphi\in H^1 \mid K(\varphi)=I(\varphi;\beta,c,0)=(m(\beta,c,0))^{\frac{p+1}{p-1}}\right\}.$$

Moreover, it is well known that the ground states are unique up to translation, so that $G(\beta,c,0)=\{\varphi_0(\cdot-y)\mid y\in\mathbf{R}\}$ for some φ_0. For example, in the case of the nonlinearity $f(u)=(-u)^p$, where $p\geq 2$ is an integer, we have the explicit formula

$$\varphi_0(x)=-\left(\frac{-c(p+1)}{2}\right)^{\frac{1}{p-1}}\operatorname{sech}^{\frac{2}{p-1}}\left(\frac{p-1}{2}\sqrt{\frac{-c}{\beta}}x\right).$$

Therefore the analogue of Theorem 2.1 takes the following form.

THEOREM 2.4. *Let $\beta>0$ and $c<0$. Let ψ_k be a sequence in H^1 such that*

$$\lim_{k\to\infty} I(\psi_k;\beta,c,0)=\lim_{k\to\infty} K(\psi_k)=(m(\beta,c,0))^{\frac{p+1}{p-1}}.$$

Then there exists a subsequence (renamed ψ_k), and there exist scalars $y_k\in\mathbf{R}$ such that $\psi_k(\cdot+y_k)\to\varphi_0$ in H^1.

THEOREM 2.5. *Fix $\beta>0$ and $c<0$ and consider any sequence $\gamma_k\to 0^+$. Denote by φ_k any element of $G(\beta,c,\gamma_k)$. Then there exists a subsequence (renamed γ_k) and translations y_k so that*

$$\varphi_k(\cdot+y_k)\to\varphi_0 \tag{2.12}$$

in H^1, as $\gamma_k\to 0^+$. That is, the generalized KdV solitary waves are the limits in H^1 of solitary waves of the generalized Ostrovsky equation.

To prove this theorem, we will show that the sequence of Ostrovsky solitary waves is a minimizing sequence for the KdV variational problem. The following lemma is proved in [15].

LEMMA 2.6. *The space X_1 is dense in H^1.*

Using the lemma, we next prove that the function m is continuous at $\gamma=0$.

LEMMA 2.7. *Fix $\beta>0$ and $c<0$. Then*

$$\lim_{\gamma\to 0^+} m(\beta,c,\gamma)=m(\beta,c,0). \tag{2.13}$$

Proof. Since m is strictly increasing in γ, it suffices to show that $m(\beta,c,\gamma_k)\to m(\beta,c,0)$ for some sequence $\gamma_k\to 0$. By the density of X_1 in H^1 we may choose ψ_k in X_1 such that $\|\psi_k-\varphi_0\|_{H^1}<\frac{1}{k}$ and define

$$\gamma_k=\min\left(\frac{1}{k},\frac{1}{k\int|D_x^{-1}\psi_k|^2\,dx}\right).$$

Then

$$\begin{aligned} m(\beta,c,\gamma_k)&\leq\frac{I(\psi_k;\beta,c,\gamma_k)}{K(\psi_k)^{\frac{2}{p+1}}}\\ &=\frac{I(\psi_k;\beta,c,0)+\gamma_k\int|D_x^{-1}\psi_k|^2\,dx}{K(\psi_k)^{\frac{2}{p+1}}}\\ &\leq\frac{I(\psi_k;\beta,c,0)+\frac{1}{k}}{K(\psi_k)^{\frac{2}{p+1}}}. \end{aligned}$$

Since $I(\cdot;\beta,c,0)$ and K are both continuous on H^1, we therefore have

$$\limsup_{k\to\infty} m(\beta,c,\gamma_k) \le \frac{I(\varphi_0;\beta,c,0)}{K(\varphi_0)^{\frac{2}{p+1}}} = m(\beta,c,0).$$

On the other hand, given $\varphi_k \in G(\beta,c,\gamma_k)$ we have $\varphi_k \in H^1$, so

$$\begin{aligned} m(\beta,c,0) &\le \frac{I(\varphi_k;\beta,c,0)}{K(\varphi_k)^{\frac{2}{p+1}}} \\ &= \frac{I(\varphi_k;\beta,c,\gamma_k) - \gamma_k \int |D_x^{-1}\varphi_k|^2\,dx}{K(\varphi_k)^{\frac{2}{p+1}}} \\ &< \frac{I(\varphi_k;\beta,c,\gamma_k)}{K(\varphi_k)^{\frac{2}{p+1}}} = m(\beta,c,\gamma_k). \end{aligned}$$

Thus

$$\liminf_{k\to\infty} m(\beta,c,\gamma_k) \ge m(\beta,c,0),$$

and the lemma follows. □

Proof of Theorem 2.5. By continuity of m at $\gamma = 0$, we have

$$\lim_{k\to\infty} K(\varphi_k) = \lim_{k\to\infty} m(\beta,c,\gamma_k)^{\frac{p+1}{p-1}} = m(\beta,c,0)^{\frac{p+1}{p-1}}$$

and

$$\begin{aligned} \limsup_{k\to\infty} I(\varphi_k;\beta,c,0) &= \limsup_{k\to\infty} I(\varphi_k;\beta,c,\gamma_k) - \gamma_k \int (D_x^{-1}\varphi_k)^2\,dx \\ &\le \lim_{k\to\infty} I(\varphi_k;\beta,c,\gamma_k) \\ &= \lim_{k\to\infty} m(\beta,c,\gamma_k)^{\frac{p+1}{p-1}} \\ &= m(\beta,c,0)^{\frac{p+1}{p-1}}. \end{aligned}$$

Thus φ_k is a minimizing sequence for the KdV variational problem, and the result follows from Theorem 2.4. □

3. Stability. The main result of this section is the following.

THEOREM 3.1. *Let $\beta > 0$, $\gamma > 0$, and $c < 2\sqrt{\beta\gamma}$. Let d be defined as in* (3.1). *If $d''(c) > 0$, then the set of ground states $G(\beta,c,\gamma)$ is X_1-stable.*

The proof is based on arguments in [14], which makes use of the method of [5]. We remark here that the condition $d''(c) > 0$ may be replaced with strict convexity of d in a neighborhood of c. See [24].

Given $\beta > 0$, $\gamma > 0$, and $c < 2\sqrt{\beta\gamma}$, we define

$$d(c) = d(\beta,c,\gamma) = E(\varphi) - cV(\varphi), \tag{3.1}$$

where φ is any element of $G(\beta,c,\gamma)$. Since

$$E(u) - cV(u) = \frac{1}{2}I(u) - \frac{1}{p+1}K(u), \tag{3.2}$$

it follows from (2.8) that

$$d(\beta,c,\gamma)=\frac{p-1}{2(p+1)}I(\varphi)=\frac{p-1}{2(p+1)}K(\varphi)=\frac{p-1}{2(p+1)}\left(m(\beta,c,\gamma)\right)^{\frac{p+1}{p-1}}. \tag{3.3}$$

Therefore d is well defined, and we may deduce its properties by examining the function $m(\beta,c,\gamma)$. The following two lemmas are immediate corollaries of Lemmas 2.2 and 2.3.

LEMMA 3.2. *Let $\beta>0$, $\gamma>0$, and $c<2\sqrt{\beta\gamma}$. For any $r>0$ and $s>0$, we have*

$$d(rs^2\beta,rc,rs^{-2}\gamma)=r^{\frac{p+1}{p-1}}sd(\beta,c,\gamma).$$

LEMMA 3.3. *The function d is continuous on the domain $\beta>0$, $\gamma>0$, $c<2\sqrt{\gamma\beta}$. Furthermore, d is strictly increasing in γ and β and strictly decreasing in c.*

LEMMA 3.4. *For each fixed $\beta>0$ and $\gamma>0$, the partial derivative $\partial d/\partial c(\beta,c,\gamma)$ exists for all but countably many $c<2\sqrt{\beta\gamma}$. Similarly, $\partial d/\partial\beta$ and $\partial d/\partial\gamma$ exist for all but countably many β and γ, respectively. At points where the partials exist,*

$$\begin{aligned}\frac{\partial d}{\partial\beta}&=\frac{1}{2}\int(\varphi_x)^2\,dx,\\ \frac{\partial d}{\partial c}&=-\frac{1}{2}\int\varphi^2\,dx,\\ \frac{\partial d}{\partial\gamma}&=\frac{1}{2}\int(D_x^{-1}\varphi)^2\,dx.\end{aligned}$$

Proof. Since d is continuous and monotone with respect to each variable, it follows that the partial derivatives exist at all but countably many points. To verify the formulas above, first fix $\beta>0$ and $\gamma>0$. Then by the inequalities in the proof of Lemma 2.3,

$$-\frac{\int\varphi_{c_2}^2\,dx}{K(\varphi_{c_2})^{\frac{2}{p+1}}}\le\frac{m(\beta,c_2,\gamma)-m(\beta,c_1,\gamma)}{c_2-c_1}\le-\frac{\int\varphi_{c_1}^2\,dx}{K(\varphi_{c_1})^{\frac{2}{p+1}}}$$

for $c_1<c_2<2\sqrt{\beta\gamma}$. Let

$$\begin{aligned}g_s(\beta,c,\gamma)&=\sup\left\{\int\varphi_c^2\,dx:\ \varphi_c\in G(\beta,c,\gamma)\right\},\\ g_i(\beta,c,\gamma)&=\inf\left\{\int\varphi_c^2\,dx:\ \varphi_c\in G(\beta,c,\gamma)\right\}.\end{aligned}$$

Then, for $c_1<c_2<2\sqrt{\beta\gamma}$,

$$-\frac{g_i(\beta,c_2,\gamma)}{m(\beta,c_2,\gamma)^{\frac{2}{p-1}}}\le\frac{m(\beta,c_2,\gamma)-m(\beta,c_1,\gamma)}{c_2-c_1}\le-\frac{g_s(\beta,c_1,\gamma)}{m(\beta,c_1,\gamma)^{\frac{2}{p-1}}}.$$

We now claim that

$$\lim_{c\to c_0}\sup g_i(\beta,c,\gamma)\le g_s(\beta,c_0,\gamma).$$

To see this, choose any $c_k\to c_0$ and $\varphi_k\in G(\beta,c_k,\gamma)$. The continuity of m, the characterization (2.8), and the relation (3.3) imply that $I(\varphi_k)\to\frac{2(p+1)}{p-1}d(\beta,c_0,\gamma)$

and $K(\varphi_k) \to \frac{2(p+1)}{p-1} d(\beta, c_0, \gamma)$. Therefore φ_k is a minimizing sequence, so by Theorem 2.1, there is a translated subsequence (renamed φ_k) which converges in X_1 to some function φ in $G(\beta, c_0, \gamma)$. Hence

$$\limsup_{k\to\infty} g_i(\beta, c_k, \gamma) \le \int \varphi^2 \, dx \le g_s(\beta, c_0, \gamma).$$

Consequently

$$\frac{\partial m}{\partial c}(\beta, c^+, \gamma) = -\frac{g_s(\beta, c, \gamma)}{m(\beta, c, \gamma)^{\frac{2}{p-1}}}.$$

Now, since $d = \frac{p-1}{2(p+1)} m^{\frac{p+1}{p-1}}$, this implies

$$\frac{\partial d}{\partial c}(\beta, c^+, \gamma) = -\frac{1}{2} g_s(\beta, c, \gamma).$$

Likewise,

$$\frac{\partial d}{\partial c}(\beta, c^-, \gamma) = -\frac{1}{2} g_i(\beta, c, \gamma).$$

So at points where the partial derivative exists, we must have $g_s(\beta, c, \gamma) = g_i(\beta, c, \gamma)$, and the first formula above follows. The proofs of the other formulas are similar. ∎

We note here that the preceding proof illustrates that uniqueness of ground states up to translation would imply differentiability of d. For if $G(\beta, c, \gamma)$ consists of translates of a single ground state, then $g_s(\beta, c, \gamma) = g_i(\beta, c, \gamma)$, from which the differentiability of d follows.

For the remainder of this section we will regard d only as a function of c for fixed β and γ. So notation such as d', d'', or d^{-1} should be interpreted with respect to the variable c. A key role in the stability analysis is played by the ϵ-neighborhood of the set of ground states defined by

$$U_{c,\epsilon} = \left\{ u \in X_1 \mid \inf_{\varphi \in G(\beta,c,\gamma)} \|u - \varphi\|_{X_1} < \epsilon \right\}.$$

Since d is strictly decreasing in c, we may define

$$c(u) = d^{-1}\left(\frac{p-1}{2(p+1)} K(u) \right).$$

This associates a speed $c(u)$ with any function $u \in X_1$. The following lemma provides the key estimate involving these speeds.

LEMMA 3.5. *If $d''(c) > 0$, then there is some $\epsilon > 0$ such that for any $u \in U_{c,\epsilon}$ and $\varphi \in G(\beta, c, \gamma)$, we have*

$$E(u) - E(\varphi) - c(u)(V(u) - V(\varphi)) \ge \frac{1}{4} d''(c) |c(u) - c|^2.$$

Proof. Since $d'(c) = -V(\varphi_c)$, it follows from Taylor's theorem that

$$d(c_1) = d(c) - V(\varphi_c)(c_1 - c) + \frac{1}{2} d''(c)(c_1 - c)^2 + o\left(|c_1 - c|^2\right)$$

for c_1 near c. By choosing ϵ sufficiently small, it then follows that

$$\begin{aligned} d(c(u)) &\geq d(c) - V(\varphi_c)(c(u)-c) + \frac{1}{4}d''(c)(c(u)-c)^2 \\ &= E(\varphi_c) - c(u)V(\varphi_c) + \frac{1}{4}d''(c)(c(u)-c)^2 \end{aligned}$$

for $u \in U_{c,\epsilon}$. Now if $\varphi_{c(u)} \in G(\beta, c(u), \gamma)$, then

$$K(\varphi_{c(u)}) = \frac{2(p+1)}{p-1}d(c(u)) = K(u),$$

and $\varphi_{c(u)}$ minimizes $I(\cdot;\beta,c(u),\gamma)$ subject to this constraint, so

$$\begin{aligned} E(u) - c(u)V(u) &= \frac{1}{2}I(u;\beta,c(u),\gamma) - \frac{1}{p+1}K(u) \\ &\geq \frac{1}{2}I(\varphi_{c(u)};\beta,c(u),\gamma) - \frac{1}{p+1}K(\varphi_{c(u)}) \\ &= d(c(u)). \end{aligned}$$

This completes the proof. □

Proof of Theorem 3.1. Suppose $G(\beta,c,\gamma)$ is X_1-unstable, and choose initial data u_0^k such that

$$\inf_{\varphi \in G(\beta,c,\gamma)} \|u_0^k - \varphi\|_{X_1} < \frac{1}{k},$$

and let $u^k(t)$ be the solution of (1.2) with $u^k(0) = u_0^k$. By continuity in t, there are some $\delta > 0$ and some times t_k such that

$$\inf_{\varphi \in G(\beta,c,\gamma)} \|u^k(t_k) - \varphi\|_{X_1} = \delta. \tag{3.4}$$

By the initial assumption, we can find $\varphi_k \in G(\beta,c,\gamma)$ such that

$$\lim_{k\to\infty} \|u_0^k - \varphi_k\|_{X_1} = 0.$$

Therefore, since E and V are continuous on X_1 and invariants of (1.2),

$$\lim_{k\to\infty} E(u^k(t_k)) - E(\varphi_k) = \lim_{k\to\infty} E(u_0^k) - E(\varphi_k) = 0 \tag{3.5}$$

and

$$\lim_{k\to\infty} V(u^k(t_k)) - V(\varphi_k) = \lim_{k\to\infty} V(u_0^k) - V(\varphi_k) = 0. \tag{3.6}$$

By Lemma 3.5, if δ is sufficiently small, we have

(3.7)

$$E(u^k(t_k)) - E(\varphi_k) - c(u^k(t_k))(V(u^k(t_k)) - V(\varphi_k)) \geq \frac{1}{4}d''(c)|c(u^k(t_k)) - c|^2.$$

By (3.4) there is some $\psi_k \in G(\beta,c,\gamma)$ such that $\|u^k(t_k) - \psi_k\|_{X_1} < 2\delta$, and by (2.10), we have

$$\|u^k(t_k)\|_{X_1} \leq \|\psi_k\|_{X_1} + 2\delta \leq 2\delta + A^{-1}I(\psi_k;\beta,c,\gamma) = 2\delta + \frac{2(p+1)}{A(p-1)}d(c) < \infty.$$

Thus since K is Lipschitz continuous on X_1 and d^{-1} is continuous, it follows that $c(u^k(t_k))$ is uniformly bounded in k. Thus by (3.5), (3.6), and (3.7) it follows that

$$\lim_{k\to\infty} c(u^k(t_k)) = c.$$

Continuity of K then implies

$$\lim_{k\to\infty} K(u^k(t_k)) = \lim_{k\to\infty} \frac{2(p+1)}{p-1} d(c(u^k(t_k))) = \frac{2(p+1)}{p-1} d(c). \tag{3.8}$$

By (3.2) and (3.1), we have

$$\begin{aligned} \frac{1}{2} I(u^k(t_k)) &= E(u^k(t_k)) - cV(u^k(t_k)) + \frac{1}{p+1} K(u^k(t_k)) \\ &= d(c) + E(u^k(t_k)) - E(\varphi_k) - c(V(u^k(t_k)) - V(\varphi_k)) + \frac{1}{p+1} K(u^k(t_k)), \end{aligned} \tag{3.9}$$

so it follows from (3.5), (3.6), and (3.8) that

$$\lim_{k\to\infty} I(u^k(t_k)) = \frac{2(p+1)}{p-1} d(c).$$

Thus $u^k(t_k)$ is a minimizing sequence and therefore has a subsequence which converges in X_1 to some $\varphi \in G(\beta, c, \gamma)$. This contradicts (3.4), so the proof of the theorem is complete. □

4. Instability. In this section we present two theorems which provide conditions for orbital instability of solitary waves. The first is complementary to the stability theorem, in that it guarantees instability when $d''(c) < 0$. The second does not involve the function d, but rather gives a set of sufficient conditions for instability directly in terms of the parameters β, γ, c, and p. While this result is not sharp, it does not rely on detailed knowledge of the function d. The proof is based on the work of Gonçalves Ribeiro [10], which is a modification of Shatah and Strauss' method [25].

The first theorem requires the following assumption.

Assumption 4.1. For each fixed $\beta > 0$ and $\gamma > 0$, there exists a C^1 map $c \mapsto \varphi_c$ from $(-\infty, 2\sqrt{\beta\gamma})$ to X_1 such that $\varphi_c \in G(\beta, c, \gamma)$.

THEOREM 4.2. *Suppose $\beta > 0$, $\gamma > 0$, $c < 2\sqrt{\beta\gamma}$ and Assumption* 4.1 *holds. If $d''(c) < 0$, then the orbit $\mathcal{O}_{\varphi_c}$ is X_1-unstable.*

THEOREM 4.3. *Let $\beta > 0$, $\gamma > 0$, $c < 2\sqrt{\beta\gamma}$, and $\varphi \in G(\beta, c, \gamma)$. Then the orbit $\mathcal{O}_\varphi$ is X_1-unstable if*

(i) *$c < 0$, $p > 5$, and $\gamma < \gamma_0$ for some small $\gamma_0 > 0$;*
(ii) *$c \le 0$, $p > 5 + 4\sqrt{2}$, and $\gamma > 0$; or*
(iii) *$c > 0$ and $p > \frac{10+k+\sqrt{(10+k)^2+4(7+k)}}{2}$, where $k = 8\left(\frac{2\sqrt{\beta\gamma}}{2\sqrt{\beta\gamma}-c} - 1\right)$.*

Theorems 4.2 and 4.3 are actually both direct corollaries of Theorem 4.4, with different choices of the "unstable direction" ϕ. The choices are given in Lemmas 4.10 and 4.11. As in the proof of the stability theorem, an important role will be played by the ϵ-neighborhoods of the orbits of solitary waves. Given $\varphi \in G(\beta, c, \gamma)$ and $\epsilon > 0$, we define

$$U_{\varphi,\epsilon} = \left\{ u \in X_1 \mid \inf_{v\in\mathcal{O}_\varphi} \|u - v\|_{X_1} < \epsilon \right\}.$$

THEOREM 4.4. *Assume $\beta > 0$, $\gamma > 0$, and $c < 2\sqrt{\beta\gamma}$. Let $\varphi \equiv \varphi_c \in G(\beta, c, \gamma)$ and define $\mathcal{M} = \{u \in X_1; V(u) = V(\varphi_c)\}$. If there exists $\phi \in L^2$ such that $\phi' \in X_s$, $s > 3/2$, $\phi'' \in X_1$, ϕ' is tangent to $\mathcal{M}$ at φ, and*

$$\langle L''(\varphi)\phi', \phi'\rangle < 0, \tag{4.1}$$

then there exist an $\epsilon > 0$ and a sequence $\{u_0^j\}$ in $U_{\varphi,\epsilon}$ such that

(i) *$u_0^j \to \varphi$ in X_1 as $j \to \infty$.*

(ii) *for all positive integers j, u^j is uniformly bounded, but escapes $U_{\varphi,\epsilon}$ in finite time, where u^j is the solution of (1.2) with $u^j(0) = u_0^j$.*

The proof of Theorem 4.4 is approached via a series of lemmas. Define

$$J(u) = \int \beta u_x^2 + \gamma(D_x^{-1}u)^2 - cu^2 + (p+1)F(u)\,dx \tag{4.2}$$

and

$$L(u) = E(u) - cV(u) = \frac{1}{2}J(u) + \frac{p-1}{2(p+1)}K(u).$$

LEMMA 4.5. *Assume $c < 2\sqrt{\beta\gamma}$. Then there exists a ground state $\varphi \in X_1$ satisfying $J(\varphi) = 0$ such that*

$$L(\varphi) = \inf\{L(u) \mid u \in X_1, u \neq 0, J(u) = 0\}. \tag{4.3}$$

Proof. The lemma follows by applying the arguments from the proof of Proposition 2.3 in [17]. □

LEMMA 4.6. *Fix $c < 2\sqrt{\gamma\beta}$ and let $\varphi \in G(\beta, c, \gamma)$. There are an $\epsilon_0 > 0$ and a unique C^2 map $\alpha : U_{\varphi,\epsilon_0} \to \mathbf{R}$ such that $\alpha(\varphi) = 0$ and for all $v \in U_{\varphi,\epsilon_0}$, and any $r \in \mathbf{R}$,*

(i) *$\langle \tau_{\alpha(v)}\varphi', v\rangle = 0$,*

(ii) *$\alpha(\tau_r v) = \alpha(v) + r$, and*

(iii) *$\alpha'(v) = -\frac{1}{\langle v, \varphi''(\cdot+\alpha(v))\rangle}\varphi'(\cdot + \alpha(v))$.*

In particular, for any $w \in \mathcal{O}_\varphi$, we have $\langle \alpha'(w), w\rangle = 0$, and $\alpha'(w) = \frac{1}{|\varphi'|_2^2}w'$.

Proof. The proof is standard. See Theorem 3.1 in [10], Lemma 3.5 in [1], or Lemma 3.5 in [3]. □

Consider a function $\phi \in L^2$ such that $\phi' \in X_1$. Define another vector field B_ϕ by

$$B_\phi(v) = \tau_{\alpha(v)}\phi' - \frac{\langle v, \tau_{\alpha(v)}\phi'\rangle}{\langle v, \tau_{\alpha(v)}\varphi''\rangle}\tau_{\alpha(v)}\varphi'' \tag{4.4}$$

for $v \in U_{\varphi,\epsilon}$.

The vector field B_ϕ is an extension of formula (4.2) in Bona, Souganidis, and Strauss [5], and a similar formula was also used in [1, 3, 10]. The important properties of B_ϕ are expressed in the following auxiliary result and will be used in the proof of Theorem 4.4.

LEMMA 4.7.

(i) *Assume that $\phi \in L^2$ such that $\phi', \phi'' \in X_1$. The mapping $B_\phi : U_{\varphi,\epsilon_0} \to X_1$ is C^1 with bounded derivative.*

(ii) *B_ϕ commutes with translations.*

(iii) *$\langle B_\phi(v), v\rangle = 0$ for all $v \in U_{\varphi,\epsilon_0}$.*

(iv) *If $\langle \varphi, \phi'\rangle = 0$, then $B_\phi(\varphi) = \phi'$.*

Proof. Since $\phi', \varphi'' \in X_1$, it is easy to see from the definition of B_ϕ that $B_\phi(v) \in X_1$ for all $v \in U_{\varphi,\epsilon}$. We now prove that B_ϕ is C^1 with bounded derivative. From part (iii) of Lemma 4.6 and (4.4), we have

$$B_\phi(v) = \tau_{\alpha(v)}\phi' + \left\langle v,\, \tau_{\alpha(v)}\phi' \right\rangle \frac{d}{dx}\alpha'(v).$$

A simple calculation shows

$$\begin{aligned} B'_\phi(v)w = {} & \langle \alpha'(v),\, w\rangle\, \tau_{\alpha(v)}\phi'' + \left\langle v,\, \tau_{\alpha(v)}\phi' \right\rangle \frac{d}{dx}\alpha''(v)w \\ & + \left(\left\langle w,\, \tau_{\alpha(v)}\phi' \right\rangle + \langle \alpha'(v),\, w\rangle \left\langle v,\, \tau_{\alpha(v)}\phi'' \right\rangle\right) \frac{d}{dx}\alpha'(v) \end{aligned} \tag{4.5}$$

for all $w \in X_1$. To show that B_ϕ is a C^1 function with bounded derivative, we need to show that all terms in the right side of (4.5) are bounded in $U_{\varphi,\epsilon}$. In fact, for $w \in X_1$ and $v \in U_{\varphi,\epsilon}$, we have

$$\begin{aligned} \left\langle v',\, \tau_{\alpha(v)}\varphi' \right\rangle \frac{d}{dx}\alpha''(v)w = {} & \langle \alpha'(v),\, w\rangle\, \tau_{\alpha(v)}\varphi''' + \left\langle \tau_{\alpha(v)}\varphi'',\, w \right\rangle \frac{d}{dx}\alpha'(v) \\ & - \langle \alpha'(v),\, w\rangle \left\langle v',\, \tau_{\alpha(v)}\varphi'' \right\rangle \frac{d}{dx}\alpha'(v). \end{aligned} \tag{4.6}$$

Setting $v = \varphi$ in (4.6) and using the relation $\alpha(\varphi) = 0$ yields

$$\frac{d}{dx}\alpha''(\varphi)w = \frac{\varphi'''}{|\varphi'|_2^4} \langle w,\, \varphi'\rangle + \frac{\varphi''}{|\varphi'|_2^4} \langle w,\, \varphi''\rangle . \tag{4.7}$$

Therefore,

$$\left\| \frac{d}{dx}\alpha''(\varphi) \right\|_{L(X_1,\, X_1)} \le C_0\left(\|\varphi\|_4\right)$$

and

$$\left\| \frac{d}{dx}\alpha'(\varphi) \right\|_{X_1} \le C_1(\|\varphi\|_3).$$

Since α is C^2 and $\frac{d}{dx}\alpha'$ and $\frac{d}{dx}\alpha''$ are continuous, by taking $\epsilon > 0$ small enough, if necessary, there exists a constant $C_2 > 0$ such that

$$\|\alpha'(v)\|_{X_1} \le C_2, \quad \left\| \frac{d}{dx}\alpha'(v) \right\|_{X_1} \le C_2, \quad \text{and} \quad \left\| \frac{d}{dx}\alpha''(v) \right\|_{L(X_1,\, X_1)} \le C_2$$

with $C_2 = C_2(\|\varphi\|_4)$ and for all $v \in U_{\varphi,\epsilon_0}$. It follows that B_ϕ is a C^1 function and the derivative of B_ϕ is bounded by

$$\left\| B'_\phi(v)w \right\|_{X_1} \le C_2\left(\|\phi''\|_{X_1} + (\epsilon + \|\varphi\|_{X_1})|\phi'|_2\right) \|w\|_{X_1},$$

which implies that

$$\|B'_\phi(v)\|_{L(X_1,\, X_1)} \le C \quad \forall v \in U_{\varphi,\epsilon},$$

where the constant C depends only on C_2, $\|\phi''\|_{X_1}$, $|\phi'|_2$, and $\|\varphi\|_{X_1}$. This proves (i). Statement (ii) can be obtained immediately from the relation $\alpha(\tau_y(v)) = \alpha(v) + y$ for

any $v \in X_1$ and $y \in \mathbf{R}$. Statement (iii) can obtained directly from the definition of B_ϕ. By $\alpha(\varphi) = 0$ and by the assumption in (iv), $\langle \varphi, \phi' \rangle = 0$, we have

$$B_\phi(\varphi) = \phi' + \frac{\langle \varphi, \phi' \rangle}{|\varphi'|_2^2} \varphi'' = \phi'.$$

This completes the proof of Lemma 4.7. □

LEMMA 4.8. *Let $\varphi \in G(\beta, c, \gamma)$. Assume that $\phi \in L^2$ is defined in Theorem* 4.4. *Then there exist $\epsilon_3 > 0$ and $\sigma_3 > 0$ such that for each $v_0 \in U_{\varphi,\epsilon_3}$,*

$$L(\varphi) \le L(v_0) + P(v_0)s \tag{4.8}$$

for some $s \in (-\sigma_3, \sigma_3)$, where $P(v) = \langle L'(v), B_\phi(v) \rangle$ and $L(v) = E(v) - cV(v)$.

Proof. Let $U_{\varphi,\epsilon}$ be as in Lemma 4.6. For each $v_0 \in U_{\varphi,\epsilon_0}$, consider the initial-value problem

$$\begin{aligned} \frac{dv}{ds} &= B_\phi(v), \\ v(0) &= v_0. \end{aligned} \tag{4.9}$$

By Lemma 4.7, it has a unique maximal solution $v(v_0, s) \in C^2(U_{\varphi,\epsilon_0}, (-\sigma, \sigma))$ with $\sigma = \sigma(v_0) > 0$. Moreover, for each $\epsilon_1 < \epsilon_0$, there exists $\sigma_1 > 0$ such that $\sigma(v_0) \ge \sigma_1$ for all $v_0 \in U_{\varphi,\epsilon_1}$. Hence, for fixed ϵ_1 and σ_1, we can define the C^2-mapping $s \in (-\sigma_1, \sigma_1) \mapsto L(v(v_0, s))$. Let $P(v) = \langle L'(v), B_\phi(v) \rangle$ and

$$R(v) = \langle L''(v) B_\phi(v), B_\phi(v) \rangle + \left\langle L'(v), B_\phi'(v)(B_\phi(v)) \right\rangle. \tag{4.10}$$

Applying Taylor's theorem yields

$$L(v(v_0, s)) = L(v_0) + P(v_0)s + \frac{1}{2} R(v(v_0, \xi s)) s^2 \tag{4.11}$$

for some $\xi \in (0, 1)$. Since $L'(\varphi) = 0$ and $R(\varphi) = \langle L''(\varphi)\phi', \phi' \rangle < 0$, it follows from the continuity of P and R that there exist $\epsilon_2 \in (0, \epsilon_1)$ and $\sigma_2 \in (0, \sigma_1)$ such that

$$L\left(v(v_0, s)\right) \le L(v_0) + P(v_0)s \tag{4.12}$$

for $v_0 \in B(\varphi, \epsilon_2)$ and $s \in (-\sigma_2, \sigma_2)$. On the other hand, a simple calculation shows that

$$J(v(v_0, s))\Big|_{(v_0,s)=(\varphi,0)} = 0$$

and

$$\frac{\partial}{\partial s} J(v(v_0, s))\Big|_{(v_0,s)=(\varphi,0)} = \langle J'(\varphi), \phi' \rangle. \tag{4.13}$$

We claim that $\langle J'(\varphi), \phi' \rangle \ne 0$. Otherwise, ϕ' would be tangent to $\mathcal{N}$ at φ, where

$$\mathcal{N} = \{ u \in X_1 \mid u \ne 0, J(u) = 0 \}.$$

Hence, $\langle L''(\varphi)\phi', \phi' \rangle \ge 0$ since φ minimizes L on $\mathcal{N}$ by Lemma 4.5. But this contradicts (4.1). Therefore, by the implicit function theorem, there exist $\epsilon_3 \in (0, \epsilon_2)$

and $\sigma_3 \in (0, \sigma_2)$ such that for every $v_0 \in B(\varphi, \epsilon_3)$, there exists a unique $s = s(v_0) \in (-\sigma_3, \sigma_3)$ such that

$$J\left(v(v_0, s(v_0))\right) = 0. \tag{4.14}$$

Applying (4.12) to $(v_0, s(v_0))$ given by (4.14) and taking into account that φ minimizes L on $\mathcal{N}$, we have

$$L(\varphi) \le L(v(v_0, s)) \le L(v_0) + P(v_0)s \tag{4.15}$$

for some $s \in (-\sigma_3, \sigma_3)$. The above inequality can be extended to U_{φ,ϵ_3} from the gauge invariance. □

Remark 4.9. From the relation

$$v(\varphi,\, s) = \varphi + \int_0^s \tau_{\alpha(v(\varphi,t))}\phi'\, dt - \int_0^s g(t)\tau_{\alpha(v(\varphi,\, t))}\varphi''\, dt,$$

where

$$g(t) = \frac{\left\langle v, \tau_{\alpha(v)}\phi' \right\rangle}{\left\langle v, \tau_{\alpha(v)}\varphi'' \right\rangle},$$

it is easy to see that $v(\varphi, s) \in X_l, l > \frac{3}{2}$, for all $s \in (-\sigma_3, \sigma_3)$.

Now we are in the position to prove Theorem 4.4.

Proof of Theorem 4.4. Since $v(v_0,\, s)$ commutes with τ_y, it follows by replacing v_0 with $v(\varphi, \delta)$ in (4.15) that

$$L(v(v(\varphi, \delta),\, s)) \le L(v(\varphi, \delta)) + P(v(\varphi,\, \delta))s \tag{4.16}$$

for any $s \in (-\sigma_2,\, \sigma_2)$ and $\delta \in (-\sigma_3, \sigma_3)$ with $0 < \sigma_3 < \sigma_2$. Taking $s = -\delta$, it thus transpires from (4.16) that

$$L(\varphi) \le L(v(\varphi,\, \delta)) - P(v(\varphi,\, \delta))\delta \tag{4.17}$$

for all $\delta \in (-\sigma_3,\, \sigma_3)$. Moreover, it follows from (4.11) and the fact that $P(\varphi) = 0$ that the mapping $\delta \mapsto L(v(\varphi,\, \delta))$ has a strict maximum locally at $\delta = 0$. Hence, we have

$$L(v(\varphi,\, \delta)) < L(\varphi) \tag{4.18}$$

for all $\delta \neq 0$ and $\delta \in (-\sigma_4,\, \sigma_4)$ with $0 < \sigma_4 \le \sigma_3$. This in turn implies from (4.17) that

$$P(v(\varphi,\, \delta)) < 0 \tag{4.19}$$

for all $\delta \in (0, \sigma_4)$. Let $\delta_j \in (0, \sigma_4)$ such that $\delta_j \to 0$ as $j \to \infty$. Consider the sequences of initial data $u_{0,j} = v(\varphi,\, \delta_j)$. Then by Remark 4.9, $u_{0,j} \in X_s$, $s > 3/2$ for all positive integers j and $u_{0,j} \to \varphi$ in X_1 as $j \to \infty$, which proves (i). For all integers j, we need only verify that the solution $u_j(t)$ of (1.2) with $u_j(0) = u_{0,j}$ escapes from U_{φ,ϵ_3} for some $\epsilon_3 > 0$ and for all positive integers j in finite time. To see this, let ϵ_3 be defined as in Lemma 4.8. Define

$$T_j = \sup\{\lambda > 0; \quad u_j(t) \in U_{\varphi,\epsilon_3}\ \forall t \in (0,\, \lambda)\}$$

and

$$P_- = \{v \in U_{\varphi,\epsilon_3}; \quad L(v) < L(\varphi), P(v) < 0\}.$$

Consider the case of the maximal existence time $T = +\infty$ by the definition of stability. It now follows from Lemma 4.8 that for all integers j and $t \in (0, T_j)$, there exists $s = s_j(t) \in (-\sigma_3, \sigma_3)$ satisfying

$$L(\varphi) \le L(u_j(t)) + P(u_j(t))s = L(u_{0,j}) + P(u_j(t))s. \tag{4.20}$$

By (4.18) and (4.19), $u_{0,j} \in P_-$. Then we deduce that $u_j(t) \in P_-$ for all $t \in [0, T_j]$. In fact, if $P(u_j(t_0)) > 0$ for some $t_0 \in [0, T_j]$, then the continuity of P implies that there exists some $t_1 \in [0, T_j]$ satisfying $P(u_j(t_1)) = 0$, and it thus follows from (4.20) that $L(\varphi) \le L(u_{0,j})$, which contradicts $u_{0,j} \in P_-$. Hence, by (4.20), P is bounded away from zero and

$$-P(u_j(t)) \ge \frac{L(\varphi) - L(u_{0,j})}{\sigma_3} = \eta_j > 0 \quad \forall t \in [0, T_j]. \tag{4.21}$$

We now define a Liapunov function

$$A(t) = \int_{\mathbf{R}} \phi(x + \alpha(u_j(t)))u_j(x,t)dx, \quad t \in [0, T_j]. \tag{4.22}$$

Then by the Cauchy–Schwarz inequality,

$$|A(t)| \le |\phi|_2 |u_j(t)|_2 = |\phi|_2 |u_{0,j}|_2 < +\infty, \quad t \in [0, T_j]. \tag{4.23}$$

On the other hand, using the Hamiltonian formulation

$$\frac{du_j}{dt} = -\partial_x E'(u_j)$$

of (1.2), we have

$$\begin{aligned}
\frac{dA}{dt} &= \alpha'(u_j(t)) \int_{\mathbf{R}} \phi'\left(x + \alpha(u_j(t))\right) u_j(t) + \int_{\mathbf{R}} \phi\left(x + \alpha(u_j(t))\right) \frac{du_j}{dt} \\
&= \left\langle \alpha'(u_j(t)), \frac{du_j}{dt} \right\rangle \left\langle \tau_{\alpha(u_j(t))}\phi', u_j \right\rangle + \left\langle \tau_{\alpha(u_j(t))}\phi, \frac{du_j}{dt} \right\rangle \\
&= \left\langle \left\langle \tau_{\alpha(u_j(t))}\phi', u_j(t) \right\rangle \frac{d\alpha'(u_j(t))}{dx} + \tau_{\alpha(u_j(t))}\phi', E'(u_j(t)) \right\rangle \\
&= \langle B_\phi(u_j(t)), E'(u_j(t)) \rangle \\
&= \langle B_\phi(u_j(t)), L'(u_j(t)) \rangle + c \langle B_\phi(u_j(t)), u_j(t) \rangle \\
&= P(u_j(t))
\end{aligned}$$

for $t \in [0, T_j]$, where $\langle B_\phi(u_j(t)), u_j(t) \rangle = 0$. Hence (4.21) yields the lower bound

$$-\frac{dA}{dt} \ge \eta_j > 0 \quad \forall t \in [0, T_j]. \tag{4.24}$$

Comparing (4.23) and (4.24), we conclude that $T_j < +\infty$ for all j. This completes the proof. □

In view of Theorem 4.4 we now look for functions ϕ that satisfy the inequality (4.1).

LEMMA 4.10. *Suppose $\beta > 0$, $\gamma > 0$, $c < 2\sqrt{\beta\gamma}$, and Assumption 4.1 holds. If $d''(c) < 0$, then there exists ϕ satisfying the conditions of Theorem 4.4.*

Proof. By Assumption 4.1, d is differentiable and by Lemma 3.4, $d'(c) = -V(\varphi_c)$. So if we define

$$g(h, \sigma) = V(\varphi_h + \sigma\varphi_c),$$

then $g(c, 0) = -d'(c)$ and

$$\frac{\partial g}{\partial h}(c, 0) = \left\langle V'(\varphi_c), \frac{\partial \varphi_c}{\partial c} \right\rangle = -d''(c) > 0.$$

Thus the implicit function theorem implies there is a C^2-mapping $h : (-\epsilon, \epsilon) \to (-\infty, 2\sqrt{\beta\gamma})$ such that $h(0) = c$ and

$$g(h(\sigma), \sigma) = V(\varphi_{h(\sigma)} + \sigma\varphi_c) = V(\varphi_c).$$

Therefore

$$0 = \frac{d}{d\sigma} g(h(\sigma), \sigma)\bigg|_{\sigma=0} = \left\langle V'(\varphi_c), h'(0)\frac{\partial \varphi_c}{\partial c} + \varphi_c \right\rangle, \tag{4.25}$$

so if we define

$$\phi(x) = \int_{-\infty}^{x} h'(0)\frac{\partial \varphi_c}{\partial c}(y) + \varphi_c(y)\, dy,$$

then ϕ' is tangent to $\mathcal{M}$ at φ_c and it follows from Assumption 4.1 that $\phi \in L^2$, $\phi' \in X_s$ for some $s > 3/2$ and $\phi'' \in X_1$. It remains to show that ϕ satisfies (4.1). First, observe that

$$L''(\varphi) = E''(\varphi_c) - cV''(\varphi_c) = -\beta\partial_x^2 - \gamma D_x^{-2} - c + f'(\varphi_c), \tag{4.26}$$

so

$$\begin{aligned} \langle L''(\varphi_c)\phi', \phi' \rangle = \langle L''(\varphi_c)\varphi_c, \varphi_c \rangle + 2h'(0)\left\langle L''(\varphi_c)\varphi_c, \frac{\partial \varphi_c}{\partial c} \right\rangle \\ + (h'(0))^2 \left\langle L''(\varphi_c)\frac{\partial \varphi_c}{\partial c}, \frac{\partial \varphi_c}{\partial c} \right\rangle. \end{aligned} \tag{4.27}$$

We claim that

$$\langle L''(\varphi_c)\varphi_c, \varphi_c \rangle = (1 - p)K(\varphi_c), \tag{4.28}$$

$$\left\langle L''(\varphi_c)\varphi_c, \frac{\partial \varphi_c}{\partial c} \right\rangle = -2d'(c), \tag{4.29}$$

and

$$\left\langle L''(\varphi_c)\frac{\partial \varphi_c}{\partial c}, \frac{\partial \varphi_c}{\partial c} \right\rangle = -d''(c). \tag{4.30}$$

To prove the first two identities, recall that $sf'(s) = pf(s)$ and $F' = f$. Thus

$$\begin{aligned}\langle L''(\varphi_c)\varphi_c, \varphi_c\rangle &= \int_{\mathbf{R}} (-\beta(\varphi_c)_{xx} - \gamma D_x^{-2}\varphi_c - c\varphi_c + pf(\varphi_c))\varphi_c\, dx \\ &= \int_{\mathbf{R}} (p-1)\varphi_c f(\varphi_c)\, dx \\ &= (1-p)K(\varphi_c)\end{aligned}$$

and

$$\begin{aligned}\left\langle L''(\varphi_c)\varphi_c, \frac{\partial\varphi_c}{\partial c}\right\rangle &= \int_{\mathbf{R}} (-\beta(\varphi_c)_{xx} - \gamma D_x^{-2}\varphi_c - c\varphi_c + pf(\varphi_c))\frac{\partial\varphi_c}{\partial c}\, dx \\ &= \int_{\mathbf{R}} (p-1)f(\varphi_c)\frac{\partial\varphi_c}{\partial c}\, dx \\ &= (p-1)\frac{d}{dc}\int_{\mathbf{R}} F(\varphi_c)\, dx \\ &= -2d'(c).\end{aligned}$$

For the third identity, differentiate the solitary wave equation with respect to c to find

$$\beta\left(\frac{\partial\varphi_c}{\partial c}\right)_{xx} + c\frac{\partial\varphi_c}{\partial c} + \varphi_c + \gamma D_x^{-2}\frac{\partial\varphi_c}{\partial c} = f'(\varphi)\frac{\partial\varphi_c}{\partial c},$$

so $L''(\varphi_c)\left(\frac{\partial\varphi_c}{\partial c}\right) = \varphi_c$, and therefore

$$\left\langle L''(\varphi_c)\frac{\partial\varphi_c}{\partial c}, \frac{\partial\varphi_c}{\partial c}\right\rangle = \left\langle \varphi_c, \frac{\partial\varphi_c}{\partial c}\right\rangle = \frac{1}{2}\frac{d}{dc}\int_{\mathbf{R}} \varphi_c^2\, dx = -d''(c)$$

by Lemma 3.4. We next compute $h'(0)$. Using (4.25) and Lemma 3.4, we have

$$0 = h'(0)\left\langle \varphi_c, \frac{\partial\varphi_c}{\partial c}\right\rangle + \langle\varphi_c, \varphi_c\rangle = \frac{1}{2}h'(0)\frac{d}{dc}\langle\varphi_c, \varphi_c\rangle + \langle\varphi_c, \varphi_c\rangle = -\frac{1}{2}h'(0)d''(c) - d'(c)$$

and therefore

$$h'(0) = -\frac{2d'(c)}{d''(c)}.$$

Finally, (4.27), (4.28), and (4.29) give

$$\langle L''(\varphi_c)\phi', \phi'\rangle = (1-p)K(\varphi_c) + 4\frac{(d'(c))^2}{d''(c)} < 0$$

under the assumption that $d''(c) < 0$. This proves the lemma. □

LEMMA 4.11. *Assume $c < 2\sqrt{\beta\gamma}$ and let $\varphi \in G(\beta, c, \gamma)$. Define*

$$\phi(x) = \int_{-\infty}^{x} (\varphi(y) + 2y\varphi'(y))\, dy.$$

Then

(i) *$\phi \in L^2$, $\phi' \in X_s, s > 3/2, \phi'' \in X_1$, and ϕ' is tangent to $\mathcal{M}$ at φ.*
(ii) *$\langle L''(\varphi)\phi', \phi'\rangle = \frac{(p-1)(5-p)}{p+1}K(\varphi) + 16\gamma\int_{\mathbf{R}}(D_x^{-1}\varphi)^2\, dx$.*

Proof. The first part of statement (i) is obvious because $\varphi \in X_s$ for $s > 3/2$ and φ is exponentially decaying at infinity. On the other hand, a simple calculation shows that

$$\langle \phi', \varphi \rangle = \int_{\mathbf{R}} (\varphi(x) + 2x\varphi'(x))\, \varphi(x)\, dx = \int_{\mathbf{R}} (x\varphi^2)'\, dx = 0.$$

This proves (i). Now we need to estimate the quantity $\langle L''(\varphi)\phi',\, \phi' \rangle$. Differentiating the solitary wave equation

$$\beta\varphi_{xx} + c\varphi + \gamma D_x^{-2}\varphi = f(\varphi) \tag{4.31}$$

gives

$$\beta\varphi_{xxx} + c\varphi_x + \gamma D_x^{-1}\varphi = f'(\varphi)\varphi_x. \tag{4.32}$$

We now claim that

$$\langle L''(\varphi)\varphi, x\varphi' \rangle = \frac{p-1}{p+1} K(\varphi) \tag{4.33}$$

and

$$\langle L''(\varphi)\,(x\varphi')\,, x\varphi' \rangle = \frac{p-1}{2(p+1)} K(\varphi) + 4\gamma \int_{\mathbf{R}} (D_x^{-1}\varphi)^2. \tag{4.34}$$

To prove these, again recall that $f'(\varphi)\varphi = pf(\varphi)$ and $F' = f$, so that

$$\begin{aligned}
\langle L''(\varphi)\varphi, x\varphi' \rangle &= \int_{\mathbf{R}} (-\beta\varphi_{xx} - \gamma D_x^{-2}\varphi - c\varphi + pf(\varphi)) x\varphi'\, dx \\
&= \int_{\mathbf{R}} (p-1) x\varphi' f(\varphi)\, dx \\
&= (1-p) \int_{\mathbf{R}} F(\varphi)\, dx \\
&= \frac{p-1}{p+1} K(\varphi).
\end{aligned}$$

Next, we note that the identities

$$\int_{\mathbf{R}} \beta\varphi_x^2 - c\varphi^2 + \gamma(D_x^{-1}\varphi)^2\, dx = K(\varphi) \tag{4.35}$$

and

$$\int_{\mathbf{R}} \frac{1}{2}\beta\varphi_x^2 + \frac{1}{2}c\varphi^2 - \frac{3}{2}\gamma(D_x^{-1}\varphi)^2\, dx = -\frac{1}{p+1} K(\varphi) \tag{4.36}$$

follow by multiplying (4.31) by φ and $x\varphi'$, respectively, and integrating. Combining these yields

$$\int_{\mathbf{R}} c\varphi^2\, dx - 2\gamma \int_{\mathbf{R}} (D_x^{-1}\varphi)^2\, dx + \frac{p+3}{2(p+1)} K(\varphi) = 0. \tag{4.37}$$

Next, we observe that

$$\begin{aligned}
L''(\varphi)(x\varphi') &= -\beta(x\varphi')'' - \gamma D_x^{-2}(x\varphi') - cx\varphi' + f'(\varphi)x\varphi' \\
&= -\beta(x\varphi''' + 2\varphi'') - \gamma x D_x^{-1}\varphi + 2\gamma D_x^{-2}\varphi - cx\varphi' + xf'(\varphi)\varphi' \\
&= -x(\beta\varphi''' + c\varphi' + \gamma D_x^{-1}\varphi - f'(\varphi)\varphi') - 2\beta\varphi'' + 2\gamma D_x^{-2}\varphi.
\end{aligned}$$

Using (4.31) and (4.32), this simplifies to

$$L''(\varphi)(x\varphi') = 2c\varphi + 4\gamma D_x^{-2}\varphi - 2f(\varphi),$$

so

$$\begin{aligned}\langle L''(\varphi)(x\varphi'), x\varphi'\rangle &= \int_{\mathbf{R}} (2c\varphi + 4\gamma D_x^{-2}\varphi - 2f(\varphi))x\varphi'\,dx \\ &= \int_{\mathbf{R}} -c\varphi^2 + 2F(\varphi) + 6\gamma(D_x^{-1}\varphi)^2\,dx.\end{aligned}$$

Together with (4.37), this implies

$$\begin{aligned}\langle L''(\varphi)(x\varphi'), x\varphi'\rangle &= 4\gamma\int_{\mathbf{R}} (D_x^{-1}\varphi)^2\,dx + \left(\frac{p+3}{2(p+1)} - \frac{2}{p+1}\right)K(\varphi) \\ &= 4\gamma\int_{\mathbf{R}} (D_x^{-1}\varphi)^2\,dx + \frac{p-1}{2(p+1)}K(\varphi),\end{aligned}$$

as claimed. Therefore we deduce from (4.28), (4.33), and (4.34) that

$$\begin{aligned}\langle L''(\varphi)\phi', \phi'\rangle &= \langle L''(\varphi)\varphi, \varphi\rangle + 4\langle L''(\varphi)\varphi, x\varphi'\rangle + 4\langle L''(\varphi)(x\varphi'), x\varphi'\rangle \\ &= (1-p)K(\varphi) + \frac{4(p-1)}{p+1}K(\varphi) + \frac{2(p-1)}{p+1}K(\varphi) + 16\gamma\int_{\mathbf{R}} \left(D_x^{-1}\varphi\right)^2\,dx \\ &= \frac{(p-1)(5-p)}{p+1}K(\varphi) + 16\gamma\int_{\mathbf{R}} \left(D_x^{-1}\varphi\right)^2\,dx.\end{aligned}$$

This completes the proof of the lemma. □

COROLLARY 4.12. *Let ϕ be defined as in Lemma 4.11. Then $\langle L''(\varphi)\phi', \phi'\rangle < 0$ if*

(i) *$c < 0$, $p > 5$, and $\gamma < \gamma_0$ for some small $\gamma_0 > 0$;*
(ii) *$c \le 0$, $p > 5 + 4\sqrt{2}$, and $\gamma > 0$; or*
(iii) *$0 < c < 2\sqrt{\gamma\beta}$ and $p > p_0$ with*

$$p_0 = \frac{10 + k + \sqrt{(10+k)^2 + 4(7+k)}}{2}$$

and

$$k = 8\left(\frac{2\sqrt{\beta\gamma}}{2\sqrt{\beta\gamma} - c} - 1\right) > 0.$$

Proof. To prove (i), we first claim that

$$\lim_{\gamma\to 0} \gamma\int_{\mathbf{R}} \left(D_x^{-1}\varphi\right)^2\,dx = 0.$$

In fact, in view of (4.35), we have

$$\begin{aligned}\gamma\int_{\mathbf{R}} \left(D_x^{-1}\varphi\right)^2\,dx &= \int_{\mathbf{R}} c\varphi^2 - \beta\varphi_x^2\,dx + K(\varphi) \\ &= \int_{\varphi} c\varphi^2 - \beta\varphi_x^2\,dx + (m(\beta, c, \gamma))^{\frac{p+1}{p-1}}.\end{aligned}$$

It is thereby inferred from Lemma 2.7 and Theorem 2.5 that

$$\lim_{\gamma\to 0}\gamma\int_{\mathbf{R}}\left(D_x^{-1}\varphi\right)^2\,dx=\int_{\mathbf{R}}c\varphi_0^2-\beta\left(\partial_x\varphi_0\right)^2\,dx+\left(m(\beta,c,0)\right)^{\frac{p+1}{p-1}}$$
$$=-I(\varphi_0;\beta,c,0)+\left(m(\beta,c,0)\right)^{\frac{p+1}{p-1}}=0,$$

where φ_0 is the ground state solution of the KdV equation with $c<0$. This in turn implies that

$$\lim_{\gamma\to 0}\langle L''(\varphi)\phi',\phi'\rangle=\frac{(p-1)(5-p)}{p+1}\left(m(\beta,c,0)\right)^{\frac{p+1}{p-1}}<0$$

for $c<0$ and $p>5$. This proves (i). To prove (ii) and (iii), we use (4.37) to write

$$2\gamma\int_{\mathbf{R}}\left(D_x^{-1}\varphi\right)^2\,dx=\frac{p+3}{2(p+1)}K(\varphi)+c\int_{\mathbf{R}}\varphi^2\,dx$$
$$=\frac{p+3}{2(p+1)}K(\varphi)-K(\varphi)+\int_{\mathbf{R}}\beta\varphi_x^2+\gamma\left(D_x^{-1}\varphi\right)^2\,dx$$
$$\leq\left(\frac{p+3}{2(p+1)}-1+\max\left\{1,\frac{2\sqrt{\beta\gamma}}{2\sqrt{\beta\gamma}-c}\right\}\right)K(\varphi).$$

Therefore it follows from formula (ii) in Lemma 4.11 that

$$\langle L''(\varphi)\phi',\phi'\rangle\leq\left(\frac{(p-1)(5-p)}{p+1}+\frac{4(p+3)}{p+1}-8(1-\rho)\right)K(\varphi),\tag{4.38}$$

where

$$\rho=\max\left\{1,\frac{2\sqrt{\beta\gamma}}{2\sqrt{\beta\gamma}-c}\right\}.$$

If $c\leq 0$, then $\rho=1$ and it follows that

$$\langle L''(\varphi)\phi',\phi'\rangle\leq\left(\frac{(p-1)(5-p)}{p+1}+\frac{4(p+3)}{p+1}\right)K(\varphi)$$
$$=-\frac{1}{p+1}\left(p-(5-4\sqrt{2})\right)\left(p-(5+4\sqrt{2})\right)K(\varphi)<0$$

under the assumption of p in (ii). If assumption (iii) is satisfied, then

$$p^2-(10+k)p-(7+k)>0,$$

where $k=8(\rho-1)>0$. It thus follows from (4.34) that

$$\langle L''(\varphi)\phi',\phi'\rangle=-\frac{1}{p+1}\left(p^2-(10+k)p-(7+k)\right)K(\varphi)<0.$$

The proof of the corollary is complete. □

5. Numerical results. We now present some numerical computations of $d(c)$ for the nonlinearity $f(u)=(-u)^p$, where $p\geq 3$ is an integer. The case $p=2$ is (1.1) and was considered in [15]. The strategy for computing d is to first compute numerically the solutions of the solitary wave equation (2.1) using a shooting method.

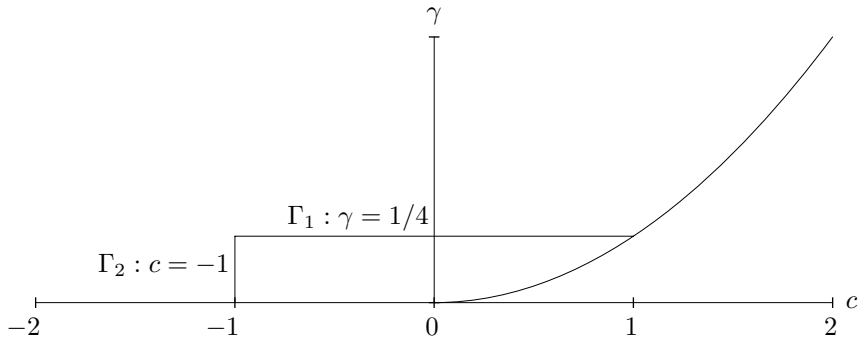
Then, using (3.3), we compute $d(c)$ and use a difference quotient to approximate $d''(c)$. The scaling property of d in Lemma 3.2 helps to reduce the calculations from the domain $c < 2\sqrt{\beta\gamma}$ to two finite line segments. To see this, substitute $r = 1/2\sqrt{\beta\gamma}$ and $s = (4\gamma/\beta)^{1/4}$ in the relation in Lemma 3.2 to get

$$d(\beta, c, \gamma) = \beta^{\frac{p+1}{2(p-1)}+\frac{1}{4}}(4\gamma)^{\frac{p+1}{2(p-1)}-\frac{1}{4}} d\left(1, \frac{c}{2\sqrt{\beta\gamma}}, \frac{1}{4}\right). \tag{5.1}$$

Thus the value of d along any surface of the form

$$S_\alpha = \{(\beta, c, \gamma) \mid c/2\sqrt{\beta\gamma} = \alpha\}$$

is determined by its value at any single point on that surface. We therefore need only compute d along some set of paths which crosses every such surface. We make the following choice. Let $\Gamma_1 = \{(1, c, 1/4) \mid -1 \le c < 1\}$ and $\Gamma_2 = \{(1, -1, \gamma) \mid 0 < \gamma \le 1/4\}$. Then Γ_1 crosses S_α for $-1 \le \alpha < 1$, and Γ_2 crosses S_α for $\alpha \le -1$. The paths Γ_1 and Γ_2 are shown below in the plane $\beta = 1$.



We now consider the sign of d_{cc} along these curves.

Along Γ_1. Differentiating (5.1) twice with respect to c gives

$$d_{cc}(\beta, c, \gamma) = \beta^{\frac{p+1}{2(p-1)}-\frac{3}{4}}(4\gamma)^{\frac{p+1}{2(p-1)}-\frac{5}{4}} d_{cc}\left(1, \frac{c}{2\sqrt{\beta\gamma}}, \frac{1}{4}\right).$$

Since $\beta > 0$ and $\gamma > 0$, it follows that the sign of d_{cc} within S_α is determined by the sign of $d_{cc}(1, \alpha, 1/4)$.

Along Γ_2. Using Lemma 3.2 again, we deduce that

$$d(1, c, \gamma) = (-c)^{\frac{p+3}{2(p-1)}} d\left(1, -1, \frac{\gamma}{c^2}\right)$$

for $c < 0$, so setting $q = \frac{p+3}{2(p-1)}$ and differentiating with respect to c gives

$$\begin{aligned} d_c(1, c, \gamma) &= -q(-c)^{q-1} d\left(1, -1, \frac{\gamma}{c^2}\right) - \frac{2\gamma}{c^3}(-c)^q d_\gamma\left(1, -1, \frac{\gamma}{c^2}\right) \\ &= -q(-c)^{q-1} d\left(1, -1, \frac{\gamma}{c^2}\right) + 2\gamma(-c)^{q-3} d_\gamma\left(1, -1, \frac{\gamma}{c^2}\right) \end{aligned}$$

and

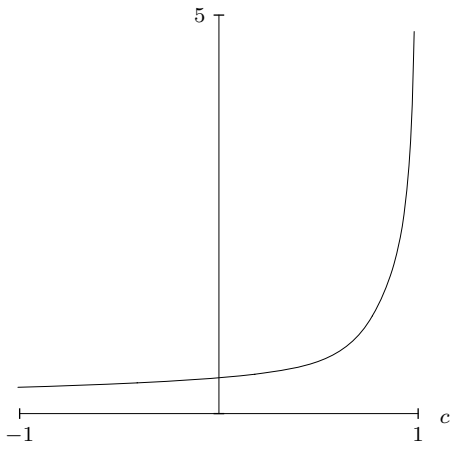
$$\begin{aligned} d_{cc}(1,c,\gamma) &= q(q-1)(-c)^{q-2}d\left(1,-1,\frac{\gamma}{c^2}\right) + \frac{2q\gamma}{c^3}(-c)^{q-1}d_\gamma\left(1,-1,\frac{\gamma}{c^2}\right) \\ &\quad - 2\gamma(q-3)(-c)^{q-4}d_\gamma\left(1,-1,\frac{\gamma}{c^2}\right) - \frac{4\gamma^2}{c^3}(-c)^{q-3}d_{\gamma\gamma}\left(1,-1,\frac{\gamma}{c^2}\right) \\ &= q(q-1)(-c)^{q-2}d\left(1,-1,\frac{\gamma}{c^2}\right) - 2\gamma(2q-3)(-c)^{q-4}d_\gamma\left(1,-1,\frac{\gamma}{c^2}\right) \\ &\quad + 4\gamma^2(-c)^{q-6}d_{\gamma\gamma}\left(1,-1,\frac{\gamma}{c^2}\right). \end{aligned}$$

Setting $r = \gamma/c^2$, this simplifies to $d_{cc}(1,c,\gamma) = (-c)^{q-2}g(r)$, where

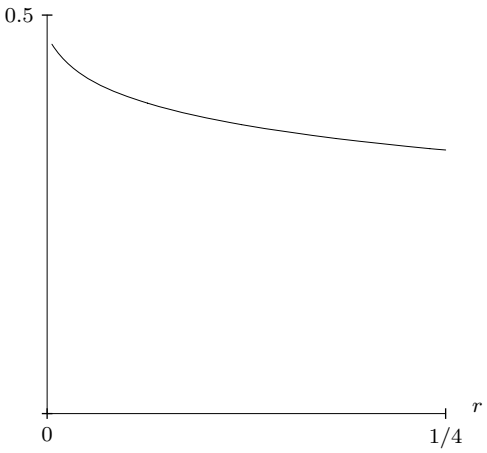
$$g(r) = 4r^2 d_{\gamma\gamma}(1,-1,r) - 2(2q-3)rd_\gamma(1,-1,r) + q(q-1)d(1,-1,r).$$

So it suffices to determine the sign of $g(r)$ for $0 < r \le \frac{1}{4}$. In the following 10 graphs, we consider the nonlinearity $f(u) = (-u)^p$ for several values of p. We remark that the case $p = 2$ is the classical Ostrovsky equation (1.1), for which it was shown by the authors in [15] using the same numerical method that $d''(c)$ is positive for all β, γ, and $c < 2\sqrt{\beta\gamma}$.

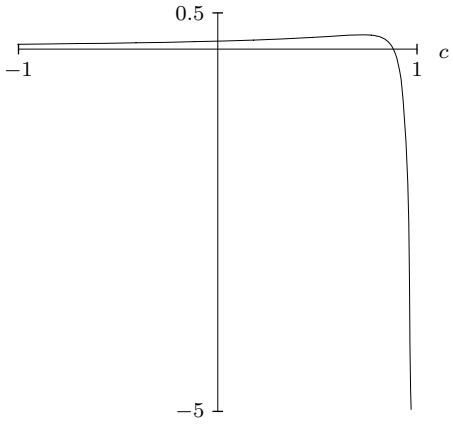
d_{cc}, $p = 3$, $\beta = 1$, $\gamma = 0.25$



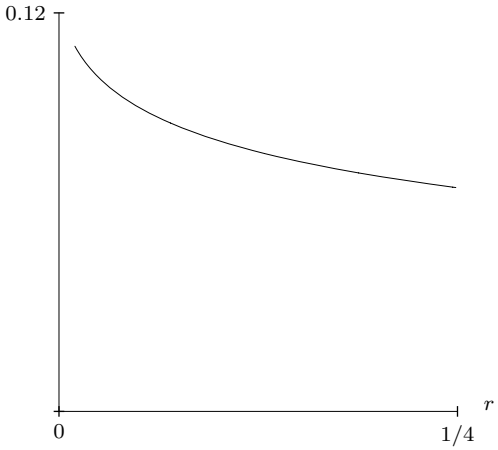
$g(r)$, $p = 3$, $\beta = 1$, $c = -1$

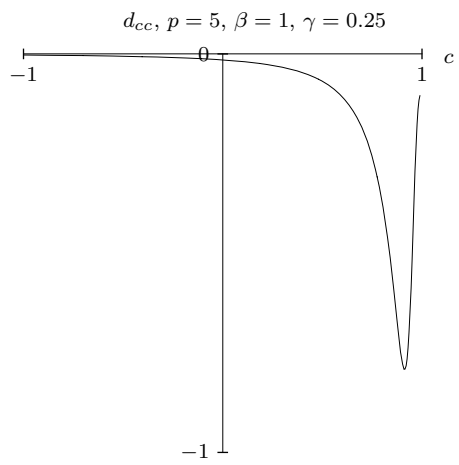


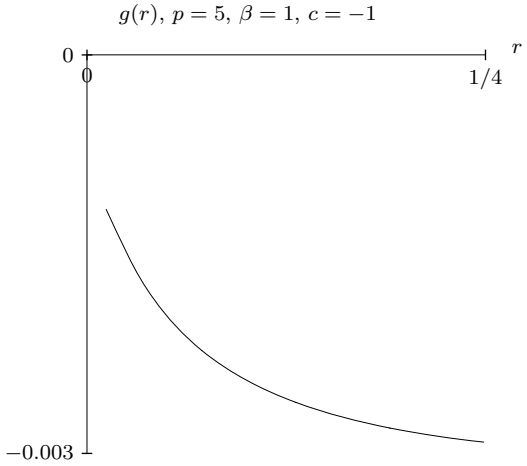
d_{cc}, $p = 4$, $\beta = 1$, $\gamma = 0.25$

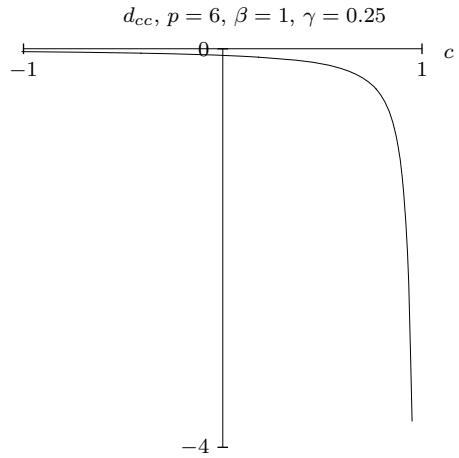


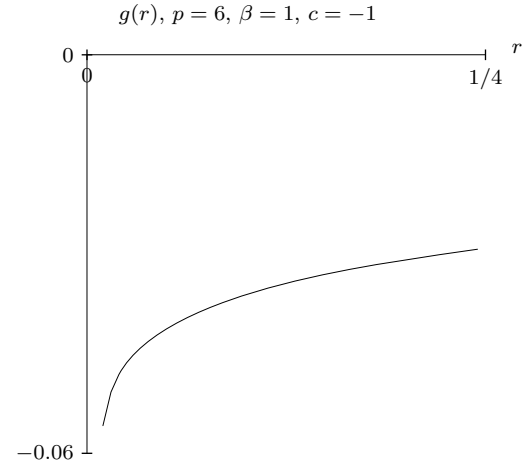
$g(r)$, $p = 4$, $\beta = 1$, $c = -1$

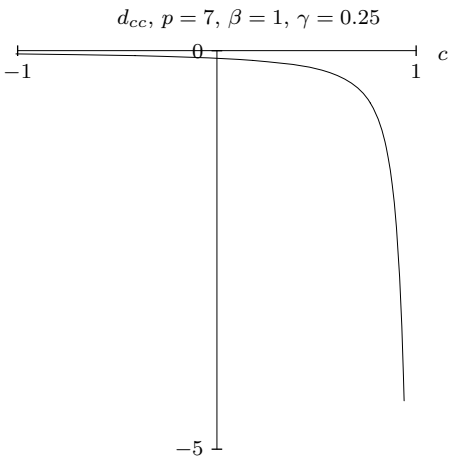


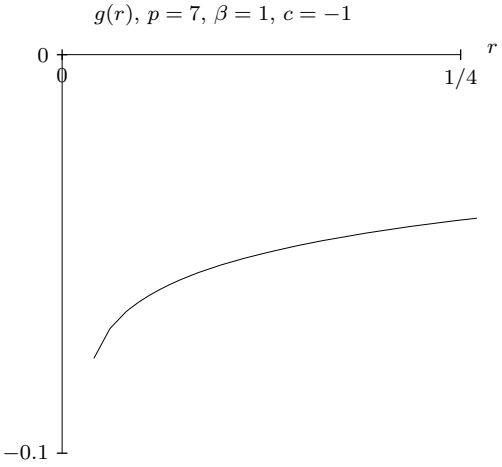

d_{cc}, $p = 5$, $\beta = 1$, $\gamma = 0.25$
0
c
-1
1
-1


$g(r)$, $p = 5$, $\beta = 1$, $c = -1$
0
r
0
1/4
-0.003


d_{cc}, $p = 6$, $\beta = 1$, $\gamma = 0.25$
0
c
-1
1
-4


$g(r)$, $p = 6$, $\beta = 1$, $c = -1$
0
r
0
1/4
-0.06


d_{cc}, $p = 7$, $\beta = 1$, $\gamma = 0.25$
0
c
-1
1
-5


$g(r)$, $p = 7$, $\beta = 1$, $c = -1$
0
r
0
1/4
-0.1

Using Theorems 3.1 and 4.3, we arrive at the following conclusions.

1. When $p = 3$, all solitary waves are stable for $c < 2\sqrt{\beta\gamma}$.
2. When $p = 4$, there exists α_0 (≈ 0.88) such that solitary waves are stable for $\frac{c}{2\sqrt{\beta\gamma}} < \alpha_0$ and solitary waves are unstable for $\alpha_0 < \frac{c}{2\sqrt{\beta\gamma}} < 1$.
3. When $p = 5$, 6, or 7, all solitary waves are unstable for $c < 2\sqrt{\beta\gamma}$.

The case $p = 4$ seems most interesting due to the change of stability. We conjecture that for all $p \geq 5$, solitary waves are unstable.

Acknowledgments. The authors are grateful for the constructive suggestions made by the referees.

REFERENCES

[1] J. Angulo, *On the instability of solitary waves solutions of the generalized Benjamin equation*, Adv. Differential Equations, 8 (2003), pp. 55–82.
[2] E. S. Benilov, *On the surface waves in a shallow channel with an uneven bottom*, Stud. Appl. Math., 87 (1992), pp. 1–14.
[3] J. Bona and Y. Liu, *Instability of solitary wave solutions of the* 3*-dimensional Kadomtsev-Petviashvili equation*, Adv. Differential Equations, 7 (2002), pp. 1–23.
[4] A. de Bouard and J.-C. Saut, *Solitary waves of generalized Kadomtsev-Petviashvili equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 211–236.
[5] J. Bona, P. Souganidis, and W. Strauss, *Stability and instability of solitary waves of Korteweg-de Vries type*, Proc. Roy. Soc. London Ser. A, 411 (1987), pp. 395–412.
[6] J. Fröhlich, E. H. Lieb, and M. Loss, *Stability of Coulomb systems with magnetic fields* I. *The one-electron atom*, Comm. Math. Phys., 104 (1986), pp. 251–270.
[7] V. N. Galkin and Y. A. Stepanyants, *On the existence of stationary solitary waves in a rotating field*, J. Appl. Math. Mech., 55 (1991), pp. 939–943.
[8] O. A. Gilman, R. Grimshaw, and Y. A. Stepanyants, *Approximate and numerical solutions of the stationary Ostrovsky equation*, Stud. Appl. Math., 95 (1995), pp. 115–126.
[9] R. Grimshaw, *Evolution equations for weakly nonlinear long internal waves in a rotating fluid*, Stud. Appl. Math., 73 (1985), pp. 1–33.
[10] J. Gonçalves Ribeiro, *Instability of symmetric stationary states for some nonlinear Schrödinger equations with an external magnetic field*, Ann. Inst. H. Poincaré, Phys. Théor., 54 (1991), pp. 403–433.
[11] Y. Guo and G. Rein, *Existence and stability of Camm type steady states in galactic dynamics*, Indiana U. Math. J., 48 (1999), pp. 1237–1255.
[12] B. B. Kadomtsev and V. I. Petviashvili, *On the stability of solitary waves in weakly dispersive media*, Sov. Phys. Dokl., 15 (1970), pp. 539–541.
[13] A. Leonov, *The effect of the earth's rotation on the propagation of weak nonlinear surface and internal long oceanic waves*, Ann. New York Acad. Sci., 373 (1981), pp. 150–159.
[14] S. Levandosky, *A stability analysis of fifth-order water wave models*, Phys. D, 125 (1999), pp. 222–240.
[15] S. Levandosky and Y. Liu, *Stability and Weak Rotation Limit of Solitary Waves of the Ostrovsky Equation*, preprint.
[16] P. L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case, Part* 1 *and Part* 2, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 1 (1984), pp. 109–145; 223–283.
[17] Y. Liu, *Stability of solitary waves for the Ostrovsky equation with weak rotation*, submitted.
[18] Y. Liu and V. Varlamov, *Stability of solitary waves and weak rotation limit for the Ostrovsky equation*, J. Differential Equations, 203 (2004), pp. 159–183.
[19] R. M. Miura, C. S. Gardner, and M. D. Kruskal, *Korteweg-de Vries equation and generalizations* II. *Existence conservation laws and constant of motion*, J. Math. Phys., 9 (1968), pp. 1204–1209.
[20] L. A. Ostrovsky, *Nonlinear internal waves in a rotating ocean*, Okeanologia, 18 (1978), pp. 181–191.
[21] L. A. Ostrovsky and Y. A. Stepanyants, *Nonlinear surface and internal waves in rotating fluids*, in Nonlinear Waves 3, Res. Rep. Phys., Springer, Berlin, Heidelberg, 1990, pp. 106–128.

[22] V. Shrira, *Propagation of long nonlinear waves in a layer of a rotating fluid*, Iza. Akad. Sci. USSR Atmospher. Ocean. Phys., 17 (1981), pp. 76–81 (in Russian).
[23] V. Shrira, *On long essentially non-linear waves in a rotating ocean*, Iza. Akad. Sci. USSR Atmospher. Ocean. Phys., 22 (1986), pp. 395–405 (in Russian).
[24] J. Shatah, *Stable standing waves of nonlinear Klein-Gordon equations*, Comm. Math. Phys., 91 (1983), pp. 313–327.
[25] J. Shatah and W. Strauss, *Instability of nonlinear bound states*, Comm. Math. Phys., 100 (1985), pp. 173–190.
[26] V. Varlamov and Y. Liu, *Cauchy problem for the Ostrovsky equation*, Discrete Contin. Dyn. Syst., 10 (2004), pp. 731–753.

SIAM J. MATH. ANAL.
Vol. 38, No. 3, pp. 1012–1034

THE DISCRETE PULSE TRANSFORM*

C. H. ROHWER† AND D. P. LAURIE†

Abstract. We investigate a recent algorithm, here called a discrete pulse transform (DPT), for the multiresolution analysis of a sequence. A DPT represents a sequence as a sum of pulses, where a pulse is a sequence which is zero everywhere except for a certain number of consecutive elements which have a constant nonzero value. Unlike the discrete Fourier and wavelet transforms, the DPT is not a discretization of an underlying continuous model, but is inherently discrete. The DPT is composed of nonlinear morphological filters based only on the order relations between elements of the sequence. It is comparable to, but computationally more efficient than, the median transform, and more amenable to theoretical analysis. In particular, we show that a DPT has remarkable shape preserving and consistency properties.

Key words. nonlinear smoothing, variation preservation, fully trend preserving, multiresolution decomposition

AMS subject classifications. Primary, 94A12, 41A46; Secondary, 47H99, 65T99, 42C99

DOI. 10.1137/040620862

1. Introduction.

The *discrete Fourier transform* (DFT) has often been called the most important mathematical tool of modern technology. There are three main reasons for this high esteem:

Multiresolution. The DFT allows us to decompose a given finite sequence into a sum of component subsequences, each of which can be interpreted as a feature appearing at a particular frequency level.

Predictability. The theoretical properties of the DFT are well understood, making it possible to predict its behavior over a wide range of typical applications.

Efficiency. There exists a very efficient procedure (the fast fourier transform) for the computation of the DFT whenever the period of the sequence has no large prime factors.

However, the DFT is primarily intended for the analysis of smooth periodic phenomena and is ineffective when the data contains discontinuities (as attested by the well-known Gibbs phenomenon) or is mainly composed of local features. In an attempt to overcome these problems, wavelet decompositions were developed over the last twenty years (an excellent overview of the state of the art is [Chu97]) to the point where it would be fair to say that all three of the reasons for using the DFT are largely applicable to wavelet decompositions too.

In addition, wavelets have the following desirable property:

Locality. The component sequences are local in a precisely definable sense.

In particular, B-splines [Sch67] fit very well into the wavelet framework. Instead of frequency levels, as in the case of the FFT, we have *resolution levels*, so that a wavelet decomposition is a technique for *multiresolution analysis* (MRA). The aim of MRA is to decompose a signal as a sum of separate signals, each of which is meaningful at its own resolution level. Ideally, the signal at each resolution level is undisturbed

*Received by the editors December 15, 2004; accepted for publication (in revised form) March 21, 2006; published electronically November 14, 2006.
http://www.siam.org/journals/sima/38-3/62086.html
†Department of Mathematics, University of Stellenbosch, Matieland 7602, South Africa (chr@sun.ac.za, dpl@sun.ac.za).

by those at the other resolution levels, allowing easy recognition of local features at each level.

In some crucial applications, such as image processing, the expectations raised by the more local nature of wavelet decompositions may have been too high. Wavelet transforms are conceptually based on orthogonality properties of continuous functions, and their application to discrete data can be thought of as a discretization of a process in which an underlying continuous function is decomposed into a sum of other continuous functions. Discontinuities can be handled, but only if their location can be pinpointed. The combination of an implied smooth model with a linear filter causes the well-known phenomenon of erosion of edges, which in image processing is a most undesirable side effect.

In these applications wavelet transforms are not fully satisfactory, and one must look elsewhere. Unavoidably, linearity of the transform must be sacrificed. The consensus among practitioners [BMS98] seems to be that the *median transform* works better than wavelet transforms.

Conceptually, the averaging filters of wavelet decomposition are replaced by median smoothers, and the analogy with wavelet decomposition is heuristically developed [BMS98]. The idea for the median transform stems from the theory of nonlinear smoothers, where running medians have for a long time [Vel77, Mal80] (longer, in fact, than wavelets) been the method of choice for the removal of impulsive noise.

A median filter has one very important property from the point of view of image processing:

Incisiveness. A sharp spike is recognized as impulsive noise that can be surgically removed without harming the signal; that same ability can be exploited to detect an edge as an important feature that should not be blurred.

In other words, a median filter does not exhibit the Gibbs phenomenon. Despite this obvious advantage, median transforms have some serious drawbacks. In particular, they fail in comparison to wavelets on two counts: they are computationally expensive and, perhaps more serious, there is almost no theory which can predict their behavior. In particular, they do not even have the property of idempotence, as can be seen from the well-known example where x is the square wave consisting of alternating trains of k zeros and k ones: if $n = 2k + 1$, the median-of-n operator produces $M_n x = 1 - x$.

In this paper we investigate a recent morphological transform which we call the *discrete pulse transform* (DPT). The DPT was proposed and motivated (although not under that name) in [Roh02b], and has among others the following notable properties:

Multiresolution. The DPT is a multiresolution decomposition in a precise sense. A sequence of nonlinear smoothing operators from the $LULU$ family [Roh89, Roh99, Roh02a, Roh02c, Roh03, Roh02b, RT91, RW02] are used to strip off the features of a sequence at different resolution levels.

Efficiency. There exists an efficient procedure for the computation of the DPT, much faster than the median transforms.

Locality. The component sequences are local in the following precisely definable sense: each can be seen as a pulse train, that is, a sum of (block-) pulses of the same sign and of equal width n, separated by at least n zeros. (A (block-) pulse of width n is a sequence that has a constant nonzero value at n contiguous elements and is zero elsewhere.)

Incisiveness. It behaves at least as well as the median transforms in the presence of discontinuities.

Predictability. The DPT is consistent in the following sense: if you analyze a sequence x by the DPT, form a new sequence z by a linear combination of those components *with all components multiplied by arbitrary nonnegative constants*, and analyze z by the DPT, you will recover the same components. Roughly speaking, one can say that this property is the analogue of the orthogonality property of the DFT.

In view of the nonlinearity of the smoothing operators used, the consistency of the DPT is perhaps the most surprising result of all. The main reason for it is that the underlying $LULU$ filters are stable in the sense that there is a clearly identifiable set of sequences (the n-monotone sequences, defined below) which pass unchanged through the first n filters. This property is shared by the median filters, but unlike the DPT, they do not necessarily generate such sequences.

Some of these properties were already given in earlier work [Roh89, Roh99, Roh02a, Roh02c, Roh03, Roh02b, RT91, RW02]. Others will be demonstrated in this article. The new results in this article deal with:

- some remarkable shape preserving properties of the smoothers involved in the decomposition,
- consistency of the decomposition: since it is not obvious in advance what is meant by a component, we work gradually towards a definition. Our main result, Theorem 8, shows that the DPT is consistent when a component is defined as a positive or negative pulse train filtered off at each resolution level.

We end with some remarks on the pulse structure of the decomposition, including the conjecture that in fact a component could be defined as an individual pulse without impairing the consistency of the decomposition.

2. A sneak preview. In this section, we show what the DPT can do, before going into details of what it is and how it does it, for the sake of those readers who believe that one example is worth a hundred theorems.

The research leading to the DPT was originally motivated by a particular application: that of recovering the time of arrival of an impulsive signal from a background of white noise. For example, any piece of equipment that relies on pulse length modulation (such as radio-controlled aircraft) leads to this problem.

In Figure 1 we show simulated data of this type: a time series with 100 points consisting of Gaussian white noise with mean 0 and variance 1 added to a positive pulse with amplitude 2 and length 11 in positions 63–73. Note that the amplitude of the noise occasionally exceeds that of the signal. The width of the pulse, and that it is positive, is in practice known, but not its position or amplitude.

Following the advice of Bijaoni, Murtagh, and Starck [BMS98], we dismiss all linear filters and try a median filter. In Figure 2 we show the results of the median-of-n filters with $n = 11$, with the original signal superimposed. This does not look very good, so we also try the idea of applying median-of-n filters with $n = 3, 5, 7, 9, 11$ successively. This looks better, and many people might be satisfied with the result, particularly those who have the advantage of knowing what the correct signal is. One could continue to $n = 13, 15, 17, 19, 21$, since all these filters should pass the signal through unharmed, and some of those look even better to the well-prepared human eye. However, there is no theory that tells us in advance where to stop. Moreover, the mere fact that all these smoothed values look good implies that little useful information can be had by scrutinizing the differences between them.

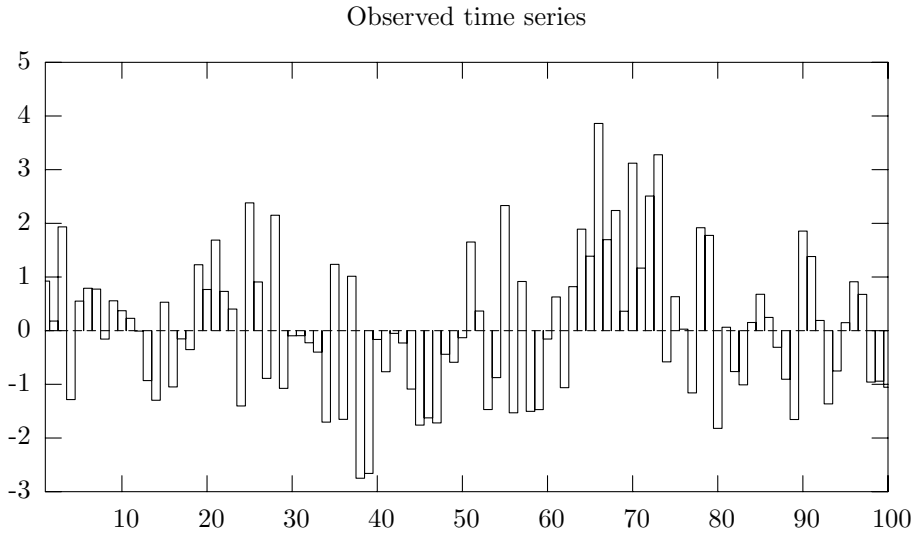


FIG. 1. *Observed time series of simulated data obtained by adding Gaussian random noise to a piecewise constant signal.*

There are two main flavors of the DPT:

$$\mathcal{C} = (C_1, C_2, C_3, \dots) \quad \text{and} \quad \mathcal{F} = (F_1, F_2, F_3, \dots),$$

each of which works like the iterated median in the sense that $F_n x$ is obtained by applying a certain nonlinear filter to $F_{n-1}x$, etc. In general, $F_n x \leqslant C_n x$; in an application, as in this example, one would apply both DPTs and base one's conclusion on what they agree on. In Figure 3 we show the result of the DPT filters C_9 and F_9, also with the original signal superimposed. The reason for stopping at this stage, before pulses of length 10 have been stripped off, will be given shortly, when we discuss conservation of variation. One could argue that both cases show a slightly sharper leading edge than the iterated median filter, but the argument would hardly be conclusive, as here, too, it would be based on knowledge of the actual signal.

However, the main point of the DPT is not the quality of the individual filters but the information given by the pulse decomposition, obtained from the differences $F_n x - F_{n-1}x$, $C_n x - C_{n-1}x$.

First, we look at the decomposition of total variation, shown in Table 1. This decomposition is analogous to the power spectrum of Fourier analysis. Note the property that the total variation of the various decomposition levels sums to the total variation of the data, in the same way that the squared norm of the components of a Fourier analysis sums to the squared norm of the data. The variation conservation property explains why the DPT is in practice found to be stable, even though it relies on order information.

Each component in the decomposition naturally splits into positive and negative pulses, analogous to the way that Fourier components split into odd and even parts. We expect a pulse of length 11 to show up at monotonicity level 11, with some spillover to levels 10 and 12. In fact, both DPT decompositions show no positive pulses at levels

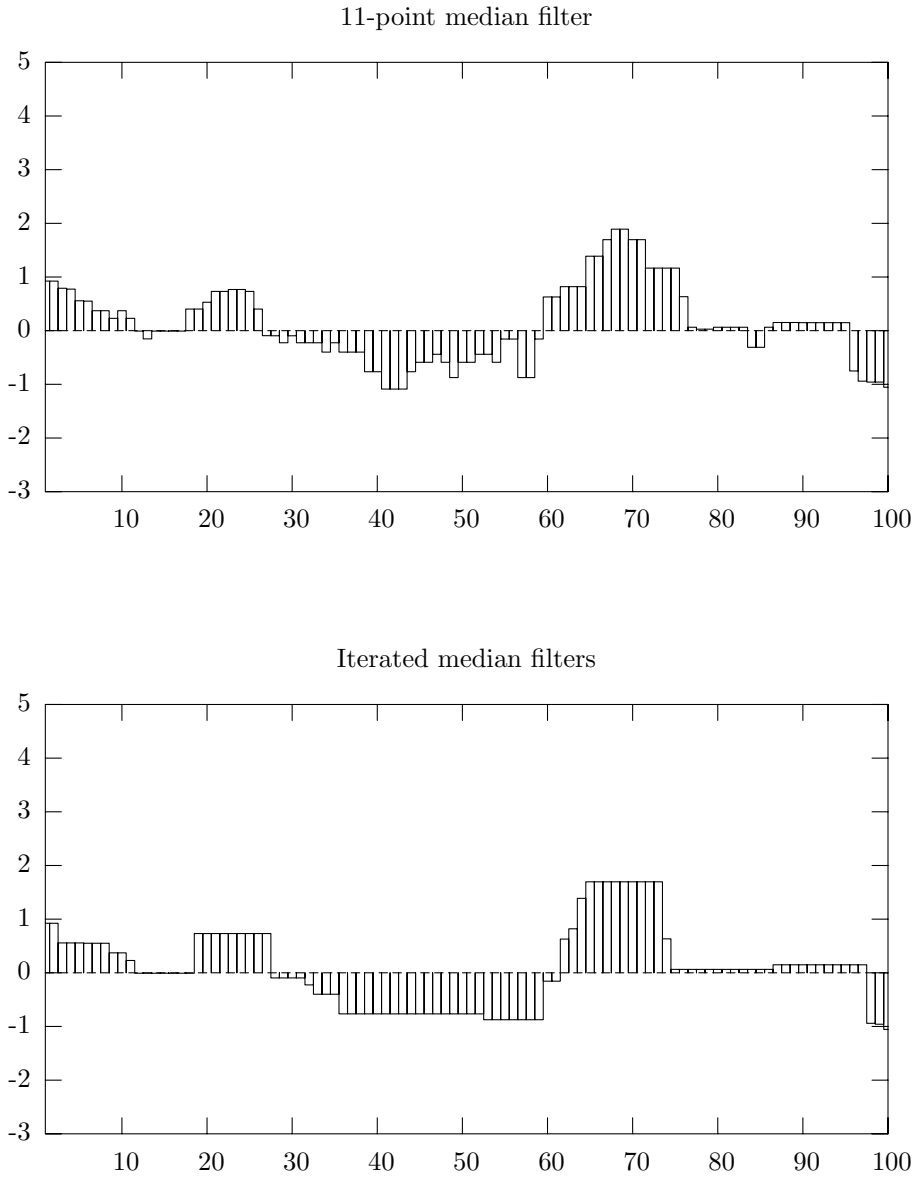


FIG. 2. *Median filters applied to the time series in Figure* 1. *In the first graph, the* 11*-point median filter was applied directly to the original time series; in the second graph, median filters of length* $3, 5, 7, 9,$ *and* 11 *were applied successively.*

9 and 12, so we form our reconstruction of the signal by combining the positive parts of levels 10 and 11. The result is shown in Figure 4.

It is not necessary to know in advance where to look to recognize a clear and unambiguous agreement between the two DPT transforms that a signal of the required kind is present. Although the DPT has not recovered the full amplitude of the signal,

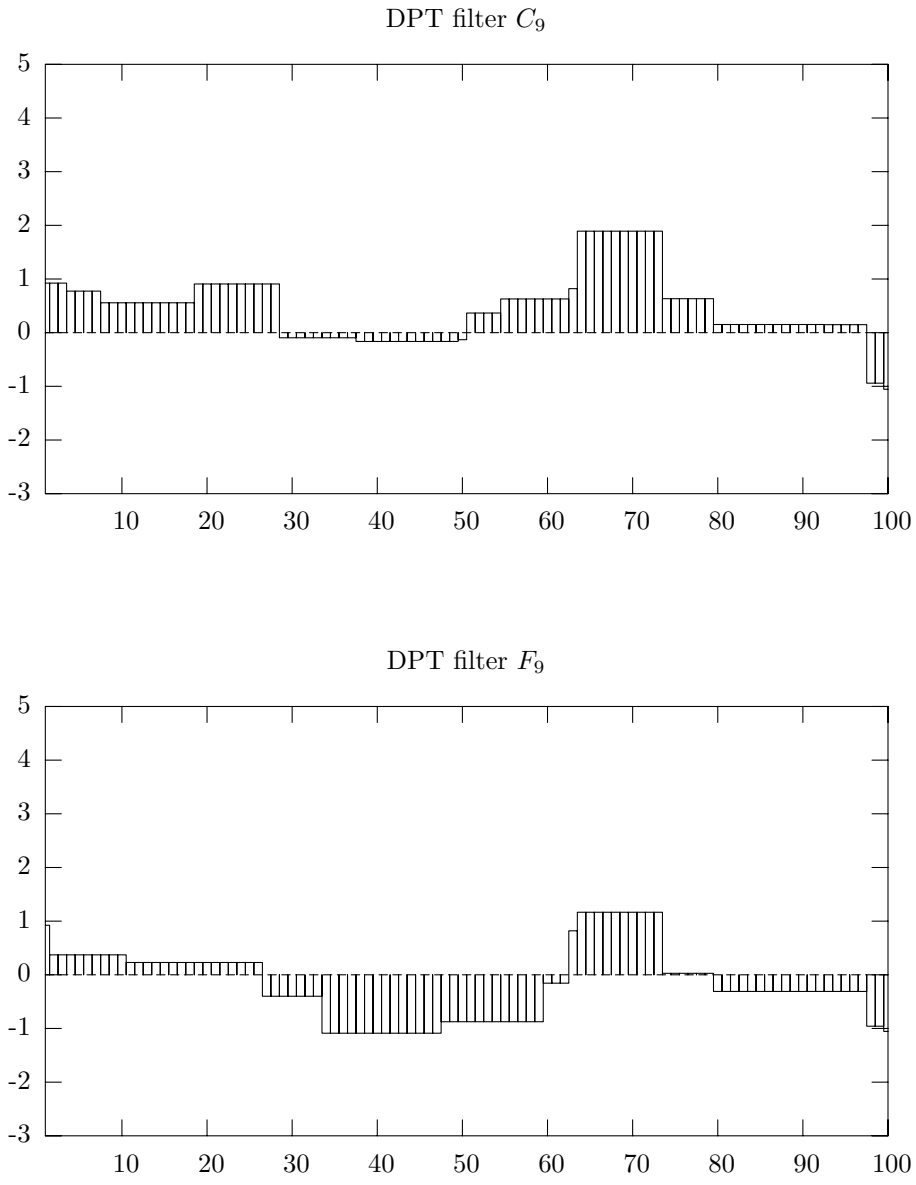


FIG. 3. *DPT filters applied to the time series in Figure* 1.

its position has been pinpointed. Also note that noise has been almost completely eliminated.

Now have another look at the figures arising from the median transform. Sure, there is a pulse somewhere in the midsixties. But can one say with confidence where it starts? Or how wide it is?

For this application, we are not aware of any other filter, linear or nonlinear, that can rival the DPT.

TABLE 1

Total variation v_m of the pulses, and p_m of the positive pulses, of width m identified by the DPT. Width ∞ refers to the final monotonic sequence. The omission of a width level means that no pulses were identified at that level. Note that in both cases $\sum v_m$ equals the total variation of the original observed sequence.

DPT using C_n			DPT using F_n		
m	v_m	p_m	m	v_m	p_m
1	84.00	24.43	1	87.57	65.15
2	24.87	5.91	2	18.60	13.59
3	7.69	3.51	3	6.24	3.31
4	2.01	0.54	4	4.56	1.30
6	0.71	0.00	5	1.01	1.01
7	0.60	0.00	6	1.82	0.00
8	0.98	0.69	7	0.47	0.00
9	0.37	0.00	8	1.26	1.26
10	2.84	2.84	10	0.69	0.69
11	0.37	0.37	11	1.58	1.58
12	0.07	0.00	14	0.43	0.00
13	0.07	0.00	17	0.37	0.37
17	0.01	0.01	20	0.31	0.31
22	0.92	0.00	26	0.95	0.00
25	0.53	0.53	33	0.18	0.00
∞	1.98		∞	1.98	

3. Analogy with the FFT. The easiest way to understand the DPT is by analogy with the well-known DFT. For the purpose at hand it is convenient to view the DFT in the following framework.

Conceptually, a finite sequence of length N,

$$x = \langle x_i,\ i = 1, 2, \ldots, N \rangle, \tag{1}$$

is extended to an infinite periodic sequence by the stipulation that $x_i = x_{i+N}$ holds for all i. The periodic extension is required for theoretical purposes: all computations involve only a sequence of length N. The practical implementation of the DFT is known as the FFT when all prime factors of N are small, but for our theoretical discussion we have no need for that hypothesis.

The DFT decomposes x into $M = \lfloor \frac{N}{2} + 1 \rfloor$ component sequences,

$$x = s^{(0)} + s^{(1)} + \cdots + s^{(M)},$$

where $s^{(0)}$ is a constant sequence and $s^{(n)}$ is a periodic sequence with frequency n times the fundamental frequency. As is well known, the mappings D_n that map x onto $s^{(n)}$ are projections onto a one-dimensional space when $n = 0$ (and for even N, also when $n = N/2$) and onto a two-dimensional space otherwise. The typical component sequence can thus be represented by one complex number, or by two real numbers, interpreted either as coefficients with respect to a chosen basis or an amplitude and phase pair. It is convenient to view the DFT as a vector of sequences

$$D(x) = DFT(x) = [D_0(x), D_1(x), \ldots, D_M(x)],$$

where $s^{(i)} = D_i(x)$.

Since the process is essentially a basis transformation it is obvious that the DFT is component consistent in the following sense:

$$\text{If } z = \sum_{i=0}^{M} \alpha_i D_i(x), \text{ then } D_n(z) = \alpha_n D_n(x) \quad \text{for each } n = 0, 1, \ldots, M.$$

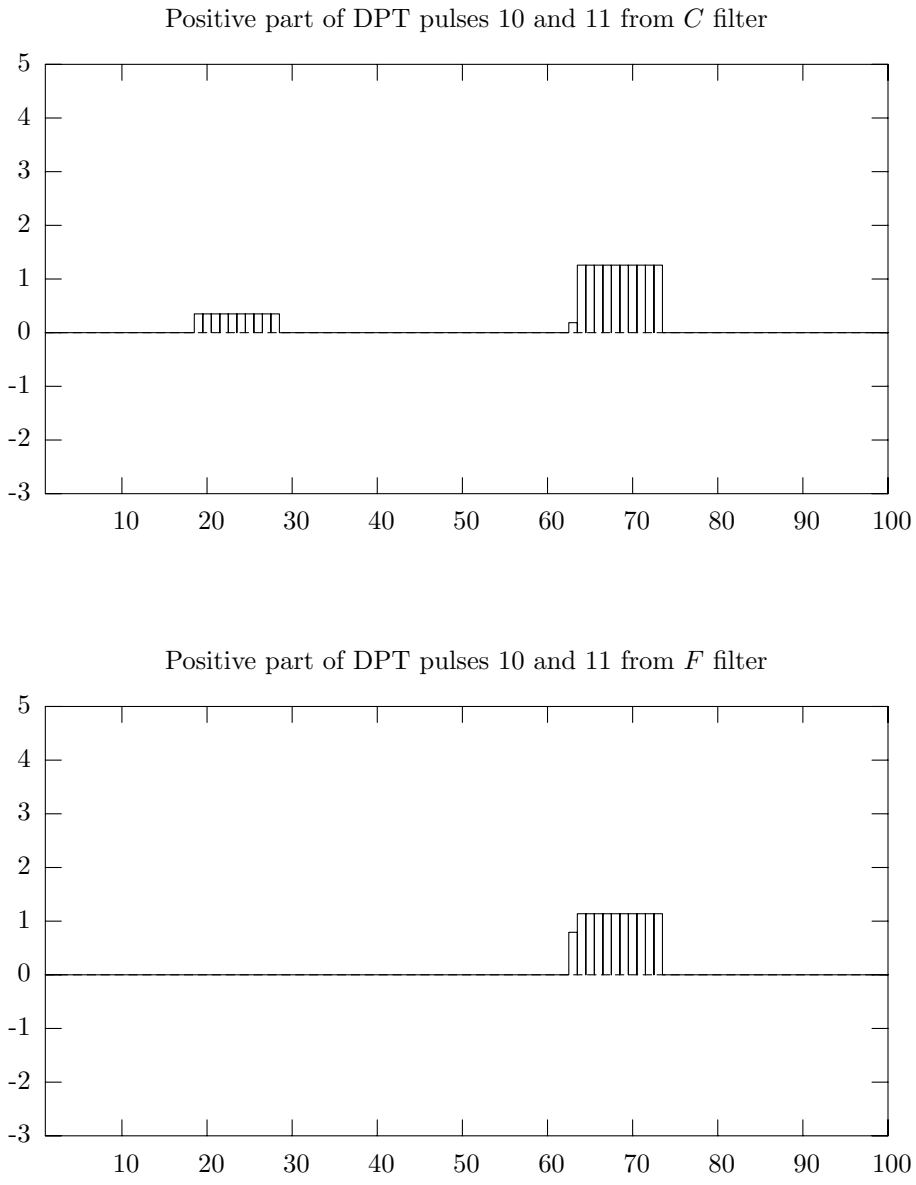


FIG. 4. *Positive pulses extracted from DPT filters applied to the time series in Figure* 1. *The pulses of length* 10 *and* 11 *have been added together.*

Furthermore, the orthonormal basis yields the so-called *energy preservation law*

$$\|x\|_2^2 = \sum_{i=0}^{M} \|D_i(x)\|_2^2,$$

so that it is possible to allocate a percentage of the energy in x to each frequency level.

In the same way, in deriving the DPT we aim to decompose a sequence into a sum of component sequences, but replacing periodic components with sequences consisting of pulses of length n at each resolution level n. From this point of view, the DPT proposed here can be seen as a mapping of a sequence x onto a vector of $M+1$ component sequences, $r^{(n)} = D_n(x)$, representing the different resolution levels, such that

$$x = D_1(x) + D_2(x) + \cdots + D_M(x) + D_\infty(x),$$

which we write as

$$D(x) = DPT(x) = [D_1(x), D_2(x), \ldots, D_M(x), D_\infty(x)]. \tag{2}$$

The final resolution level is numbered D_∞ rather than D_{M+1}, since the final sequence is constant and equal to zero if $x \in \ell_1$, which we shall assume.

There are two equivalent natural primary choices for such a decomposition procedure, which we call $\mathcal{C}$ and $\mathcal{F}$, starting respectively from one of the smoothers L_nU_n or U_nL_n that will be defined below. To avoid saying everything twice, we describe only the case where $DPT(x) = \mathcal{C}(x)$.

The decomposition proceeds recursively. In the first step, the operator L_1U_1 is applied to x to yield a sequence L_1U_1x smoother than x; the residual $(I - L_1U_1)x = D_1(x)$, where I is the identity operator, is the component of x at the finest resolution level. Thus L_1U_1 is not only a smoother, but also a separator (these terms will be defined precisely later). The smoothed part L_1U_1x is then separated by L_2U_2 to yield the second resolution component $r^{(2)} = D_2(x) = (I - L_2U_2)L_1U_1x$ and the smoothed part $L_2U_2L_1U_1x$, etc. The process is continued until only a constant sequence $D_0(x)$ remains.

Note that the DPT components are numbered in order of increasing smoothness, whereas the DFT components are numbered in order of decreasing smoothness.

The theory holds with certain changes if the periodicity condition is replaced by defining $x_i = x_1$ for $i < 1$ and $x_i = x_N$ for $i > N$: such sequences are *ultimately constant*. For example, the lowest resolution component $D_0(x)$ will then be monotone, but constant only when $x_1 = x_N$. For the sake of clarity of exposition, we will restrict ourselves in this paper to the periodic case.

In the course of the paper we shall refine this preliminary formulation of the DPT as more and more of its properties are discovered.

4. Properties of the *LULU* operators. Clearly the properties of the DPT depend crucially on those of the $LULU$ smoothers L_nU_n and U_nL_n. The properties listed in this section have been explored in previous papers [Roh89, Roh99, Roh02a, Roh02c, Roh03, Roh02b, RT91, RW02] and will not be rederived here. To start with, we need some definitions.

DEFINITION 1. *The L_n and U_n (mnemonic: lower and upper) operators are given by*

$$(U_nx)_i = \min\{\max\{x_{i-n}, \ldots, x_i\}, \ldots, \max\{x_i, \ldots, x_{i+n}\}\},$$
$$(L_nx)_i = \max\{\min\{x_{i-n}, \ldots, x_i\}, \ldots, \min\{x_i, \ldots, x_{i+n}\}\}.$$

DEFINITION 2. *The sequence x is* n-monotone *if either*

$$x_i \leqslant x_{i+1} \leqslant \cdots \leqslant x_{i+n} \leqslant x_{i+n+1}$$

or

$$x_i \geqslant x_{i+1} \geqslant \cdots \geqslant x_{i+n} \geqslant x_{i+n+1}$$

for all values of i such that x_i and x_{i+n+1} are both members of the sequence. The set of all n-monotone sequences is denoted by $\mathcal{M}_n$.

$\mathcal{M}_0$ is just the set of all sequences. Note that

$$\mathcal{M}_0 \supset \mathcal{M}_1 \supset \mathcal{M}_2 \supset \cdots, \tag{3}$$

and that the set $\mathcal{M}_n$ is not a vector space except when $n = 0$.

DEFINITION 3. *An operator A is* idempotent *if $A^2 = A$ and* coidempotent *if $I - A$ is idempotent.*

Note that the condition $A(I - A) = 0$ (we do not use a different notation than 0 for the zero operator) is equivalent to coidempotence, even when A is not a linear operator.

DEFINITION 4. *An operator A is a* smoother *if $AE = EA$, where $Ex_i = x_{i+1}$, $A(x + b) = Ax + b$ for all constant sequences b, and $A(cx) = c(Ax)$ for all scalars $c \geqslant 0$.*

In other words, a smoother is location invariant and scale invariant. Our definition differs from that of Mallows [Mal80] in disallowing negative scale factors.

DEFINITION 5. *A smoother A is a* separator *if it is both idempotent and coidempotent.*

DEFINITION 6. *When $x \in \mathcal{M}_0$ and $y \in \mathcal{M}_0$, $x \geqslant y$ means that $x_i \geqslant y_i$ for each i; when A and B are operators on a sequence, $A \geqslant B$ means that $Ax \geqslant Bx$ for all $x \in \mathcal{M}_0$.*

DEFINITION 7. *The nth* median smoother M_n *is given by*

$$(M_n x)_i = \text{median}\{x_{i-n}, \ldots, x_i, \ldots, x_{i+n}\}.$$

DEFINITION 8. *An operator A is* syntone *if $Ax \leqslant Ay$ whenever $x \leqslant y$. Serra* [Ser84] *uses the term "monotone" for this property.*

DEFINITION 9. *An operator A is* neighbor trend preserving *(NTP) if for each sequence x,*

$$x_i \geqslant x_{i+1} \implies (Ax)_i \geqslant (Ax)_{i+1},$$
$$x_i \leqslant x_{i+1} \implies (Ax)_i \leqslant (Ax)_{i+1}.$$

DEFINITION 10. *An operator A is* fully trend preserving *(FTP) if A is NTP and*

$$|(Ax)_i - (Ax)_{i+1}| \leqslant |x_i - x_{i+1}|.$$

DEFINITION 11. *An operator A is* total variation preserving *if $T(x) = T(Ax) + T(x - Ax)$ for each x, where*

$$T(x) = \sum |x_{i-1} - x_i|$$

is the total variation *of the sequence x. In the case of a periodic sequence, the sum is taken over one period.*

It can be shown [Roh03] that the property of total variation preservation is equivalent to that of being FTP.

DEFINITION 12. *The C_n and F_n operators (mnemonic: ceiling and floor) are given by*

$$C_0 = L_0 U_0 = I = U_0 L_0 = F_0,$$
$$C_{n+1} = L_{n+1} U_{n+1} C_n, \qquad F_{n+1} = U_{n+1} L_{n+1} F_n.$$

Here, then, is a list of properties of the $LULU$ operators, referred to below as property 1, etc.:

1. For each integer n, $U_n x = L_n x = x$ whenever $x \in \mathcal{M}_n$ [Roh89].
2. $L_n \leqslant U_n L_n \leqslant C_n \leqslant F_n \leqslant L_n U_n \leqslant U_n$.
3. For each integer n, the operators $L_n U_n$ (and $U_n L_n$) map onto $\mathcal{M}_n$ [Roh89].
4. $U_n U_k = U_m$ and $L_n L_k = L_m$, where $m = \max\{n, k\}$ [RW02].
5. $L_n U_n$ (and $U_n L_n$) are idempotent and coidempotent, and therefore are separators [Roh99].
6. $U_n L_n \leqslant M_n \leqslant L_n U_n$ [Roh89].
7. $L_n U_n$ (and $U_n L_n$) are syntone operators [Roh89].
8. $L_n U_n$ (and $U_n L_n$) are NTP operators [Roh02c].
9. $L_n U_n$ (and $U_n L_n$) are FTP operators [Roh02c].
10. U_n and L_n are variation preserving.
11. $(I - L_{n+1}U_{n+1})L_n U_n x \in \mathcal{M}_n$ (and $(I - U_{n+1}L_{n+1})L_n U_n x \in \mathcal{M}_n$) for each $x \in \mathcal{M}_0$.
12. The operators L_n and U_n are duals in that $U_n(-x) = -L_n(x)$ [Roh89].
13. $U_n(x + c) = U_n x + c$ (and $L_n(x + c) = L_n x + c$) for any constant sequence c [Roh89].
14. $U_n(\alpha x) = \alpha U_n(x)$ (and $L_n(\alpha x) = \alpha L_n x$) for any $\alpha \geqslant 0$ [Roh89].
15. F_n and C_n are separators.
16. If $x \in \mathcal{M}_{n-1}$ and A is FTP, then $U_n A U_n x = A U_n x$ and $L_n A L_n x = A L_n x$. Also $C_j C_k = C_{\max\{j,k\}}$ and $F_j F_k = F_{\max\{j,k\}}$.

5. How does one define consistency? For the DPT to be of practical (or theoretical) use, the heuristics that lead to its construction have to be supported by consistency in behavior and properties. But what do we mean by "consistency"?

In the case of a linear decomposition like the DFT, which is just a basis transformation, there is only one way to define consistency. The decomposition should recover the coordinates, given any linear transformation of basis vectors.

In the case of a nonlinear decomposition, it is not at all obvious what we should aim for. There is no vector space and therefore no basis. A level of consistency that is certainly out of our reach is that one of the two properties of linear operators might hold in general. If one checks whether $DPT(x + y) = DPT(x) + DPT(y)$ with some random sequences, a counterexample will be found quickly. Consistent scaling (i.e., $DPT(\alpha x) = \alpha DPT(x)$) is true (by property 14) when $\alpha \geqslant 0$ but not when $\alpha < 0$. Nevertheless, this rather weak property of consistent nonnegative scaling contains the germ of a strong consistency property.

The concept of consistency of a two-level decomposition can be made unambiguous. If an operator S is to separate a sequence into a "signal" Sx and (additive) "noise" $(I - S)x$, then consistency means that S and $(I - S)$ applied to the components separately should leave that component unchanged. For example, an idempotent linear operator (i.e., a projection) is a consistent separator: $S(Sx) = Sx$ and $(I - S)Sx = 0$. Similarly $(I - S)((I - S)x) = (I - S)x$ as a projection is linear and $S(I - S)x = 0$.

Nonlinear operators are much more intractable: the well-known median operators are not even idempotent. Idempotence by itself gives only consistent signal extraction (i.e., $S(Sx) = Sx$), not consistent noise extraction: $S(I - S)x$ does not necessarily equal $Sx - S^2x$. For consistent separation, both idempotence and coidempotence (the idempotence of $(I - S)$) are required; i.e., the operator must be a separator (Definition 5). A separator can be shown [Roh99] to have only two eigenvalues 1 and 0, with eigensequences corresponding respectively to signal and noise. Since $L_n U_n$ and $U_n L_n$ are separators (property 5), we hope to find some nontrivial consistency

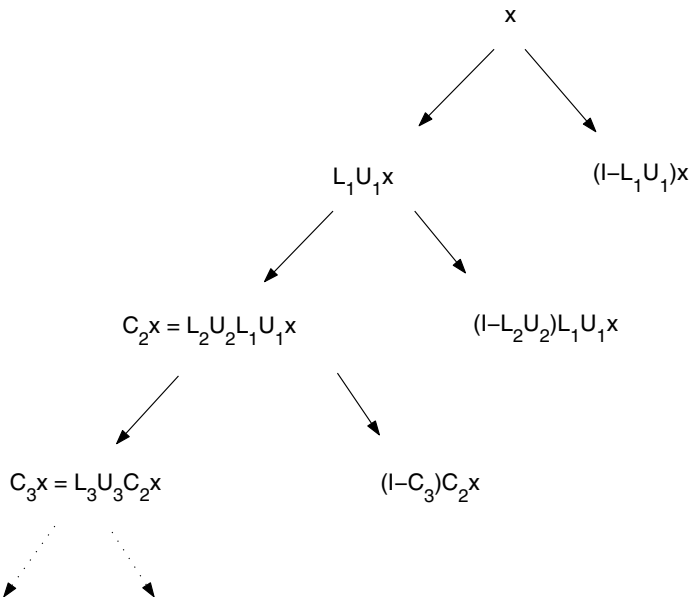


FIG. 5. *A schematic representation of the sequential separation in the DPT.*

properties of the DPT.

But first some words of caution.

Although $\mathcal{M}_n$ is exactly the set of eigensequences with respect to 1 of both L_nU_n and U_nL_n, it is not a subspace, since the sum of two sequences in $\mathcal{M}_n$ need not be in $\mathcal{M}_n$. Moreover (and this may come as a surprise), the operators $I - L_nU_n$ and $I - U_nL_n$ do not share their eigensequences w.r.t. the eigenvalue 0. In other words, L_nU_n and U_nL_n do not always agree on what is noise, although they agree on what is signal, since $L_nU_n(U_nL_n) = U_nL_n$ and $U_nL_n(L_nU_n) = L_nU_n$.

In general, $\mathcal{C}(x)$ and $\mathcal{F}(x)$ are not the same. This is true even when the input is only a single resolution component: if the nth resolution component $r^{(n)}$ of $\mathcal{C}(x)$ is fed to the $\mathcal{F}$ decomposition, then it will in general not be its own nth resolution component!

All of this makes the following innocent question highly problematic:

> When $\mathcal{C}(x)$ or $\mathcal{F}(x)$ is viewed as separating a given sequence x into a number of component sequences at different resolution levels, what should "consistency" mean?

Even so modest a demand as that when an individual resolution component $r^{(n)}$ of x is fed to the same DPT, its decomposition should yield $r^{(n)}$ at level n and zero at the other levels, is not obvious.

At this stage it is illuminating to consider the DPT in Figure 5, to illustrate the sequential separation by the separators L_nU_n.

In [Roh03] it is shown that the sequence $r^{(n)}$ at resolution level n consists of a sum of pulses of length n, far enough apart so that $L_nU_nr^{(n)} = 0$. In fact, [Roh03] gives a stricter characterization, but the latter property is all that is required here. An example of a randomly generated sequence x decomposed into the highest three resolution levels and the rest is depicted in Figure 6, to illustrate the concepts used.

The example demonstrates some essential features of what is achieved in prac-

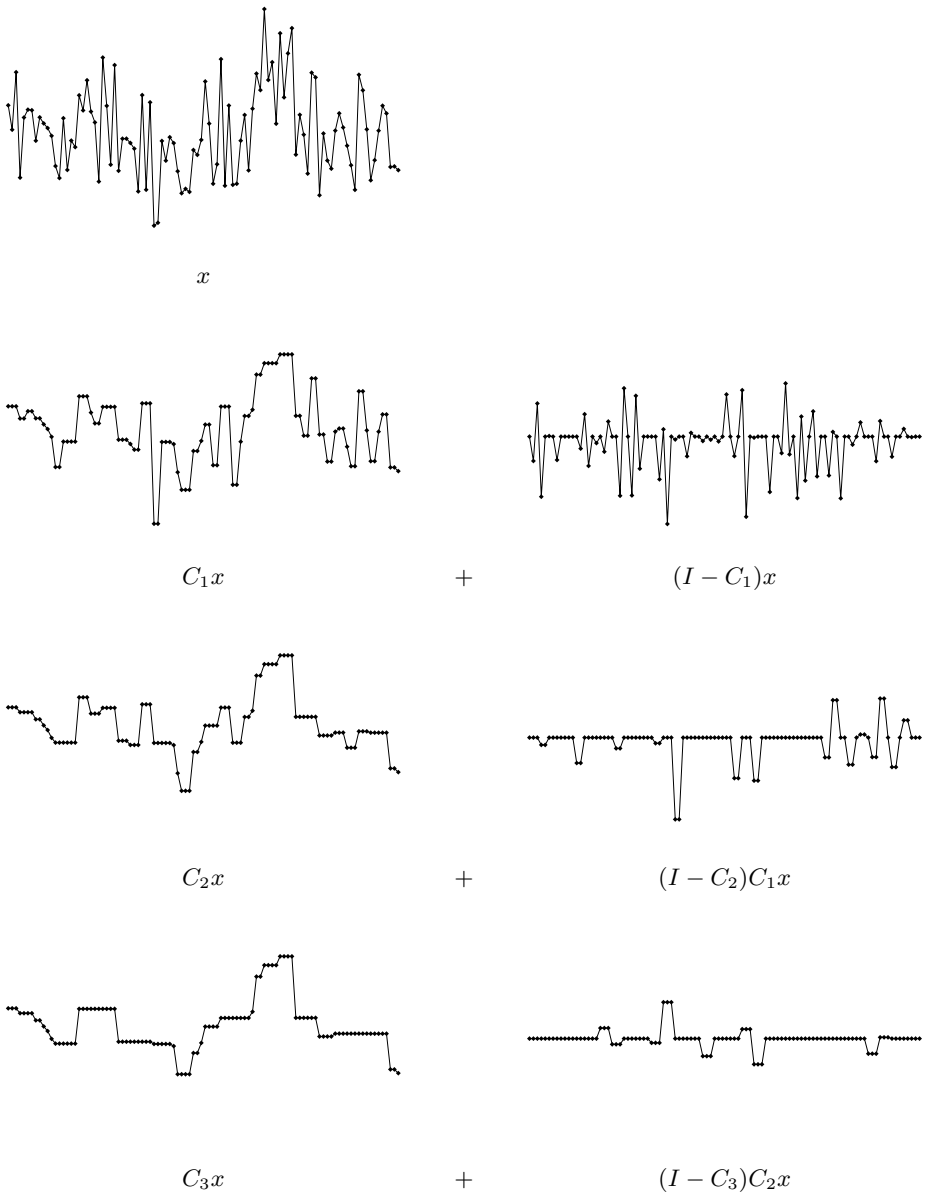


FIG. 6. *The sequence of Figure* 1 *decomposed into the three highest resolution components.*

tice. A sequence is decomposed into sequences of different resolution levels. $\mathcal{F}$ yields similar-looking decompositions that are usually very close to those obtained by $\mathcal{C}$. In Figure 3 we show both DPT transformations to illustrate the general observation that the differences are mainly due to, and proportional to, the random noise present.

It is necessary to proceed very carefully. Some of the theorems in the next sections may look very innocuous to an eye trained on linear operators, but each is an important step on the way to a remarkable consistency that to our knowledge is not shared generally by decomposition based on nonlinear morphological operators.

6. Basic consistency properties. We put

$$D_n x = (C_{n-1} - C_n)x = (I - L_n U_n)C_{n-1}x \tag{4}$$

in (2) to fix the definition of the DPT; in other words, we work with the $\mathcal{C}(x)$ form of the DPT. In view of the duality property (property 12), corresponding results and proofs hold also for $\mathcal{F}$.

Most of the rest of this article is devoted to answering the question of how to define consistency. We start with the most modest of aims, and each time that we achieve our aim the goalposts are shifted, until we reach a conjecture that as yet we cannot prove.

For a start, we demand that a signal consisting of only the output of a single resolution level should be reproduced by the DPT at that level, with zero components elsewhere. This is easy to show.

THEOREM 1. *Let the decomposition DPT be defined by*

$$DPT(x) = [D_1 x, D_2 x, \ldots, D_N x, D_0 x].$$

Then $D_j(D_i x) = \delta_{ij} D_i x$, where δ_{ij} is the Kronecker delta.

Proof. Let $n \geqslant 1$ and $z = D_n x$. Then $z \in \mathcal{M}_{n-1}$ by (4) and property 11, which by (3) and property 1 implies that $L_k U_k z = z$ for $k = 1, 2, \ldots, n-1$, and therefore that $C_k z = z$ for $k = 1, 2, \ldots, n-1$. It follows from (4) that $D_k z = (I - C_k)z$ for $k = 1, 2, \ldots, n$, giving $D_k z = 0$ for $k = 1, 2, \ldots, n-1$ and $D_n z = z - C_n z$. But since C_n is coidempotent, we have $C_n(I - C_n) = 0$, by property 15 and hence $C_n z = C_n(I - C_n)C_{n-1}x = 0$. Thus $D_n z = z$. Since $D_k z$ for $k > n$ is obtained by applying operators to $C_n z$, it follows that the higher resolution levels are all zero. ∎

Thus individual resolution components of a sequence are consistently decomposed. We now ask more: is the DPT consistent as a "low-pass" smoother? In other words, if z consists of all but the first n components of x, does the DPT applied to z yield the same later components as when applied to x? This is also easy.

THEOREM 2. *Let $r^{(m)} = D_m(x)$ and $z = x - \sum_{i=1}^{n} r^{(i)}$. Then*

$$D_m(z) = \begin{cases} 0 & \text{for } m < n, \\ r^{(m)} & \text{for } m \geqslant n. \end{cases}$$

Proof. The same argument as in the proof of Theorem 1 gives $D_k z = 0$ for $k = 1, 2, \ldots, n-1$. Thus, $C_{n-1} z = C_{n-1} x$, so that the n-stage of the DPT of z has the same input as the nth stage of the DPT of x. ∎

Emboldened by these quick successes, we now ask whether the DPT is also consistent as a high-pass smoother; i.e., if the input consists of the sum of only the first n resolution levels, are these levels reproduced unchanged?

Such a result cannot be derived in the same way as Theorem 2. It is true that when x has only two nonzero resolution levels, $r^{(1)} = D_1 x$ and $r^{(2)} = D_2 x$, since if we omit either of the resolution levels to form a sequence z, the decomposition would by Theorem 1 be consistent with that of x. But if there are more than two nonzero resolution levels, it is not so simple, as the following example shows.

Assume $z = r^{(1)} + r^{(2)} = x - r^{(3)}$. Then

$$D_1 z = (I - L_1 U_1)(I - L_1 U_1)x + (I - L_2 U_2)(C_1 x),$$

and since neither L_1U_1 nor $I - L_1U$ is linear, it is not clear that $D_1z = D_1x = r^{(1)}$. This seems to hinge on the fact that L_1U_1 acts linearly on the particular sum $r^{(1)} + r^{(2)} = x - C_2x = (I - C_2)x$. Omitting the resolution component $r^{(2)}$ yields similar problems in the proof of the consistent decomposition of the partial sum. Since experimentation suggests consistency of this type, there must exist some theory not yet explored. The observed consistency seems too good not to be true!

The clue to further development of the theory is suggested by a remarkable property, shared by any composition of different operators L_n and U_n [Roh02c]. The following lemma for **ftp** operators (which by property 9 includes all the $LULU$ operators) is as important as it may be surprising. The proof from [Roh02c] is repeated here to illustrate the concepts involved.

LEMMA 1. (a) *If A is FTP, then $I - A$ is FTP.*

(b) *If A and B are FTP, the composition, or any convex combination of A and B, is also FTP.*

Proof. For each sequence x, consider any index i. Suppose first that $x_i \geqslant x_{i+1}$. Since A is FTP, $Ax_i \geqslant Ax_{i+1}$ and $|Ax_i - Ax_{i+1}| \leqslant |x_i - x_{i+1}| = x_i - x_{i+1}$. Therefore $Ax_i - Ax_{i+1} \leqslant x_i - x_{i+1}$, giving $x_i - Ax_i \geqslant x_{i+1} - Ax_{i+1}$, so that $(I - A)x_i \geqslant (I - A)x_{i+1}$; in other words, $(I - A)$ is NTP. Then $|(I - A)x_i - (I - A)x_{i+1}| = (I - A)x_i - (I - A)x_{i+1} = x_i - x_{i+1} - (Ax_i - Ax_{i+1}) \leqslant x_i - x_{i+1} = |x_i - x_{i+1}|$, which shows that $(I - A)$ is FTP. The rest is easy.

A similar argument holds if $x_i \leqslant x_{i+1}$, completing the proof of (a). The proof of (b) is simple and is left as an exercise for the reader. ∎

COROLLARY 1. *A is FTP $\iff$ $I - A$ is FTP; a convex combination of any finite number of FTP operators is FTP; and αA is NTP if $0 \leqslant \alpha$.*

It is interesting to note that the only linear smoothers that have full trend preservation are the trivial ones I and O. This may explain the fact that the concept of FTP operators is unfamiliar.

In order to progress further, we need two results proved in [Roh03].

RESULT 1. *If $x \in M_{n-1}$ and A is FTP, then*

$$U_n(I - AU_n)x = U_nx - U_nAU_nx = U_nx - AU_nx,$$
$$L_n(I - AL_n)x = L_nx - L_nAL_nx = L_nx - AL_nx.$$

RESULT 2. *$C_j(I - C_n) = C_j - C_n$ and $F_j(I - F_n) = F_j - F_n$ for all $j \leqslant n$.*

Note that the expressions appearing in Result 2 are clearly equal to 0 if $j \geqslant n$, thus proving the coidempotence of F_n and C_n.

THEOREM 3. *Let $z = D_m(x) + D_{m+1}(x) + \cdots + D_n(x) = (C_{m-1} - C_n)x$. Then z decomposes consistently.*

Proof. The same argument as in the proof of Theorem 1 gives $D_jz = 0$ for $j = 1, 2, \ldots, m - 1$. For $m \leqslant j \leqslant n$, we have

$$C_j(z) = (I - C_n)C_jC_{m-1}x = (I - C_n)C_j(x)$$

and therefore, since $C_kC_l = C_{\max\{k,l\}}$,

$$\begin{aligned} D_j(z) &= (C_{j-1} - C_j)z \\ &= (I - C_n)C_{j-1}x - (I - C_n)C_jx \\ &= C_{j-1}x - C_jx - C_nC_{j-1}x + C_nC_jx \\ &= C_{j-1}x - C_jx \\ &= D_j(x). \end{aligned}$$

Finally $C_n(z) = C_n(I - C_n)C_{m-1}x = 0$, since $C_n(I - C_n) = 0$ by the coidempotence of C_n. Thus the components of z come out fully consistently. □

This theorem shows that the DPT is consistent as a "high-pass" smoother (the case where $m = 1$) and as a "band pass" smoother (general $m < n$).

7. Pseudolinearity. The difficulty in proving stronger consistency results comes as might be expected from the nonlinear nature of the filters used. We now establish some properties of the filter operators that resemble those of linear operators, but only in certain special circumstances.

The first of these theorems describes the interaction between U_n (or L_n) and NTP operators when applied to sequences known to be $(n-1)$-monotone.

THEOREM 4. *Let $x \in \mathcal{M}_{n-1}$, and A, B be NTP. Then for all $\alpha, \beta \geqslant 0$ the following hold:*

(a) $U_n(\alpha A + \beta B U_n)x \geqslant \alpha U_n A x + \beta B U_n x$;
$L_n(\alpha A + \beta B L_n)x \leqslant \alpha L_n A x + \beta B L_n x$.

(b) *If $AU_n x = U_n A x$, then $U_n(\alpha A + \beta B U_n)x = \alpha U_n A x + \beta B U_n x$.*

(c) *If $AL_n x = L_n A x$, then $L_n(\alpha A + \beta B L_n)x = \alpha L_n A x + \beta B L_n x$.*

Proof. (a) With the notation $U = U_n$, let $x \in \mathcal{M}_{n-1}$ and i be an index. Consider first the case when $Ux_i = x_i$. Then there is an index $j \in [i-n, i]$ such that $\max\{x_j, \dots, x_{j+n}\} \leqslant x_i$. However, the subsequences $\{x_j, \dots, x_i\}$ and $\{x_i, x_{j+n}\}$ are monotone. Thus

$$x_j \leqslant x_{j+1} \leqslant \cdots \leqslant x_{i-1} \leqslant x_i \geqslant x_{i+1} \cdots \geqslant x_{j+n}.$$

$A, BU, UA, UBU, \alpha A + \beta B$, and $U(\alpha A + \beta BU)$ all transfer these inequalities to their respective outputs, since they are NTP. Therefore

$$\begin{aligned} UAx_i &= Ax_i, \\ UBUx_i &= BUx_i, \\ U(\alpha A + \beta BU)x_i = \alpha Ax_i + \beta BUx_i &= \alpha UAx_i + \beta UBUx_i. \end{aligned}$$

This establishes the first part of (a) in the case where $Ux_i = x_i$.

The other possibility is that $Ux_i \neq x_i$ (and therefore $Ux_i > x_i$). Since each of the sets $\{x_j, x_{j+1}, \dots, x_{j+n}\}$ containing x_i has a maximum larger than x_i (from the definition of U_n) and the sequence x is $(n-1)$-monotone, there is a $j \in [i-n, i]$ such that

$$x_{j-1} > x_j = x_{j+1} = \cdots = x_i = \cdots = x_{j+n-1} < x_{j+n},$$

so that $Ux_j = \min\{x_{j-1}, x_{j+n}\}$. But since U is NTP,

$$Ux_{j-1} \geqslant Ux_j = Ux_{j+1} = \cdots = Ux_i = \cdots = Ux_{j+n-1} \leqslant Ux_{j+n},$$

and we obtain $Ux_i = \min\{x_{j-1}, x_{j+n}\}$.

All the operators in question are NTP so that again the inequalities are inherited by the outputs and

$$\begin{aligned} UAx_i &= \min\{Ax_{j-1}, Ax_{j+n}\}, \\ UBUx_i &= \min\{BUx_{j-1}, BUx_{j+n}\} = BUx_i \qquad \text{since } B \text{ is NTP.} \end{aligned}$$

Therefore

$$\begin{aligned} U(\alpha A + \beta BU)x_i &= \min\{(\alpha A + \beta BU)x_{j-1}, (\alpha A + \beta BU)x_{j+n}\} \\ &= \min\{\alpha Ax_{j-1} + \beta BUx_{j-1}, \alpha Ax_{j+n} + \beta BUx_{j+n}\}. \end{aligned}$$

Now $\alpha UAx_i + \beta BUx_i \leqslant \alpha Ax_{j-1} + \beta BUx_{j-1}$ and $\alpha UAx_i + \beta BUx_i \leqslant \alpha Ax_{j+n} + \beta BUx_{j+n}$. The inequality therefore also holds for the minimum of the two, proving that $\alpha UAx_i + \beta BUx_i \leqslant U(\alpha A + \beta B)x_i$. A similar argument proves the other part.

(b) Noting that $UAx_i = \min\{Ax_{j-1}, Ax_{j+1}\}$, assume that $Ax_{j-1} \leqslant Ax_{j+1}$ so that $UAx_i = Ax_{j-1}$ (a similar argument holds if the other is smaller). Then $Ux_{j-1} = Ux_j = \cdots = Ux_i = \cdots = Ux_{j+n-1} \leqslant Ux_{j+n}$. Since A and B are NTP the equalities are inherited by AUx and BUx, so that $AUx_i = AUx_{j-1} \leqslant AUx_{j+n}$ and $BUx_i = BUx_{j-1} \leqslant BUx_{j+n}$, so that

$$U(\alpha A + \beta BU)x_i \leqslant \alpha Ax_{j-1} + \beta BUx_{j-1} \leqslant \alpha UAx_{j-1} + \beta BUx_i.$$

If now $UAx_{j-1} = AUx_{j-1}$, then since $AUx_{j-1} = AUx_i$, we get

$$U(\alpha A + \beta BU)x_i \leqslant \alpha AUx_i + \beta BUx_i = \alpha UAx_i + \beta BUx_i.$$

This, together with the inequality of part (a) of the proof yields the result $U(\alpha B + \beta BU)x_i = \alpha UAx_i + \beta BUx_i$, which proves part (b) of the theorem.

(c) A similar proof to the above, or using a duality argument, yields the required equality. □

The next result says that U_n behaves like a linear operator when applied to a positive linear combination of the signal and noise obtained by separating an $(m-1)$-monotone sequence with L_nU_n. For completeness, we state the corresponding result with L_n and U_n interchanged.

THEOREM 5. *Let $x \in \mathcal{M}_{n-1}$ and $\alpha, \beta \geqslant 0$. Then*

$$U_n(\alpha(I - L_nU_n) + \beta L_nU_n)x = \alpha U_n(I - L_nU_n)x + \beta L_nU_nx,$$
$$L_n(\alpha(I - U_nL_n) + \beta U_nL_n)x = \alpha L_n(I - U_nL_n)x + \beta U_nL_nx.$$

Proof. L_n and $I - L_nU_n$ are FTP and thus NTP. By result 1, $I - L_nU_n$ commutes with U_n, and by Theorem 4(b) the first equality holds. A similar proof holds for the other equality. □

Now we prove that under the same conditions, L_nU_n also acts like a linear operator.

THEOREM 6. *Let $x \in \mathcal{M}_{n-1}$ on $\alpha, \beta \geqslant 0$. Then $L_nU_n(\alpha(I - L_nU_n) + \beta L_nU_n) = \beta L_nU_n$ and $U_nL_n(\alpha(I - U_nL_n) + \beta U_nL_n) = \beta U_nL_n$.*

Proof.

$$\begin{aligned}
&L_n(U_n(\alpha(I - L_nU_n) + \beta L_nU_n))x \\
&= L_n(\alpha U_n(I - L_nU_n) + \beta L_nU_n)x && \text{by Theorem 5} \\
&= L_n(\alpha(U_n - L_nU_n) + \beta L_nU_n)x && \text{by Theorem 4} \\
&= L_n(\alpha(I - L_n) + \beta L_n)U_nx \\
&= (\alpha L_n(I - L_n) + \beta L_n)U_nx && \text{by Theorem 4} \\
&= \beta L_nU_nx && \text{by the co-idempotence of } L_n.
\end{aligned}$$

A similar proof holds for the other equality. □

8. Strong consistency properties. We are now in a position to prove the following strong result: *the DPT acts like a basis transformation when the input is a linear combination of outputs from another application of the DPT, provided that the coefficients involved are nonnegative.*

THEOREM 7. *Let $x \in \mathcal{M}_0$ and $DPT(x) = [D_1x, D_2x, \ldots, D_Nx, D_0x]$, with $D_0x = C_Nx$. If $\alpha_i \geqslant 0$ and $\sum_{i=1}^{N} \alpha_i = 1$, then $z = \sum_{i=1}^{N} \alpha_i D_i x$ is decomposed consistently.*

Proof. Let $z = \sum_{i=1}^{n} \alpha_i D_i x$. Define $A_m = (\sum_{i=m}^{n} \alpha_i (I - C_i)C_{i-1})$. Since $D_i = (I - C_i)C_{i-1},\ i = 1, 2, \ldots, n$, are all FTP, it follows from Lemma 1 that A_m is FTP. Therefore

$$\begin{aligned}\sum_{i=m}^{n} \alpha_i D_i x &= \left(\sum_{i=m}^{n} \alpha_i (I - C_i)C_{i-1} \right) x \\ &= A_m C_{m-1} x \in \mathcal{M}_{m-1}\end{aligned}$$

since $C_{m-1}x \in \mathcal{M}_{m-1}$. Consider the induction hypothesis that $C_{j-1}z = (\sum_{i=j}^{n} \alpha_i D_i)$ $C_{j-1}x$ for $j < n$. (Clearly it is true for $j = 1$.) The operators $A_k = \sum_{i=k}^{n} \alpha_i D_i$ are all FTP, as they are convex combinations of FTP operators. Thus

$$\begin{aligned}C_j z &= L_j U_j \left(\alpha_j D_j + \left(\sum_{i=j+1}^{n} \alpha_i D_i \right) \right) C_{j-1} x \\ &= L_j U_j \left(\alpha_j (I - L_j U_j) + \left(\sum_{i=j+1}^{n} \alpha_i D_i \right) L_j U_j \right) C_{j-1} x \\ &= L_j (\alpha_j U_j D_j + (A_{j+1}) L_j U_j) C_{j-1} x \qquad \text{by Theorem 4} \\ &= L_j (\alpha_j U_j (I - L_j U_j) + (A_{j+1}) L_j U_j) C_{j-1} x \\ &= L_j (\alpha_j (I - L_j) U_j + (A_{j+1}) L_j U_j) C_{j-1} x \qquad \text{by result 1} \\ &= L_j (\alpha_j (I - L_j) + (A_{j+1}) L_j) U_j C_{j-1} x \\ &= (\alpha_j L_j (I - L_j) + (A_{j+1}) L_j) U_j C_{j-1} x \\ &= 0 + A_{j+1} C_j x.\end{aligned}$$

Thus $D_j z = (C_{j-1} - C_j)z = \alpha_j D_j x$ and $C_j z = (\sum_{i=j+1}^{n} \alpha_i D_i) C_j x$, which completes the induction. □

COROLLARY 2. *Any sequence $\sum_{i=0}^{N} \alpha_i D_i x$, with $\alpha_i \geqslant 0$, is decomposed consistently.*

The $\mathcal{F}$ form of the DPT has a similar consistency, and this can be proved in the same way as above or by using property 12. What is important to note is that the $\mathcal{F}$ and $\mathcal{C}$ forms generally give different resolution components, with a corresponding "interval of ambiguity." When the sequence is strongly correlated, the two decompositions yield essentially similar results, but the addition of additive random noise yields some separation, proportional to the amplitude of this noise. Practical experience and initial theoretical analysis demonstrates that this ambiguity will occur mainly in the first few (highest) resolution levels, becomes small in the middle resolutions, and may appear lower down again, especially if there is a strong specific frequency in the sequence.

An important side effect of the theorems proved here is that we can provide some theoretical support for the popular median decomposition, which hitherto has been motivated purely heuristically. The two DPT processes produce level 1 outputs that bracket the resolution 1 output that would appear out of an equivalent median decomposition, since the median operators involved are in the $LULU$ interval, i.e., $U_n L_n \leqslant M_n \leqslant L_n U_n$ [RT91].

However, at later resolution levels this strict inclusion need not hold. The most we can say is that some order holds on average, since the sequence that is passed on

from the mth separator to the $(m+1)$th separator lies between those of the other two. The median decomposition has in fact almost none of the strong consistency involved in the $LULU$ cases. Consensus is that it "works well" [BMS98]. Our analysis suggests that it can be expected to work well when both $LULU$ transforms work well, and give decompositions close to each other.

9. Consistent one-sided scaling. Although the result of the previous section appears to be what we sought, there is still further room for improvement.

We again turn to the DFT as a source for analogy. In the DFT, each "resolution level" (read: frequency level) is a two-dimensional space of sequences. Since the mapping from a sequence x onto the component sequence $s^{(n)}$,

$$s_i^{(n)} = \alpha_n^- \sin(ni\theta) + \alpha_n^+ \cos(ni\theta),$$

where $\theta = \pi/N$ for some natural number N, simply involves a particular choice of basis for the two-dimensional subspace, it is clear that a sequence consisting of arbitrary linear combinations of the sine and cosine components is decomposed consistently, so that each component comes out with its corresponding amplitude.

The pulse decomposition decomposes a sequence x onto different resolution levels. As noted before, the output at each resolution level is a union of nonoverlapping (in fact, progressively more well-separated) pulses. Here, too, it is possible to think of the output as two-dimensional, given the importance that the sign of the coefficients has in the theorems we have been able to prove.

Define the positive part B^+ and the negative part B^- of an operator B by

$$(B^+x)_i = \begin{cases} Bx_i & \text{if } Bx_i > 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$(B^-x)_i = \begin{cases} Bx_i & \text{if } Bx_i < 0, \\ 0 & \text{otherwise.} \end{cases}$$

At each resolution level, $D_n(x)$ can uniquely be decomposed into a positive and a negative part:

$$D_n(x) = D_n^+(x) + D_n^-(x).$$

Thus D_n^+x contains all the positive pulses, and D_n^-x contains all the negative pulses.

We can thus refine the meaning of "component" so that D_n^+x and D_n^-x are regarded as separate components. This immediately raises the question of whether the DPT is consistent in terms of these components. That is, is it true that for all $\alpha_j^-, \alpha_j^+ \geqslant 0$ the input $z = \sum_{i=1}^n \alpha_i^- D_i^-(x) + \alpha_i^+ D_i^+(x)$ is decomposed consistently to yield $D_n^-(z) = \alpha_n^- D_n^-(x)$ and $D_n^+(z) = \alpha_n^+ D_n^+(x)$ for each $n > 0$?

Theorem 8, our main result, states precisely this. We shall need the following lemma.

LEMMA 2.

$$D_n^- = ((I - L_nU_n)C_{n-1})_- = (I - U_n)C_{n-1},$$
$$D_n^+ = ((I - L_nU_n)C_{n-1})_+ = (I - L_n)U_nC_{n-1}.$$

Proof. The first equation in each case follows from the definition of D_n. Let

$$\hat{D}_n^- = ((I - U_n)C_{n-1}) \quad \text{and} \quad \hat{D}_n^+ = (I - L_n)U_nC_{n-1}.$$

Clearly $\hat{D}_n^- + D_n^+ = D_n$, since

$$(I - U_n)C_{n-1} + (I - L_n)U_nC_{n-1} = (I - L_nU_n)C_{n-1}.$$

Now $(I - U_n)C_{n-1} \leqslant 0$, since $I - U_n \leqslant 0$; and $(I - L_n)U_nC_{n-1} \geqslant 0$, since $I - L_n \geqslant 0$. What is needed to complete the proof is to show that the set of indices i where $\hat{D}_n^- x_i < 0$ and the set of indexes where $\hat{D}_n^+ x_i > 0$ are disjoint.

Let x be a sequence. Then $C_{n-1}x = z$ is in M_{n-1}. Assume that $\hat{D}_n^- z_i < 0$ and $\hat{D}_n^- z_{i-1} = z_{i-1}$. Then

$$(U_n z)_i = \min\{\max\{z_{i-n}, \ldots, z_i\}, \ldots, \max\{z_i, \ldots, z_{i+n}\}\} > z_i.$$

To the left of z_i and to the right of z_i there are $z_\ell, z_r > z_i$ with $|r - \ell| < n + 1$. Since $z \in M_{n-1}$ we have a constant section $z_i = z_{i+1} = \cdots = z_{i+n-1}$, with $U_n z_i = \min\{z_{i-1}, z_{i+n}\} > z_i$. This is because $z_\ell > z_i$ are both in $\{z_{i-n}, \ldots z_i\}$, which is therefore monotone decreasing as $z \in M_{n-1}$. Similarly $\{z_i, \ldots, z_r, \ldots, z_{i+n}\}$ is monotone increasing, and the intersection must be constant. But if $j \in [i, i + n]$, then

$$(L_nU_nz)_j = \max\{\min\{U_nz_{i-n}, \ldots, U_nz_i\}, \ldots, \min\{U_nz_i, \ldots, U_nz_{i+n}\}\};$$

therefore $L_nU_nz_j = U_nz$ and $\hat{D}_n^+ z_j = 0$.

Thus $\hat{D}_n^+ z$ is zero where $\hat{D}_n^-$ is negative, and $\hat{D}_n^+ z$ can be positive only where $\hat{D}_n^- z$ is zero. Thus $\hat{D}_n^- = D_n^-$ and $\hat{D}_n^+ = D_n^+$.

The second part is proved by a generally similar argument, or by appealing to duality. □

Since $(I - U_n)C_{n-1}$ and $(I - L_n)U_nC_{n-1}$ are FTP, the following theorem is easy to prove in analogy with the previous one.

THEOREM 8. *Let $x \in \mathcal{M}_0$ and $DPT(x) = [D_1x, D_2x, \ldots, D_Nx, D_0x]$, with $D_0x = C_N(x)$. Then $z = \sum_{i=1}^N (\alpha_i^- D_n^-(x) + \alpha_i^+ D_n^+(x))$ is decomposed consistently.*

Proof. The proof proceeds as in Theorem 7, with $\alpha_i D_i x$ replaced by $\alpha_i^- D_i^- x + \alpha_i^+ D_i^+ x$, giving

$$C_j z = L_j U_j \left(\alpha_j^- D_j^- + \alpha_j^+ D_j^+ + \sum_{i=j+1}^{N} (\alpha_i^- D_i^- + \alpha_i^+ D_i^+) \right) C_{j-1}(x).$$

Applying U_j first, we get

$$\begin{aligned} U_j z &= U_j \left(\alpha_j^- D_j^- + \left(\alpha_j^+ D_j^+ + \sum_{i=j+1}^{N} \left(\alpha_i^- D_i^- + \alpha_i^+ D_i^+\right) \right) \right) C_{j-1}(x) \\ &= U_j(\alpha_j^-(I - U_j) + (A_{j+1}L_j)U_j)C_{j-1}(x) \end{aligned}$$

with A_{j+1} FTP. This implies by Theorem 4 (since $C_{j-1}(x) \in \mathcal{M}_{j-1}$) that

$$\begin{aligned} U_j z &= (\alpha_j^- U_j(I - U_j) + A_{j+1}L_jU_j)C_{j-1}(x) \\ &= A_{j+1}L_jU_jC_{j-1}(x) \end{aligned}$$

since $U_j(I - U_j) = 0$.

Applying L_j to $U_j z$ yields in a similar fashion that

$$C_j z = \sum_{i=j+1}^{n} (\alpha_i^- D_i^- + \alpha_i^+ D_i^+) C_{j-1}(x).$$

The resolution components that are removed at resolution layer j are therefore $\alpha_j^- D_j^- C_{j-1}(x) = \alpha_j^- D_j^-(x)$ and $\alpha_j^+ Dj^+(x)$, and the induction argument used in Theorem 6 completes the proof. □

10. The pulse structure of a DPT. Each individual component $D_n x$ decomposes into a number of nonoverlapping pulses of width exactly n. Moreover, we can show [Roh02c] that the pulses in $D_n^+ x$ and $D_n^- x$ are separated from each other by at least n zeros.

It is easy to obtain an upper bound on the total number of pulses in the decomposition. The difference sequence of x, i.e., $\{d_i = x_i - x_{i-1},\ i = 1, 2, \ldots, N\}$, obviously contains at most N distinct values. It is shown in [Roh02c] that for every pulse appearing in $D_n x$, the number of distinct values that the difference set of $C_n x$ can take is reduced by at least one. Thus, there can be at most N pulses in total in the entire decomposition.

Once the DPT of x has been computed, it is therefore possible to represent it by $K \leqslant N$ triples $(p_j, w_j, h_j),\ j = 1, \ldots, K$, representing position, width, and height of each pulse. That is,

$$x = \sum_{j=1}^{K} h_j P(p_j, w_j),$$

where

$$P_i(p, w) = \begin{cases} 1, & p \leqslant i < p + w, \\ 0 & \text{elsewhere.} \end{cases}$$

This economical representation in terms of at most $3N$ numbers is important from the point of view of storage and transmission of the data.

The temptation is strong to refine the notion of a component still further, provoking the audacious idea:

> *Can we define each individual pulse in a resolution level as a component and still obtain a consistent decomposition?*

This would give the DPT a consistency comparable to that of a wavelet decomposition, in which each individual wavelet is a member of a basis for a vector space. But that a similar property might hold for a transform as nonlinear as the DPT seems incredible.

Yet, if this idea is correct, the implications for image processing would be enormous. For example, if a square white slab were barely visible against a background that is nearly as white, a pulse decomposition on the rows (or columns) of a matrix of luminosity values obtained from a photo can locate the slab by its expected ratio of width to height. Amplifying this particular pulse would make the slab clearly visible, without distorting any of the surroundings.

To get a feeling for what is meant here, let us look at a simple example. A random sequence x is generated, and one level of the DPT is applied; the components $D_1^- x$ and $D_1^+ x$ are shown in Figure 7. Then each separate pulse is amplified by its own random positive factor. All the resolution levels are then added together again, and

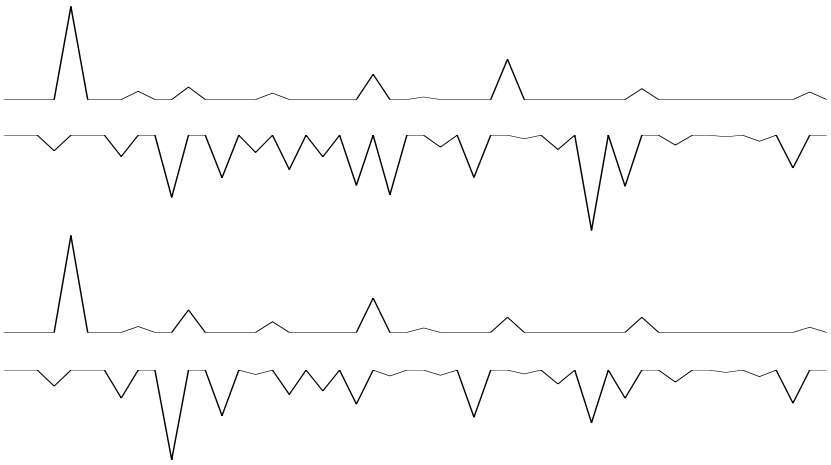


FIG. 7. *The sequences* $(I - U_1)x$ *and* $(I - L_1)U_1x$ *and their modified versions.*

the sum is decomposed. Comparing the first resolution levels demonstrates that the modified sequences have pulses in exactly the same positions.

We have overwhelming evidence from computer simulations that this behavior always occurs, but a general proof eluded us until this article was in revision. We think we now have a proof [LR06] based on a totally different approach to the problem, but the correctness of this proof is still to be checked. Therefore, we do not claim here that this idea is more than a formal conjecture, as follows.

> *Conjecture.* Let $(p_j, w_j, h_j),\ j = 1, \ldots, K$, be pulses in the DPT of x, and let $z = \sum_{j=1}^{K} h'_j P(p_j, w_j)$, where $h_j h'_j \geqslant 0$. Then the pulses in the DPT of z are $(p_j, w_j, h'_j),\ j = 1, \ldots, K$.

Of course, $h'_j = 0$ is understood to mean that the pulse is absent from the DPT of z.

11. Conclusion. We have shown that the discrete pulse transform has many of the properties of consistent component identification that are normally associated with linear transforms like the discrete Fourier transform. In particular, despite its essential nonlinearity, it acts like a basis transformation in a positive cone, with the basis comprising the positive and negative components extracted at each resolution level. This property already has significant potential in the field of image processing.

One area for future research is the characterization problem: what are the conditions under which some set of pulses forms a possible set of basis components? In the DFT and wavelet transforms, the basis is chosen in advance. In the DPT, the basis is constructed by the process itself. We know how to characterize the set of pulse trains that could appear as outputs at any particular resolution level n—pulses of the same sign must be at least n positions from each other—but we do not know how the various levels interact. In fact, the only way we know to find out whether a given set of pulse trains forms a "basis" is to combine them and to analyze that combination by the DPT to see whether they are recovered.

More work is required to establish the conjecture that the true basis components in the DPT are the individual pulses, in the sense that any nonnegative linear combination of them can be recovered by the DPT. Such a result would not only neatly round off the theory, but also provide compelling theoretical backing for the use of the DPT as a device for image enhancement. We are working on a promising idea that may also shed light on the characterization problem.

Acknowledgement. An anonymous referee found many small technical inaccuracies and prompted us to add section 2, which (we must admit) makes the final product more attractive.

REFERENCES

[BMS98] A. Bijaoni, F. Murtagh, and J. L. Starck, *Image Processing and Data Analysis—The Multiple Approach*, Cambridge University Press, Cambridge, UK, 1998.

[Chu97] C. K. Chui, *Wavelets: A Mathematical Tool for Signal Analysis*, SIAM Monogr. Math. Model. Comput. 1, SIAM, Philadelphia, 1997.

[LR06] D. P. Laurie and C. H. Rohwer, *Fast implementation of the discrete pulse transform*, in International Conference on Numerical and Applied Mathematics 2006, T. E. Simos, G. Psihoyios, Ch. Tsitouras, eds., Wiley-VCH, Weinheim, Germany, 2006, pp. 484–487.

[Mal80] C. L. Mallows, *Some theory of nonlinear smoothers*, Ann. Statist., 8 (1980), pp. 695–715.

[Roh89] C. H. Rohwer, *Idempotent one-sided approximation of median smoothers*, J. Approx. Theory, 58 (1989), pp. 151–163.

[Roh99] C. H. Rohwer, *Projections and separators*, Quaest. Math., 22 (1999), pp. 219–230.

[Roh02a] C. H. Rohwer, *Fast approximation with locally monotone sequences*, Suppl. Rend. Circ. Mat. Palermo, Serie II, 68 (2002), pp. 777–790.

[Roh02b] C. H. Rohwer, *Multiresolution analysis with pulses*, in Advanced Problems in Constructive Approximation, M. D. Buhmann and D. A. Mache, eds., Internat. Ser. Numer. Anal. 142, Birkhäuser-Verlag, Basel, 2002, pp. 165–186.

[Roh02c] C. H. Rohwer, *Variation and LULU-smoothing*, Quaest. Math., 25 (2002), pp. 163–176.

[Roh03] C. H. Rohwer, *Fully trend preserving operators*, Quaest. Math., 27 (2004), pp. 217–229.

[RT91] C. H. Rohwer and L. M. Toerien, *Locally monotone robust approximation of sequences*, J. Comput. Appl. Math., 36 (1991), pp. 399–408.

[RW02] C. H. Rohwer and M. Wild, *Natural alternative for one dimensional median filtering*, Quaest. Math., 25 (2002), pp. 135–162.

[Sch67] I. J. Schoenberg, *On spline functions*, in Inequalities, O. Shisha, ed., Academic Press, New York, 1967, pp. 255–291.

[Ser84] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, London, 1984.

[Vel77] P. F. Velleman. *Robust nonlinear data smoothers: Definitions and recommendations*, Proc. Natl. Acad. Sci. USA, 74 (1977), pp. 434–436.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1035–1047

REMARKS ON THE BLOWUP FOR THE L^2-CRITICAL NONLINEAR SCHRÖDINGER EQUATIONS*

TAOUFIK HMIDI† AND SAHBI KERAANI†

Abstract. This paper is dedicated to the formation of singularities of the solutions of L^2-critical nonlinear Schrödinger equations. We prove a refined compactness lemma adapted to the blowup analysis, and we use it to improve the recent result by Colliander et al. [*Math. Res. Lett.*, 12 (2005), pp. 357–375] on the concentration in the L^2-critical nonlinear Schrödinger equation below H^1.

Key words. time dependent Schrödinger equation, blowup, mass concentration

AMS subject classifications. 35Q55, 35B40, 35B05

DOI. 10.1137/050624054

1. Introduction.

We consider the L^2-critical nonlinear Schrödinger equation:

$$\begin{cases} i\partial_t u + \Delta u + |u|^{\frac{4}{d}}u = 0, & x \in \mathbb{R}^d,\ t > 0, \\ u(0,x) = u_0(x). \end{cases} \tag{1.1}$$

Here, $\Delta = \sum_{i=1}^{d} \partial_{x_i}^2$ is the Laplace operator on $\mathbb{R}^d$ and $u_0 : \mathbb{R}^d \longrightarrow \mathbb{C}$. It is well known (see [5], for instance) that the Cauchy problem (1.1) is locally well-posed in H^s for every $s \geq 0$. The unique solution has the following conservation law:

$$\int_{\mathbb{R}^d} |u(t,x)|^2 dx = \int_{\mathbb{R}^d} |u_0(x)|^2 dx. \tag{1.2}$$

Also, if $s \geq 1$, the energy

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

is conserved as t varies. For $s > 0$, (1.1) is subcritical: the lifespan of the solution depends only on the H^s norm of the data. This yields the following blowup alternative: either $T^* = \infty$ or $T^* < +\infty$ and

$$\lim_{t \uparrow T^*} \|u(t,\cdot)\|_{H^s} = +\infty.$$

The space L^2 and the equation have the same scaling. More precisely, if u solves (1.1), then for every $\lambda > 0$, so does $u_\lambda(x,t) = \lambda^{d/2}u(\lambda^2 t, \lambda x)$ with data $u_\lambda(0,x) = \lambda^{d/2}u_0(\lambda x)$. But $\|u_\lambda(0,\cdot)\|_{L^2} = \|u_0\|_{L^2}$, and from this point of view (1.1) is L^2-critical. In this case the situation is more subtle, and the time of existence depends on the shape of the data.

The local theory relies heavily on some integrability properties of the solution of the associated linear Schrödinger equation, called Strichartz estimates. In fact, by

*Received by the editors February 9, 2005; accepted for publication (in revised form) April 3, 2006; published electronically November 21, 2006.
http://www.siam.org/journals/sima/38-4/62405.html
†IRMAR, UMR 6625 du CNRS, Université de Rennes 1, Campus de Beaulieu, 35 042 Rennes cedex, France (thmidi@univ-renne1.fr, sahbi.keraani@univ-rennes1.fr).

using Fourier analysis in connection with the work by Tomas [23], as in [22], or an abstract operators theory as in [9], it was proved that

$$\|e^{it\Delta}u_0\|_{L^{\frac{4}{d}+2}(\mathbb{R}^{d+1})} \leq C\|u_0\|_{L^2(\mathbb{R}^d)}. \tag{1.3}$$

The local solution follows from solving the equivalent integral equation

$$u(t,x) = e^{it\Delta}u_0(x) + i\int_0^t e^{i(t-s)\Delta}|u|^{\frac{4}{d}}u(s,x)ds,$$

by a standard Picard iteration method. The small data theory asserts that there exists a $\delta > 0$ (related to the constant C in (1.3)) such that if

$$\|u_0\|_{L^2(\mathbb{R}^d)} < \delta,$$

the initial value problem (1.1) has a unique global solution. This follows by solving the Cauchy problem (1.1) directly in the whole space (the first step of the iteration method suffices to reach $T^* = \infty$). However, for large data, *blowup* may occur. The blowup or "wave collapse" corresponds to self-trapping of beams in laser propagation. A lot of theoretical and numerical works are dedicated to this subject when the initial data belongs to H^1. In fact, in this space energy arguments apply, and a blowup theory has been developed in the last two decades (see [5], [21], [16] and the references therein). This theory is connected to the notion of ground state: the unique positive radial solution of the elliptic problem

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

In [25], Weinstein exhibited the following refined Gagliardo–Nirenberg inequality:

$$\|\psi\|^{\frac{4}{d}+2}_{L^{\frac{4}{d}+2}} \leq C_d\|\psi\|^{\frac{4}{d}}_{L^2}\|\nabla\psi\|^2_{L^2} \qquad \forall\psi \in H^1, \tag{1.4}$$

with $C_d = \frac{d+2}{d}\|Q\|^{-\frac{4}{d}}_{L^2}$. Combined with the conservation of energy, this implies that $\|Q\|_{L^2}$ is the critical mass for the formation of singularities: for every $u_0 \in H^1$ such that

$$\|u_0\|_{L^2} < \|Q\|_{L^2}$$

the solution of (1.1) with initial data u_0 is global. Also, this bound is optimal. By using the conformal invariance, one constructs

$$S(t,x) = (T^*-t)^{-d/2}e^{[(i/(T^*-t))+(-i|x|^2/T^*-t)]}Q\left(\frac{x}{T^*-t}\right),$$

a solution of (1.1) with $\|u\|_{L^2} = \|Q\|_{L^2}$ that blows up in a finite time T^*. In [14], Merle has proved that $S(t)$ is the unique minimal blowup solution in H^1 in the following sense: let $u_0 \in H^1$ with $\|u_0\|_{L^2} = \|Q\|_{L^2}$, and assume that $u(t)$ blows up at finite time; then $u(t) = S(t)$ up to symmetries of the equation in H^1. It is also proved (see [18] and [24]) that at the blowup there is a concentration phenomenon in the L^2 norm: there exist continuous functions $x(t)$ such that

$$\forall R > 0, \quad \liminf_{t\to T^*}\int_{|x-x(t)|\leq R}|u(t,x)|^2dx \geq \int Q^2. \tag{1.5}$$

For the case $u_0 \in H^s$, with $0 \le s < 1$, the classical energy arguments don't work. Nevertheless, the general consensus is that, in this case too, the same concentration phenomenon happens (the blowup solutions concentrate a minimal amount of mass which is equal to $\|Q\|_{L^2}$). The first result in this direction is due to Bourgain [3] in the case of two dimensions and $u_0 \in L^2$. In fact, by using a refined version of the Strichartz inequality (1.3) proved in [20] and harmonic analysis techniques, this author proved that if u is a blowup solution of (1.1) at finite time $T^* > 0$, then

$$\lim_{t\uparrow T^*} \left(\sup_{y\in\mathbb{R}^2} \int_{\{|x-y|<\sqrt{T^*-t}\}} |u(t,x)|^2 dx \right) > \alpha(\|u_0\|_{L^2}) > 0,$$

with $\alpha(\lambda) \to 0$ as $\lambda \to \infty$. Using the work by Bourgain [3], Merle and Vega [19] have proved, among other things, an asymptotic compactness property in $L^2(\mathbb{R}^2)$ up to the invariance of the equation.

In [11] the second author defined the minimal mass δ_0 as the L^2 norm necessary to ignite a wave collapse[1] and stressed its role in the blowup mechanism. This yields a better description of the blowup solutions of (1.1) in *one* and *two* space dimensions (see also [4]). It is worth noting that these results can be generalized to higher dimensions, thanks to recent work by Bégout and Vargas [2]. However, unlike the H^1-case, in this level of regularity there was no result connecting the blowup mechanism to the ground state Q, and an explicit quantification of the minimal amount of mass concentrated remains an open problem. Recently, very important progress in this direction has been made by Colliander et al. [6]. These authors have proved that blowup solutions, which are radially symmetric, concentrate at least the mass of the ground state, for an intermediate case $1 > s > s_Q$.[2] Their proof is based on the so-called I-method introduced by Colliander et al. (see [7] and the references therein).

In this paper we prove a compactness lemma adapted to the analysis of the blowup phenomenon of the nonlinear Schroödinger equation, and we use it to improve the results of [6]: we remove the assumption of radial symmetry of the initial data, and we prove that Q is a profile for the singular solutions with minimal mass. The main tools of the proof of this compactness lemma are an argument of profile decomposition, introduced by Gérard [8] to study the defect of compactness for Sobolev embedding, and the sharp Gagliardo–Nirenberg inequalities (1.4). We prove the following result.

THEOREM 1.1. *Let $\{v_n\}_{n=1}^\infty$ be a bounded family of $H^1(\mathbb{R}^d)$ such that*

$$\limsup_{n\to\infty} \|\nabla v_n\|_{L^2} \le M \qquad \text{and} \qquad \limsup_{n\to\infty} \|v_n\|_{L^{\frac{4}{d}+2}} \ge m. \tag{1.6}$$

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

$$v_n(\cdot + x_n) \rightharpoonup V \in H^1 \qquad \text{weakly}$$

with $\|V\|_{L^2} \ge (\frac{d}{d+2})^{d/4} \frac{m^{\frac{d}{2}+1}}{M^{d/2}} \|Q\|_{L^2}$.

Remark 1.2. The lower bound on the L^2 norm of V is optimal. In fact, if we take $v_n = Q$, then we get equality.

Remark 1.3. If $d = 2$, then we can interchange the roles of $\|\nabla v_n\|_{L^2}$ and $\|v_n\|_{L^2}$. More precisely, if we assume that $\limsup \|v_n\|_{L^2} \le M$ and that $\limsup \|v_n\|_{L^4} \ge m$,

[1] We have $\delta_0 \le \|Q\|_{L^2}$, and the equality is strongly conjectured.

[2] Some nonoptimal index $\le \frac{1+\sqrt{11}}{5}$.

then we get the same conclusion with $\|\nabla V\|_{L^2} \geq \frac{1}{\sqrt{2}} \frac{m^2}{M} \|Q\|_{L^2}$ instead of the lower bound on the L^2 norm.

Remark 1.4. In the H^1 context this theorem allows us to easily obtain the results on the concentration and uniqueness of the profile of concentration proved by Weinstein [26] using the concentration-compactness lemma of Lions [13]. To see this, take $u_0 \in H^1$ such that the corresponding solution u of (1.1) blows up in finite time $T^* > 0$ and $t_n \uparrow T^*$ as $n \to \infty$. We set

$$v_n(x) = \lambda_n^{d/2} u(t_n, \lambda_n x), \qquad \lambda_n = 1/\|\nabla u(t_n, .)\|_{L^2}.$$

Using conservation of energy, we get trivially that $\{v^n\}_{n=1}^{\infty}$ satisfies the assumptions of Theorem 1.1 with $M = 1$ and $m = (\frac{d+2}{d})^{\frac{d}{2d+4}}$, which implies that

$$\lambda_n^{d/2} u(t_n, \lambda_n(\cdot - x_n)) \rightharpoonup V,$$

with $\|V\|_{L^2} \geq \|Q\|_{L^2}$. This yields, in particular, the concentration estimate (1.5). If we assume, in addition, that $\|u_0\|_{L^2} = \|Q\|_{L^2}$, then the limit above becomes strong in H^1, and the variational characterization of the ground state Q implies the universality of the profile of the blowup solutions with minimal mass. It is worth noting that these arguments are indeed standard, and the novelty is just the use of Theorem 1.1 to avoid discussion of the concentration, vanishing, and dichotomy cases of the concentration-compactness lemma of Lions and then simplifying the proof.

As an application of Theorem 1.1 and the results of [6] we obtain the following result.

THEOREM 1.5. *Assume $d = 2$ and $s > s_Q$. Let $u_0 \in H^s(\mathbb{R}^2)$ be such that the corresponding solution u of (1.1) blows up in finite time $T^* > 0$. Then there exists a sequence $t_n \longrightarrow T^*$ such that the following holds true: there exists a function $V \in H^1$ with $\|V\|_{L^2} \geq \|Q\|_{L^2}$, and a sequence $\{\rho_n, x_n\}_{n=1}^{\infty} \subset \mathbb{R}_+^* \times \mathbb{R}^2$ satisfying*

$$\rho_n \leq A(T^* - t_n)^{s/2}$$

for some $A > 0$, such that

$$\rho_n u(t_n, \rho_n x + x_n) \rightharpoonup V \quad \textit{weakly}.$$

Remark 1.6. We do not know if ψ or $\|\psi\|_{L^2}$ depends on the time sequence $\{t_n\}_{n=1}^{\infty}$. Recently, Merle and Raphael [17] have proved that, in the H^1 context, Q is the universal blowup profile (for the strong $\dot{H}^1$ convergence) for the near-critical mass solutions.

Remark 1.7. Note that, despite the fact that $\{u(t_n, \cdot)\}_{n=1}^{\infty}$ belongs to H^s, the blowup profile V is in H^1. This fact corroborates the expectation that $V = Q$.

Theorem 1.5 and the variational characterization of the ground state allow us to prove the following theorem.

THEOREM 1.8. *Assume $d = 2$ and $s > s_Q$. Let $u_0 \in H^s(\mathbb{R}^2)$ with $\|u_0\|_{L^2} = \|Q\|_{L^2}$ such that the solution u of (1.1) blows up in finite time $T^* > 0$. Then there exists a sequence $t_n \longrightarrow T^*$ satisfying*

$$\rho_n \leq A(T^* - t_n)^{s/2}$$

for some $A > 0$, and a sequence $\{\rho_n, \theta_n, x_n\}_{n=1}^{\infty} \subset \mathbb{R}_+^* \times [0, 2\pi[\times\mathbb{R}^2$ such that

$$\rho_n e^{i\theta_n} u(t_n, \rho_n x + x_n) \longrightarrow Q$$

strongly in $H^{\tilde{s}-}$, *where* $\tilde{s} = \frac{s+1}{4-2s}$.

Remark 1.9. This result is the analogue of the one proved by Weinstein [26] in the H^1 case.[3] However, contrary to our case, the H^1 result of Weinstein holds for all t instead of along the sequence t_n, and the scaling parameter ρ_n is explicitly given by $\rho = \|\nabla Q\|_{L^2}/\|\nabla u(t,.)\|_{L^2}$.

As an application we obtain the next claim.

COROLLARY 1.10. *Under the assumptions of Theorem* 1.5, *let* $\lambda(t) > 0$ *such that* $\frac{(T^*-t)^{s/2}}{\lambda(t)} \longrightarrow 0$ *as* $t \to T^*$. *There exists* $x(t) \in \mathbb{R}^2$ *such that*

$$\limsup_{t\to T^*} \int_{|x-x(t)|\le\lambda(t)} |u(t,x)|^2 dx \ge \int Q^2.$$

Remark 1.11. As already remarked in [6], the fact that Theorems 1.5 and 1.8 hold for only a time sequence $\{t_n\}$ and the lim sup (instead of lim inf) in Corollary 1.10 is due to the lack of information on the monotonicity of the H^s norm of the blowup solutions near the collapse time.

The rest of this paper is structured as follows. In section 2 we prove Theorem 1.1. Section 3 is devoted to the proofs of blowup results.

2. Proof of Theorem 1.1. In what follows we set $2^* = \infty$ if $d = 1, 2$, and $2^* = \frac{2d}{d-2}$ if $d \ge 3$. Theorem 1.1 is a consequence of a profile decomposition of the bounded sequences in H^1 following the work by Gérard [8] (see also [1] and [10]). More precisely, we have the following result.

PROPOSITION 2.1. *Let* $\mathbf{v} = \{v_n\}_{n=1}^{\infty}$ *be a bounded sequence in* $H^1(\mathbb{R}^d)$. *Then there exist a subsequence of* $\{v_n\}_{n=1}^{\infty}$ (*still denoted* $\{v_n\}_{n=1}^{\infty}$), *a family* $\{\mathbf{x}^j\}_{j=1}^{\infty}$ *of sequences in* $\mathbb{R}^d$, *and a sequence* $\{V^j\}_{j=1}^{\infty}$ *of* H^1 *functions such that*

(i) *for every* $k \ne j$, $|x_n^k - x_n^j| \underset{n\to\infty}{\longrightarrow} +\infty$;

(ii) *for every* $\ell \ge 1$ *and every* $x \in \mathbb{R}^d$,

$$v_n(x) = \sum_{j=1}^{\ell} V^j(x - x_n^j) + v_n^\ell(x),$$

with

$$\limsup_{n\to\infty} \|v_n^\ell\|_{L^p(\mathbb{R}^d)} \underset{\ell\to\infty}{\longrightarrow} 0, \tag{2.1}$$

for every $p \in]2, 2^*[$.

Moreover, we have, as $n \to +\infty$,

$$\|v_n\|_{L^2}^2 = \sum_{j=1}^{\ell} \|V^j\|_{L^2}^2 + \|v_n^\ell\|_{L^2}^2 + o(1) \tag{2.2}$$

and

$$\|\nabla v_n\|_{L^2}^2 = \sum_{j=1}^{\ell} \|\nabla V^j\|_{L^2}^2 + \|\nabla v_n^\ell\|_{L^2}^2 + o(1). \tag{2.3}$$

[3] The asymptotic result by Weinstein was completed by the result of Kwong on the uniqueness of the ground state [12].

Proof. Let $\mathcal{V}(\mathbf{v})$ be the set of functions obtained as weak limits of subsequences of the translated $v_n(.+x_n)$ with $\{x_n\}_{n=1}^{\infty}\subset\mathbb{R}^d$. We denote

$$\eta(\mathbf{v})=\sup\{\|V\|_{H^1},\quad V\in\mathcal{V}(\mathbf{v})\}.$$

Clearly

$$\eta(\mathbf{v})\leq\limsup_{n\to\infty}\|v_n\|_{H^1}.$$

We will prove the existence of a sequence $\{V^j\}_{j=1}^{\infty}$ of $\mathcal{V}(\mathbf{v})$ and a family $\{\mathbf{x}^j\}_{j=1}^{\infty}$ of sequences of $\mathbb{R}^d$ such that

$$k\neq j\Longrightarrow|x_n^k-x_n^j|\underset{n\to\infty}{\longrightarrow}\infty,$$

and, up to extracting a subsequence, the sequence $\{v_n\}_{n=1}^{\infty}$ can be written as

$$v_n(x)=\sum_{j=1}^{\ell}V^j(x-x_n^j)+v_n^{\ell}(x),\quad \eta(\mathbf{v}^{\ell})\underset{\ell\to\infty}{\longrightarrow}0,$$

such that the identities (2.2)–(2.3) hold. Indeed, if $\eta(\mathbf{v})=0$, we can take $V^j\equiv0$ for all j; otherwise we choose $V^1\in\mathcal{V}(\mathbf{v})$ such that

$$\|V^1\|_{H^1}\geq\frac{1}{2}\eta(\mathbf{v})>0.$$

By definition, there exists some sequence $\mathbf{x}^1=\{x_n^1\}_{n=1}^{\infty}$ of $\mathbb{R}^d$ such that, up to extracting a subsequence, we have

$$v_n(\cdot+x_n^1)\rightharpoonup V^1\qquad\text{weakly.}$$

We set

$$v_n^1=v_n-V^1(\cdot-x_n^1).$$

Since $v_n^1(\cdot+x_n^1)\rightharpoonup0$, we get, as $n\to\infty$,

$$\begin{aligned}\|v_n\|_{L^2}^2&=\|V^1\|_{L^2}^2+\|v_n^1\|_{L^2}^2+o(1),\\ \|\nabla v_n\|_{L^2}^2&=\|\nabla V^1\|_{L^2}^2+\|\nabla v_n^1\|_{L^2}^2+o(1).\end{aligned}$$

Now, we replace $\mathbf{v}$ by $\mathbf{v}^1$ and repeat the same process. If $\eta(\mathbf{v}^1)>0$, we get V^2, $\mathbf{x}^2$, and $\mathbf{v}^2$. Moreover, we have

$$|x_n^1-x_n^2|\longrightarrow\infty\qquad\text{as}\quad n\to\infty.$$

Otherwise, up to extracting of a subsequence, we get

$$x_n^1-x_n^2\longrightarrow x_0$$

for some $x_0\in\mathbb{R}^d$. Since

$$v_n^1(\cdot+x_n^2)=v_n^1(\cdot+(x_n^2-x_n^1)+x_n^1)$$

and $v_n^1(\cdot+x_n^1)$ converge weakly to 0, then $V^2=0$. A contradiction. An argument of iteration and orthogonal extraction allows us to construct the families $\{\mathbf{x}^j\}_{j=1}^{\infty}$

and $\{V^j\}_{j=1}^{\infty}$ satisfying the claims above. Furthermore, the convergence of the series $\sum_{j=1}^{\infty}\|V^j\|_{H^1}^2$ implies that

$$\|V^j\|_{H^1} \underset{j\to\infty}{\longrightarrow} 0.$$

However, by construction, we have

$$\eta(\mathbf{v}^j) \le 2\|V^{j-1}\|_{H^1},$$

which proves that $\eta(\mathbf{v}^j) \to 0$k, as claimed. To complete the proof of Proposition 2.1, (2.1) remains to be proved. For that purpose let us introduce $\chi_R \in \mathcal{S}(\mathbb{R}^d)$ such that

$$\hat{\chi}_R(\xi) = 1 \quad \text{if } |\xi| \le R, \qquad \hat{\chi}_R(\xi) = 0 \quad \text{if } |\xi| \ge 2R.$$

Here $\hat{}$ denotes the Fourier transform. One has

$$v_n^\ell = \chi_R * v_n^\ell + (\delta - \chi_R) * v_n^\ell,$$

where $*$ stands for the convolution and δ for the Dirac distribution.

Let $p \in]2, 2^*[$ be fixed. On the one hand, using Sobolev embedding, we get

$$\|(\delta - \chi_R) * v_n^\ell\|_{L^p} \precsim \|(\delta - \chi_R) * v_n^\ell\|_{\dot{H}^\beta} \precsim R^{\beta-1}\|v_n^\ell\|_{H^1}$$

for $\beta = d(\frac{1}{2} - \frac{1}{p}) < 1$. On the other hand, one can estimate

$$\begin{aligned}\|\chi_R * v_n^\ell\|_{L^p} &\precsim \|\chi_R * v_n^\ell\|_{L^2}^{2/p}\|\chi_R * v_n^\ell\|_{L^\infty}^{1-2/p}\\ &\precsim \|v_n^\ell\|_{L^2}^{2/p}\|\chi_R * v_n^\ell\|_{L^\infty}^{1-2/p}.\end{aligned}$$

Now, observe that

$$\limsup_{n\to\infty}\|\chi_R * v_n^\ell\|_{L^\infty(\mathbb{R}^d)} = \sup_{\{x_n\}_{n=1}^\infty}\limsup_{n\to+\infty}|\chi_R * v_n^\ell(x_n)|.$$

Thus, in view of the definition of $\mathcal{V}(\mathbf{v}^\ell)$, we infer

$$\limsup_{n\to\infty}\|\chi_R * v_n^\ell\|_{L^\infty(\mathbb{R}^d)} \le \sup\left\{\left|\int_{\mathbb{R}^d}\chi_R(-x)V(x)dx\right|,\ V \in \mathcal{V}(\mathbf{v}^\ell)\right\}.$$

Therefore, by Hölder's inequality, it follows that

$$\limsup_{n\to\infty}\|\chi_R * v_n^\ell\|_{L^\infty(\mathbb{R}^d)} \le C(R)\sup\{\|V\|_{L^2(\mathbb{R}^d)},\ V \in \mathcal{V}(\mathbf{v}^\ell)\}.$$

Thus, we obtain

$$\limsup_{n\to\infty}\|\chi_R * v_n^\ell\|_{L^\infty(\mathbb{R}^d)} \le C(R)\eta(\mathbf{v}^\ell)$$

for every $\ell \ge 1$. Finally, we get

$$\|v_n^\ell\|_{L^p(\mathbb{R}^d)} \precsim R^{\beta-1}\|v_n^\ell\|_{H^1} + C(R)\|v_n^\ell\|_{L^2}^{2/p}\eta(\mathbf{v}^\ell)^{1-2/p}.$$

We successively let ℓ and R go to infinity, and since $\eta(\mathbf{v}^\ell) \underset{\ell\to\infty}{\longrightarrow} 0$ and the family of sequences $\{v_n^\ell\}$ are uniformly bounded in $H^1(\mathbb{R}^d)$, we infer

$$\limsup_{n\to\infty}\|v_n^\ell\|_{L^p} \underset{\ell\to\infty}{\longrightarrow} 0,$$

as claimed. This completes the proof of Proposition 2.1. □

Let us now prove Theorem 1.1. By extracting a subsequence, we may replace lim sup in the assumptions (1.6) by lim. According to Proposition 2.1, the sequence $\{v_n\}_{n=1}^{\infty}$ can be written, up to a subsequence, as

$$v_n(x) = \sum_{j=1}^{\ell} V^j(x - x_n^j) + v_n^{\ell}(x)$$

such that (2.1), (2.2), and (2.3) hold. This implies, in particular, that

$$m^{\frac{4}{d}+2} \leq \limsup_{n\to\infty} \|v_n\|_{L^{\frac{4}{d}+2}}^{\frac{4}{d}+2} = \limsup_{n\to\infty} \left\| \sum_{j=1}^{\infty} V^j(\cdot - x_n^j) \right\|_{L^{\frac{4}{d}+2}}^{\frac{4}{d}+2}.$$

The elementary inequality

$$\left| \left| \sum_{j=1}^{l} a_j \right|^{4/d+2} - \sum_{j=1}^{l} |a_j|^{4/d+2} \right| \leq C \sum_{j\neq k} |a_j||a_k|^{4/d+1},$$

and the pairwise orthogonality of the family $\{\mathbf{x}^j\}_{j=1}^{\infty}$ leads the mixed terms in the sum above to vanish, so that we get

$$m^{\frac{4}{d}+2} \leq \sum_{j=1}^{\infty} \|V^j\|_{L^{\frac{4}{d}+2}}^{\frac{4}{d}+2}.$$

However, in view of the precise version of the Gagliardo–Nirenberg inequality (1.4), we have

$$\sum_{j=1}^{\infty} \|V^j\|_{L^{\frac{4}{d}+2}}^{\frac{4}{d}+2} \leq C_d \sup\{\|V^j\|_{L^2}^{4/d}, j \geq 1\} \sum_{j=1}^{\infty} \|\nabla V^j\|_{L^2}^2.$$

Also, from (2.3), we infer

$$\sum_{j=1}^{\infty} \|\nabla V^j\|_{L^2}^2 \leq \limsup_{n\to\infty} \|\nabla v_n\|_{L^2}^2 \leq M^2.$$

Therefore,

$$\sup_{j\geq 1} \|V^j\|_{L^2}^{4/d} \geq \frac{m^{\frac{4}{d}+2}}{(M^2 C_d)^{d/4}}.$$

Since the series $\sum \|V^j\|_{L^2}^2$ converges, the supremum above is attained. In particular, there exists j_0 such that

$$\|V^{j_0}\|_{L^2} \geq \frac{m^{\frac{d}{2}+1}}{(C_d M^2)^{d/4}} = \left(\frac{d}{d+2}\right)^{d/4} \frac{m^{\frac{d}{2}+1}}{M^{d/2}} \|Q\|_{L^2}.$$

On the other hand, a change of variables gives

$$v_n(x + x_n^{j_0}) = V^{j_0}(x) + \sum_{\substack{1\leq j\leq \ell \\ j\neq j_0}} V^j(x + x_n^{j_0} - x_n^j) + \tilde{v}_n^{\ell}(x),$$

where $\tilde{v}_n^\ell(x) = v_n^\ell(x + x_n^{j_0})$. The pairwise orthogonality of the family $\{\mathbf{x}^j\}_{j=1}^\infty$ implies

$$V^j(\cdot + x_n^{j_0} - x_n^j) \rightharpoonup 0 \quad \text{weakly}$$

for every $j \neq j_0$. Hence, we get

$$v_n(\cdot + x_n^{j_0}) \rightharpoonup V^{j_0} + \tilde{v}^\ell,$$

where $\tilde{v}^\ell$ denote the weak limit of $\{\tilde{v}_n^\ell\}_{n=1}^\infty$. However, we have

$$\|\tilde{v}^\ell\|_{L^{\frac{4}{d}+2}} \leq \limsup_{n\to\infty} \|\tilde{v}_n^\ell\|_{L^{\frac{4}{d}+2}} = \limsup_{n\to\infty} \|v_n^\ell\|_{L^{\frac{4}{d}+2}} \underset{l\to\infty}{\longrightarrow} 0.$$

Thereby, by the uniqueness of the weak limit, we get

$$\tilde{v}^\ell = 0$$

for every $\ell \geq j_0$. Thus

$$v_n(\cdot + x_n^{j_0}) \rightharpoonup V^{j_0}.$$

The sequence $\{x_n^{j_0}\}_{n=1}^\infty$ and the function V^{j_0} now fulfill the conditions of Theorem 1.1. □

3. Proof of the main results.

3.1. The modified energy. Here we recall the result of almost conservation of the modified energy proved in [6].

I_N stands for the smoothing operators[4] $I_N : H^s \to H^1$:

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

where

$$m(\xi) = \begin{cases} 1, & |\xi| \leq N, \\ (\frac{|\xi|}{N})^{s-1}, & |\xi| > 3N, \end{cases}$$

with $m(\xi)$ smooth, radial, and monotone in $|\xi|$. The following properties of I_N are easily verified:

$$\begin{aligned} \|I_N u\|_{L^2} &\leq \|u\|_{L^2}, \\ \|u\|_{H^s} \leq &\|I_N u\|_{H^1} \leq N^{1-s}\|u\|_{H^s}. \end{aligned} \tag{3.1}$$

The *blowup parameter* associated with the H^s norm of the solution is

$$\Lambda(t) = \sup_{0\leq\tau\leq t} \|u(t)\|_{H^s}.$$

The following proposition is a restatement of the part of [6] which is relevant for us.

PROPOSITION 3.1 (see [6]). *There exists $s_Q \leq \frac{1}{5} + \frac{1}{5}\sqrt{11}$ such that for all $s > s_Q$ there exists $p(s) < 2$ with the following holding true: if $H^s \ni u_0 \longmapsto u(t)$ solves (1.1) on a maximal (forward) finite existence interval $[0, T^*)$, then for all $T < T^*$ there exists $N = N(T)$ such that*

$$|E[I_{N(T)}u(T)]| \leq C_0(\Lambda(T))^{p(s)} \tag{3.2}$$

[4] See [7] for more properties of this operator.

with $C_0 = C_0(s, T^*, \|u_0\|_{H^s})$. *Moreover,* $N(T) = C(\Lambda(T))^{\frac{p(s)}{2(1-s)}}$.

In [6], $p(s)$ is explicitly given by

$$p(s) = \frac{6^+}{2^- - 4^+(1-s)} 2(1-s), \tag{3.3}$$

where $\alpha^\pm = \alpha \pm \epsilon$ for some $\epsilon > 0$.

3.2. Proof of Theorem 1.5. As in [6], we choose $\{t_n\}_{n=1}^\infty$ to be a sequence such that $t_n \uparrow T^*$ and for each t_n

$$\|u(t_n)\|_{H^s} = \Lambda(t_n).$$

We set

$$\psi_n = \rho_n I_N u(t_n, \rho_n x),$$

where

$$\rho_n = \frac{\|\nabla Q\|_{L^2}}{\|\nabla I_N u(t_n, \cdot)\|_{L^2}}.$$

The estimate (3.1) yields

$$\rho_n \leq \frac{1}{\|u(t_n, \cdot)\|_{H^s}} = \frac{1}{\Lambda(t_n)}.$$

Also, from Corollary 3.6 in [6], it holds that

$$\rho_n \leq A(T^* - t_n)^{s/2}$$

for some constant $A > 0$. The sequence $\{\psi_n\}_{n=1}^\infty$ satisfies

$$\|\psi_n\|_{L^2} \leq \|u_0\|_{L^2}, \quad \|\nabla \psi_n\|_{L^2} = \|\nabla Q\|_{L^2}.$$

Furthermore, in view of Proposition 3.1,

$$E(\psi_n) = \rho_n^2 E[I_{N(t_n)} u(t_n)] \leq \rho_n^2 (\Lambda(t_n))^{p(s)} \leq (\Lambda(t_n))^{p(s)-2}.$$

Since $\|u(t_n)\|_{H^s} \longrightarrow +\infty$ and $p(s) < 2$, it holds that

$$E(\psi_n) \longrightarrow 0 \qquad \text{as} \quad n \to \infty,$$

which yields, in particular,

$$\|\psi_n\|_{L^4}^4 \longrightarrow 2\|\nabla Q\|_{L^2}^2 \qquad \text{as} \quad n \to \infty. \tag{3.4}$$

The family $\{\psi_n\}_{n=1}^\infty$ satisfies the conditions of Theorem 1.1 with

$$m = (2\|\nabla Q\|_{L^2}^2)^{1/4} \qquad \text{and} \qquad M = \|\nabla Q\|_{L^2}.$$

Thus, there exists $\{x_n\}_{n=1}^\infty \subset \mathbb{R}^2$ such that, up to a subsequence,

$$\psi_n(\cdot + x_n) \rightharpoonup V \in H^1, \tag{3.5}$$

with $\|V\|_{L^2} \geq \|Q\|_{L^2}$. Coming back to $\{\psi_n\}_{n=1}^{\infty}$, one obtains

$$\rho_n I_N u(t_n, \rho_n x + x_n) = V + \varepsilon_n,$$

where $\varepsilon_n \rightharpoonup 0$ in H^1.

However, for every $\bar{s} < s$, one has

$$\begin{aligned}\|\rho_n(I_N u(t_n) - u(t_n))(\rho_n \cdot + x_n)\|_{\dot{H}^{\bar{s}}(\mathbb{R}^2)} &\leq \rho_n^{\bar{s}} N^{\bar{s}-s} \|u(t_n, \cdot)\|_{H^s(\mathbb{R}^2)} \\ &\leq (\Lambda(t_n))^{\frac{p(s)(\bar{s}-s)}{2(1-s)}+1-\bar{s}}.\end{aligned}$$

Using the explicit formula of $p(s)$, an easy calculus yields that

$$\frac{p(s)(\bar{s}-s)}{2(1-s)} + 1 - \bar{s} < 0 \Longleftrightarrow \bar{s} < \tilde{s} := \frac{s+1}{4-2s}.$$

Under this choice, we get

$$\|\rho_n(I_N u(t_n) - u(t_n))(\rho_n \cdot + x_n)\|_{H^{\tilde{s}-}(\mathbb{R}^2)} \longrightarrow 0 \qquad \text{as} \quad n \to \infty.$$

Thus

$$\rho_n u(t_n, \rho_n x + x_n) = V + h_n, \tag{3.6}$$

where $h_n \rightharpoonup 0$ in $H^{\tilde{s}-}$. This concludes the proof of Theorem 1.5.

3.3. Proof of Theorem 1.8. If, in the context of the proof of Theorem 1.5, we assume also that $\|u_0\|_{L^2} = \|Q\|_{L^2}$, we get trivially (remember that we have already proved that $\|V\|_{L^2} \geq \|Q\|_{L^2}$)

$$\|V\|_{L^2} = \|Q\|_{L^2}.$$

Then the convergence (3.5) is strong in L^2 and, since $\{v_n\}$ is bounded in H^1, we have

$$v_n(\cdot + x_n) \longrightarrow V \qquad \text{in} \quad L^4.$$

Combined with (3.4) and the precise version of the Gagliardo–Nirenberg inequality (1.4), this leads to

$$\|\nabla V\|_{L^2} \geq \|\nabla Q\|_{L^2}.$$

However, one has

$$\|\nabla V\|_{L^2} \leq \limsup \|\nabla v_n\|_{L^2} = \|\nabla Q\|_{L^2},$$

which means that $\|\nabla V\|_{L^2} = \|\nabla Q\|_{L^2}$. As a result we get

$$v_n(\cdot + x_n) \longrightarrow V \qquad \text{in} \quad H^1$$

and $E(V) = 0$.

Let us summarize the properties of the profile V:

$$\|V\|_{L^2} = \|Q\|_{L^2}, \quad \|\nabla V\|_{L^2} = \|\nabla Q\|_{L^2}, \quad \text{and} \quad E(V) = 0.$$

The variational characterization of the ground state implies that

$$V(x) = e^{i\theta} Q(x + x_0)$$

for some $\theta \in [0, 2\pi[$ and $x_0 \in \mathbb{R}^2$.

Coming back to (3.6), one obtains

$$\rho_n u(t_n, \rho_n x + x_n) = \lambda e^{i\theta} Q(\lambda x + x_0) + \varepsilon_n,$$

where $h_n \to 0$ in $H^{\tilde{s}-}$. This ends the proof of Theorem 1.8.

3.4. Proof of Corollary 1.10. Let u be a blowup solution of (1.1) at finite time $T^* > 0$. According to Theorem 1.5, there exists a time sequence such that $t_n \longrightarrow T^*$, a profile $V \in H^1(\mathbb{R}^2)$ with $\|V\|_{L^2} \geq \|Q\|_{L^2}$, and a sequence $\{\rho_n, x_n\}_{n=1}^{\infty} \subset \mathbb{R}_+^* \times \mathbb{R}^2$ such that

$$\rho_n u(t_n, \rho_n x + x_n) \rightharpoonup V \tag{3.7}$$

and

$$\lim_{n\to\infty} \frac{\rho_n}{(T^* - t_n)^{s/2}} \leq A \tag{3.8}$$

for some $A \geq 0$. From (3.7), it follows that

$$\lim_{n\to\infty} \frac{1}{(\rho_n)^2} \int_{|x|\leq R} |u(t_n, \rho_n x + x_n)|^2 dx \geq \int_{|x|\leq R} |V|^2 dx$$

for every $R > 0$. Thus,

$$\lim_{n\to\infty} \sup_{y\in\mathbb{R}^2} \int_{|x-y|\leq R\rho_n} |u(t_n, x)|^2 dx \geq \int_{|x|\leq R} |V|^2 dx. \tag{3.9}$$

Since $\frac{(T^*-t)^{s/2}}{\lambda(t)} \to 0$ as $t \to T^*$, it follows from (3.8) that $\frac{\rho_n}{\lambda(t_n)} \to 0$, and then

$$\lim_{n\to\infty} \sup_{y\in\mathbb{R}^2} \int_{|x-y|\leq \lambda(t_n)} |u(t_n, x)|^2 dx \geq \int_{|x|\leq R} |V|^2 dx$$

for every R. We let R go to infinity to obtain

$$\lim_{n\to\infty} \sup_{y\in\mathbb{R}^2} \int_{\{|x-y|\leq \lambda(t_n)\}} |u(t_n, x)|^2 dx \geq \int_{\mathbb{R}^2} |V|^2 dx \geq \|Q\|_{L^2}^2.$$

This yields finally

$$\limsup_{t\uparrow T^*} \sup_{y\in\mathbb{R}^2} \int_{\{|x-y|\leq \lambda(t)\}} |u(t, x)|^2 dx \geq \|Q\|_{L^2}^2.$$

Since, for every t, the function $y \longmapsto \int_{\{|x-y|\leq\lambda(t)\}} |u(t,x)|^2 dx$ is continuous and goes to 0 at infinity, there exists a family $x(t)$ such that

$$\sup_{y\in\mathbb{R}^2} \int_{\{|x-y|\leq \lambda(t)\}} |u(t, x)|^2 dx = \int_{\{|x-x(t)|\leq \lambda(t)\}} |u(t, x)|^2 dx,$$

which concludes the proof of Corollary 1.10. ∎

Remark 3.2. While this manuscript was under review, we were informed that N. Tzirakis has written a paper following [6] in which he removed the radial symmetry assumption and established corresponding results for the focusing quintic nonlinear Schrödinger equation on $\mathbb{R}$.

Acknowledgments. The authors thank the anonymous referees for their very relevant and detailed comments, which helped to improve the manuscript.

REFERENCES

[1] H. Bahouri and P. Gérard, *High frequency approximation of solutions to critical nonlinear wave equations*, Amer. J. Math., 121 (1999), pp. 131–175.

[2] P. Bégout and A. Vargas, *Mass concentration phenomena for the L^2-critical nonlinear Schrödinger equation*, Trans. Amer. Math. Soc., to appear.

[3] J. Bourgain, *Refinements of Strichartz inequality and applications to 2D-NLS with critical nonlinearity*, Int. Math. Res. Not., 8 (1998), pp. 253–283.

[4] R. Carles and S. Keraani, *Quadratic oscillations in NLS* II. *The L^2-critical case*, Trans. Amer. Math. Soc., 359 (2007), pp. 33–62.

[5] T. Cazenave, *An Introduction to Nonlinear Schrödinger Equations*, Courant Lect. Notes Math. 10, Courant Inst. Math. Sci., New York.

[6] J. Colliander, S. Raynor, C. Sulem, and J. D. Wright, *Ground state mass concentration in the L^2-critical nonlinear Schrodinger equation below H^1*, Math. Res. Lett., 12 (2005), pp. 357–375.

[7] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Almost conservation laws and global rough solutions to a nonlinear Schrödinger equation*, Math. Res. Lett., 9 (2002), pp. 659–682.

[8] P. Gérard, *Description du défaut de compacité de l'injection de Sobolev*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 213–233.

[9] J. Ginibre and G. Velo, *Scattering theory in the energy space for a class of nonlinear Schrödinger equations*, J. Math. Pure Appl., 64 (1984), pp. 363–401.

[10] S. Keraani, *On the defect of compactness for the Strichartz estimates of the Schrödinger equations*, J. Differential Equations, 175 (2001), pp. 353–392.

[11] S. Keraani, *On the blowup phenomenon for the L^2-critical nonlinear Schrödinger equation*, J. Differential Equations, 175 (2001), pp. 353–392.

[12] M. K. Kwong, *Uniqueness of positive solutions to $\Delta u - u + u^p = 0$, in $\mathbb{R}^n$*, Arch. Ration. Mech. Anal, 105 (1989), pp. 243–266.

[13] P-L. Lions, *The concentration-compactness principle in the calculus of variations. The compact case. Part* 1, Ann. Inst. H. Poincaré Anal. Non linéaire, 1 (1984), pp. 109–145.

[14] F. Merle, *Determination of blowup solutions with minimal mass for nonlinear Schrödinger equations with critical power*, Duke Math. J., 69 (1993), pp. 203–254.

[15] F. Merle, *Construction of solutions with exactly k blowup points for nonlinear Schrödinger equations with critical nonlinearity*, Comm. Math. Phys., 129 (1990), pp. 223–240.

[16] F. Merle, *Blow-up phenomena for critical nonlinear Schrödinger and Zakharov equations*, in Proceedings of the International Congress of Mathematicians, Vol. III (Berlin, 1998), Doc. Math. 1998, Extra Vol. III, pp. 57–66.

[17] F. Merle and P. Raphael, *On universality of blow-up profile for L^2 critical nonlinear Schrödinger equation*, Invent. Math., 156 (2004), pp. 565–672.

[18] F. Merle and Y. Tsutsumi, *L^2 concentration of blowup solutions for the nonlinear Schrödinger equation with critical power nonlinearity*, J. Differential Equations, 84 (1990), pp. 205–214.

[19] F. Merle and L. Vega, *Compactness at blowup time for L^2 solutions of the critical nonlinear Schrödinger equations in 2D*, Int. Math. Res. Not., 8 (1998), pp. 399–425.

[20] A. Moyua, A. Vargas, and L. Vega, *Schrödinger maximal function and restriction properties of the Fourier transform*, Int. Math. Res. Not., 16 (1996), pp. 793–815.

[21] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation. Self-focusing and Wave Collapse*, Appl. Math. Sci. 139, Springer-Verlag, New York, 1999.

[22] R. Strichartz, *Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equation*, Duke Math. J., 44 (1977), pp. 705–714.

[23] P. Tomas, *A restriction theorem for the Fourier transform*, Bull. Amer. Math. Soc., 81 (1975), pp. 477–478.

[24] Y. Tsutsumi, *Rate of L^2 concentration of blowup solutions for the nonlinear Schrödinger equation with critical power*, Nonlinear Anal., 15 (1990), pp. 719–724.

[25] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1983), pp. 567–567.

[26] M. I. Weinstein, *On the structure of singularities in solutions to the nonlinear dispersive evolution equations*, Comm. Partial Differential Equations, 11 (1984), pp. 545–565.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1048–1059

HOMOGENIZATION AND FIELD CONCENTRATIONS IN HETEROGENEOUS MEDIA*

ROBERT LIPTON†

Abstract. A multiscale characterization of the field concentrations inside composite and polycrystalline media is developed. We focus on gradient fields associated with the intensive quantities given by the temperature and the electric potential. In the linear regime these quantities are modeled by the solution of a second order elliptic partial differential equation with oscillatory coefficients. The characteristic length scale of the heterogeneity relative to the sample size is denoted by ε and the intensive quantity is denoted by u^ε. Field concentrations are measured using the L^p norm of the gradient field $\|\nabla u^\varepsilon\|_{L^p(D)}$ for $2 \leq p < \infty$. The analysis focuses on the case when $0 < \varepsilon \ll 1$. Explicit lower bounds on $\liminf_{\varepsilon \to 0} \|\nabla u^\varepsilon\|_{L^p(D)}$ are developed. These bounds provide a way to rigorously assess field concentrations generated by the microgeometry without having to compute the actual field u^ε.

Key words. composite materials, polycrystalline media, homogenization, field concentrations, Young measures

AMS subject classifications. 35B27, 74Q05

DOI. 10.1137/050648687

1. Introduction.

The initiation of failure inside heterogeneous media is a multiscale phenomenon. Loads applied at the structural scale are often amplified by the microstructure, creating local zones of high field concentration. The local amplification of the applied field creates conditions that are favorable for failure initiation [8]. This paper focuses on gradient fields associated with the intensive quantities given by the temperature and the electric potential inside heterogeneous media. The local integrability of the gradient directly correlates with singularity strength, which influences the onset of failure such as dielectric breakdown.

In this work it is shown how to assess the L^p integrability of the gradient fields in microstructured media by investigating the multiscale integrability of suitably defined quantities. The analysis is carried out with minimal regularity assumptions on the coefficients describing the local properties inside the heterogeneous media. The results are described in terms of the pth order moments of the solution of two-scale corrector problems. The quantities are sensitive to microscopic field concentrations and can become divergent for $p > 2$. This is in contrast to the well-known effective constitutive properties, which are based upon local averages and are bounded above independently of the microgeometry.

The results given here are presented in the context of two-scale homogenization [1], [18]. We consider a bounded domain Ω in $\mathbf{R}^n$, $n \geq 2$. A common microstructure

*Received by the editors December 29, 2005; accepted for publication (in revised form) May 1, 2006; published electronically November 21, 2006. This research effort is sponsored by the NSF through grant DMS-0406374 and by the Air Force Office of Scientific Research, Air Force Material Command USAF, under grants F49620-02-1-0041 and FA9550-05-1-0008. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

http://www.siam.org/journals/sima/38-4/64868.html

†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (lipton@math.lsu.edu).

that admits a two-scale description is a simple generalization of a uniformly periodic microstructure and is described as follows. Consider a partition of the domain Ω made up of measurable subsets Ω_ℓ, $\ell = 1, 2, \ldots, K$, such that $\Omega = \cup_{\ell=1}^{K} \Omega_\ell$. Inside each subdomain Ω_ℓ we place a different periodic microstructure made from N anisotropic heat conductors. This type of microstructure will be referred to as a piecewise periodic microstructure [4]. Well-known engineering composites that are modeled by piecewise periodic microstructures include fiber reinforced laminates [6], [19], [21].

The thermal conductivity tensor for the piecewise periodic microstructure is described as follows. The indicator function for each of the subdomains Ω_ℓ is denoted by $\chi_{\Omega_\ell}(\mathbf{x})$, taking the value 1 for points in Ω_ℓ and 0 outside. In order to describe the periodic microstructure inside the ℓth subdomain we introduce the unit period cell Q. The configuration of the N phases inside Q is described by the indicator functions $\chi_\ell^i(\mathbf{y})$, $i = 1, \ldots, N$, associated with each phase. Here $\chi_\ell^i(\mathbf{y}) = 1$ for points inside the ith phase and 0 outside. The length scale of the microstructure relative to the size of the domain Ω is given by $\varepsilon_k = 1/k$, $k = 1, 2 \ldots$. The microstructure is obtained by rescaling the configuration inside the unit period cell. The indicator function of the ith conductor in the microstructured composite is given by

$$\chi_i^{\varepsilon_k}(\mathbf{x}) = \chi_i(\mathbf{x}, \mathbf{x}/\varepsilon_k) = \sum_{\ell}^{K} \chi_{\Omega_\ell}(\mathbf{x}) \chi_\ell^i(\mathbf{x}/\varepsilon_k). \tag{1.1}$$

The local conductivity tensor A^{ε_k} has a two-scale structure and is given by

$$A^{\varepsilon_k}(\mathbf{x}) = A(\mathbf{x}, \mathbf{x}/\varepsilon_k) = \sum_{i}^{N} A^i \chi_i(\mathbf{x}, \mathbf{x}/\varepsilon_k). \tag{1.2}$$

Other heterogeneous media that are amenable to similar or more general two-scale descriptions include polycrystalline materials such as metals and ceramics. We state the general hypotheses under which the two-scale homogenization theory applies; see [1] and [2]. It is assumed that $A(\mathbf{x}, \mathbf{y})$ is a matrix defined on $\Omega \times Q$ and there exist positive numbers $\alpha < \beta$ such that for every vector η in $\mathbf{R}^n$,

$$\alpha|\eta|^2 \leq A(\mathbf{x}, \mathbf{y})\eta \cdot \eta \leq \beta|\eta|^2. \tag{1.3}$$

The conductivity $A_{ij}(\mathbf{x}, \mathbf{y})$ is Q-periodic in the second variable, $A_{ij}(\mathbf{x}, \mathbf{x}/\varepsilon_k)$ is measurable and satisfies

$$\lim_{\varepsilon_k \to 0} \int_\Omega \left| A_{ij}\left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon_k}\right) \right|^2 d\mathbf{x} = \int_{\Omega \times Q} |A_{ij}(\mathbf{x}, \mathbf{y})|^2 \, d\mathbf{x} d\mathbf{y}, \tag{1.4}$$

and for any suitable two-scale trial field $\psi(\mathbf{x}, \mathbf{y})$,

$$\lim_{\varepsilon_k \to 0} \int_\Omega A_{ij}\left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon_k}\right) \psi\left(x, \frac{\mathbf{x}}{\varepsilon_k}\right) d\mathbf{x} = \int_{\Omega \times Q} A_{ij}(\mathbf{x}, \mathbf{y}) \psi(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} d\mathbf{y}. \tag{1.5}$$

The convergence given by (1.5) is a weak convergence and is known as two-scale convergence [1], [18]. The space of suitable two-scale trials is denoted by $L^2[D; C_{per}(Q)]$. Here $C_{per}(Q)$ denotes Q-periodic continuous functions defined on $\mathbf{R}^3$, and the space $L^2[D; C_{per}(Q)]$ is the space of functions $h : \Omega \to C_{per}(Q)$ which are measurable and satisfy $\int_\Omega \|h(\mathbf{x})\|^2_{C_{per}(Q)} d\mathbf{x} < \infty$. The norm $\|h(\mathbf{x})\|_{C_{per}(Q)}$ is defined by $\sup_{\mathbf{y} \in Q} |h(\mathbf{x}, \mathbf{y})|$.

In what follows, no other regularity hypothesis on the conductivity matrix $A(\mathbf{x},\mathbf{y})$ is made.

The temperature field u^{ε_k} associated with the conductivity tensor field $A^{\varepsilon_k}(\mathbf{x}) = A(\mathbf{x},\mathbf{x}/\varepsilon_k)$ is the solution of the equilibrium equation

$$-\text{div}\,(A^{\varepsilon_k}(\mathbf{x})\nabla u^{\varepsilon_k}) = f \ \text{ in } \ \Omega \tag{1.6}$$

with the boundary conditions given by $u^{\varepsilon_k} = 0$ on $\partial\Omega_D$ and $\mathbf{n}\cdot A^{\varepsilon_k}\nabla u^{\varepsilon_k} = g$ on $\partial\Omega_N$ with $\partial\Omega_D \cap \partial\Omega_N = \emptyset$ and $\partial\Omega = \overline{\partial\Omega}_D \cup \overline{\partial\Omega}_N$.

In what follows, we consider the limit as ε_k tends to zero. We fix a subdomain D of Ω and derive lower bounds on

$$\liminf_{\varepsilon_k\to 0} \|\nabla u^{\varepsilon_k}\|_{L^p(D)}. \tag{1.7}$$

The lower bound is expressed in terms of a two-scale integral that encodes the field amplification properties of the microstructure. It is formulated in terms of the solution of the homogenized problem together with a local corrector matrix that captures the interaction between the periodic microstructure and the gradients of the homogenized temperature field. The bounds introduced here provide a rigorous way to assess field concentrations generated by the microgeometry without having to compute the full solution u^{ε_k}.

We consider an orthonormal basis for $\mathbf{R}^n$ and denote the basis vectors by $\mathbf{e}^i$, $i = 1,\ldots,n$. The lower bound is given in terms of the solutions $w^i(\mathbf{x},\mathbf{y})$ to the local periodic problem. For each $\mathbf{x}$ in Ω, the function $w^i(\mathbf{x},\mathbf{y})$ is a Q-periodic function of the second variable $\mathbf{y}$ and is a solution of

$$\text{div}_{\mathbf{y}}\left(A(\mathbf{x},\mathbf{y})(\nabla_{\mathbf{y}} w^i(\mathbf{x},\mathbf{y}) + \mathbf{e}^i)\right) = 0, \tag{1.8}$$

with $\int_Q w^i(\mathbf{x},\mathbf{y})\,d\mathbf{y} = 0$. The corrector matrix $P(\mathbf{x},\mathbf{y})$ is defined by

$$P_{ij}(\mathbf{x},\mathbf{y}) = \partial_{y_j} w^i(\mathbf{x},\mathbf{y}) + \delta_{ij}, \tag{1.9}$$

where $\delta_{ij} = 1$ for $i = j$ and 0 otherwise. The associated effective conductivity tensor $A^E(\mathbf{x})$ is given by

$$A^E(\mathbf{x}) = \int_Q A(\mathbf{x},\mathbf{y})P(\mathbf{x},\mathbf{y})\,d\mathbf{y}. \tag{1.10}$$

The two-scale homogenization theory gives the following theorem [1].

THEOREM 1.1 (two-scale homogenization theorem). *The sequence of solutions* $\{u^{\varepsilon_k}\}_{\varepsilon_k>0}$ *of* (1.6) *converges weakly to* $u^H(\mathbf{x})$ *in* $H^1(\Omega)$*, where* u^H *is the solution of the homogenized problem*

$$\begin{aligned} -\text{div}\left(A^E(\mathbf{x})\nabla u^H(\mathbf{x})\right) &= f(\mathbf{x}) \ \textit{in}\ \Omega, \\ u^H(\mathbf{x}) &= 0 \ \textit{on}\ \partial\Omega_D, \ \textit{and} \\ \mathbf{n}\cdot A^E\nabla u^H &= g \ \textit{on}\ \partial\Omega_N. \end{aligned} \tag{1.11}$$

The field concentration functions of order p are defined by

$$f_p(\mathbf{x},\nabla u^H(\mathbf{x})) \equiv \left(\int_Q |P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x})|^p\,d\mathbf{y}\right)^{1/p}, \quad 2 \le p \le \infty, \tag{1.12}$$

and $f_p(\mathbf{x}, \nabla u^H(\mathbf{x})) \le f_q(\mathbf{x}, \nabla u^H(\mathbf{x}))$ for $p \le q$. It is clear that f_p corresponds to a pth order moment of the corrector matrix (1.9) and that

$$f_\infty(\mathbf{x}, \nabla u^H(\mathbf{x})) \equiv \lim_{p\to\infty} \left(\int_Q |P(\mathbf{x}, \mathbf{y}) \nabla u^H(\mathbf{x})|^p \, d\mathbf{y} \right)^{1/p}. \tag{1.13}$$

THEOREM 1.2 (lower bounds on field concentrations). *For* $2 \le p < \infty$,

$$\left(\int_D \left(f_p(\mathbf{x}, \nabla u^H(\mathbf{x})) \right)^p d\mathbf{x} \right)^{1/p} \le \liminf_{\varepsilon_k \to 0} \|\nabla u^{\varepsilon_k}\|_{L^p(D)}. \tag{1.14}$$

For multiphase conductivity problems with coefficients described by (1.2), the field concentration functions of order p are defined for each phase and are given by

$$f_p^i(\mathbf{x}, \nabla u^H(\mathbf{x})) \equiv \left(\int_Q \chi_i(\mathbf{x}, \mathbf{y}) |P(\mathbf{x}, \mathbf{y}) \nabla u^H(\mathbf{x})|^p \, d\mathbf{y} \right)^{1/p}, \ i = 1, \dots, N, \ 2 \le p \le \infty, \tag{1.15}$$

and $f_p^i(\mathbf{x}, \nabla u^H(\mathbf{x})) \le f_q^i(\mathbf{x}, \nabla u^H(\mathbf{x}))$ for $p \le q$. As before, one defines

$$f_\infty^i(\mathbf{x}, \nabla u^H(\mathbf{x})) \equiv \lim_{p\to\infty} \left(\int_Q \chi_i(\mathbf{x}, \mathbf{y}) |P(\mathbf{x}, \mathbf{y}) \nabla u^H(\mathbf{x})|^p \, d\mathbf{y} \right)^{1/p}. \tag{1.16}$$

For this case, lower bounds on

$$\liminf_{\varepsilon_k \to 0} \|\chi_i^{\varepsilon_k} \nabla u^{\varepsilon_k}\|_{L^p(D)} \tag{1.17}$$

are given by the following theorem.

THEOREM 1.3 (lower bounds for multiphase composites). *For* $2 \le p < \infty$,

$$\left(\int_D \left(f_p^i(\mathbf{x}, \nabla u^H(\mathbf{x})) \right)^p d\mathbf{x} \right)^{1/p} \le \liminf_{\varepsilon_k \to 0} \|\chi_i^{\varepsilon_k} \nabla u^{\varepsilon_k}\|_{L^p(D)}. \tag{1.18}$$

The bounds can be applied to develop a Chebyshev inequality for the distribution functions associated with the sequence $\{\chi_i^{\varepsilon_k} |\nabla u^{\varepsilon_k}|\}_{\varepsilon_k > 0}$. Here the distribution function $\lambda_i^{\varepsilon_k}(D, t)$ gives the measure of the set inside D, where $\chi_i^{\varepsilon_k} |\nabla u^{\varepsilon_k}| > t$.

Arguing as in Proposition 2.1 of [12] and combining with (1.18) gives the following.

THEOREM 1.4 (homogenized Chebyshev inequality).

$$\limsup_{\varepsilon_k \to 0} \lambda_i^{\varepsilon_k}(D, t) \le t^{-p} \left(\int_D \left(f_p^i(\mathbf{x}, \nabla u^H(\mathbf{x})) \right)^p d\mathbf{x} \right) \le t^{-p} \liminf_{\varepsilon_k \to 0} \|\chi_i^{\varepsilon_k} \nabla u^{\varepsilon_k}\|_{L^p(D)}^p. \tag{1.19}$$

We point out that Theorems 1.2 and 1.3 are obtained using the minimum regularity assumptions on the coefficients A^{ε_k}. Because of this, the hypotheses of Theorem 2.6 in [1] do not apply, and one cannot take advantage of the strong convergence given in that theorem. Instead, the theorems are proved using a perturbation approach introduced in [11] and [13]; see also our section 2.

The lower bounds are sensitive to the presence of singularities generated by the microstructure. To illustrate this we consider a microstructure made from a periodic

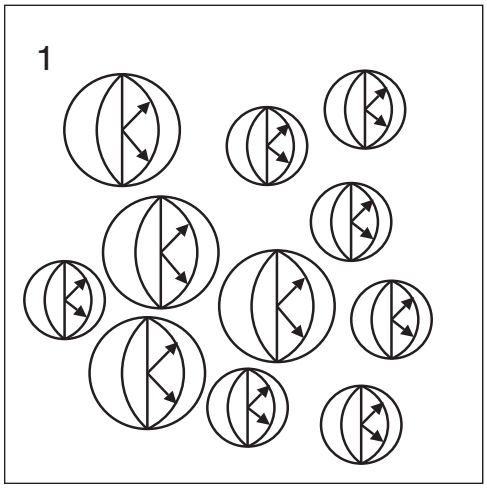


FIG. 1. *Unit period cell with Schulgasser crystallites embedded inside a material with unit thermal conductivity.*

distribution of uniaxial crystallites embedded in an isotropic matrix of unit conductivity. The period cell for the composite is illustrated in Figure 1. Each crystallite occupies a sphere and has conductivity λ_1 in the radial direction and λ_2 in the tangential direction. The dispersion of the N crystallites is specified by $\cup_\ell^N B(\mathbf{y}^\ell, r_\ell)$, where $B(\mathbf{y}^\ell, r_\ell)$ denotes the ℓth sphere centered at $\mathbf{y}^\ell$ with radius r_ℓ. Each crystallite has a conductivity tensor given by

$$A(\mathbf{y}) = \lambda_1 \mathbf{n} \otimes \mathbf{n} + \lambda_2 (I - \mathbf{n} \otimes \mathbf{n}), \tag{1.20}$$

where $\mathbf{n} = (\mathbf{y} - \mathbf{y}^\ell)/|\mathbf{y} - \mathbf{y}^\ell|$ for $\mathbf{y}$ in $B(\mathbf{y}^\ell, r_\ell)$ and I is the 3×3 identity. Outside the crystallites we set $A(\mathbf{y}) = I$. It is assumed that the aggregate of crystallites occupy a portion of the unit cell Q of volume $0 < \theta < 1$. It is noted that the conductivity inside each crystallite is precisely the one employed in the Schulgasser sphere assemblage [22].

When a constant gradient field is applied to a single isolated crystallite and when $\lambda_1 > \lambda_2$, the crystallite exhibits a gradient field singularity at its center. In what follows, we use the lower bound (1.14) to show how this local information affects the integrability of the sequence $\{\nabla u^{\varepsilon_k}\}_{\varepsilon_k > 0}$. We form $A^{\varepsilon_k} = A(\mathbf{x}/\varepsilon_k)$ and consider solutions u^{ε_k} of (1.6). To fix ideas we choose f to be in $L^r(\Omega)$ for $r > 3$ and g to be in $L^2(\partial\Omega_N)$. In what follows, λ_2 is restricted to lie in the interval $1/2 < \lambda_2 < 1$, and $\lambda_1 = 1/(2\lambda_2 - 1)$. For this choice it is shown in section 3 that the homogenized temperature field u^H is the solution of (1.11) with $A^E = I$.

For D compactly contained in Ω, it follows from the L^p theory [15] that $\|\nabla u^H\|_{L^p(D)} < \infty$ for every $1 \le p < \infty$. On the other hand, calculation and application of Theorem 1.2 show that

$$LB(p) \times \|\nabla u^H\|_{L^p(D)} \le \liminf_{\varepsilon_k \to 0} \|\nabla u^{\varepsilon_k}\|_{L^p(D)}, \tag{1.21}$$

where

$$LB(p) = \begin{cases} \frac{3p\theta(2\lambda_2 - 1)}{2(1-\lambda_2)(\frac{3}{2(1-\lambda_2)} - p)} + (1-\theta) & \text{for } p < \frac{3}{2(1-\lambda_2)}, \\ +\infty & \text{for } p \ge \frac{3}{2(1-\lambda_2)}. \end{cases} \tag{1.22}$$

For a fixed choice of λ_2, the value $p_c = \frac{3}{2(1-\lambda_2)}$ satisfies $3 < p_c < +\infty$, and we have

$$\liminf_{\varepsilon_k \to 0} \|\nabla u^{\varepsilon_k}\|_{L^p(D)} = +\infty \text{ for } p \geq p_c. \tag{1.23}$$

This is in stark contrast to the L^p integrability of the gradient of the homogenized solution, which holds for any $p < +\infty$. It is clear from this example that the information carried by the homogenized problem is not adequate and misses the singular behavior exhibited by the sequence $\{\nabla u^{\varepsilon_k}\}_{\varepsilon_k > 0}$. This example shows that failure initiation criteria based solely upon the solution of the homogenized equations will be optimistic. The inequalities given above are established in section 3.

The maximum integrability exponent for the gradient of the solution of the local problem (1.8) is referred to as the threshold exponent for the composite. The threshold exponent is introduced in the work of Milton [16] and measures the worst singularity of the gradient field. The threshold exponent is precisely p_c for the local problem considered here and corresponds to the divergence in the lower bound for $p \geq p_c$.

Next we consider an example for which the sequence $\{\nabla u^{\varepsilon_k}\}_{\varepsilon_k > 0}$ is uniformly bounded in L^p for some class of coefficients and right-hand sides f. For this case we show that the lower bound given in Theorem 1.2 is attained. In this example we make use of the a priori estimates for $\{\nabla u^{\varepsilon_k}\}_{\varepsilon_k > 0}$ developed in Theorem 4 of Avellaneda and Lin [3]. Let Ω be a $C^{1,\alpha}$ domain $(0 < \alpha \leq 1)$ and suppose for $0 < \gamma \leq 1$, $C > 0$, that $A(\mathbf{y}) \in C^\gamma(\mathbf{R}^n)$ and $\|A(\mathbf{y})\|_{C^\gamma(\mathbf{R}^n)} \leq C$. Then we choose $A^{\varepsilon_k} = A(\mathbf{x}/\varepsilon_k)$. For $\delta > 0$ suppose $2 \leq q \leq n + \delta$ and $f \in L^q$ and set $1/\hat{q} = 1/q - 1/(n+\delta)$. Given these choices, we consider the $W_0^{1,2}(\Omega)$ solutions u^{ε_k} of

$$-\text{div}\,(A^{\varepsilon_k}(\mathbf{x})\nabla u^{\varepsilon_k}) = f \quad \text{in} \quad \Omega. \tag{1.24}$$

It is shown in section 4 that (1.14) holds with equality for every p such that $p < \hat{q}$. In fact, it is seen more generally that, for $p < \hat{q}$ and any Carathéodory function $\psi : D \times \mathbf{R}^n \to \mathbf{R}$ satisfying

$$|\psi(\mathbf{x}, \eta)| \leq |\eta|^p \text{ for a.e. } \mathbf{x} \in D \text{ and } \eta \in \mathbf{R}^3, \tag{1.25}$$

we have

$$\lim_{\varepsilon_k \to 0} \int_D \psi(\mathbf{x}, \nabla u^{\varepsilon_k}(\mathbf{x}))\, d\mathbf{x} = \int_D \int_Q \psi(\mathbf{x}, P(\mathbf{y})\nabla u^H(\mathbf{x}))\, d\mathbf{y} d\mathbf{x}. \tag{1.26}$$

This is established in section 4.

It is anticipated that there are several classes of conductivity coefficients and right-hand sides f for which the lower bounds are attained. In this direction we point out the a priori estimates given in [7], [9], [10], and [23].

We conclude by noting that the analogues of the field concentration functions (1.12) and (1.15) have appeared earlier in the contexts of G-convergence and random media; see [11] and [12]. In those treatments they are shown to provide upper bounds for the distribution function of the local stress and electric field for G-convergent sequences of elasticity tensors and random dielectric tensors.

2. Derivation of the lower bounds. We recall the weak formulation of the $\varepsilon_k > 0$ problem given by (1.6). Let V denote the closure in $H^1(\Omega)$ of all smooth functions that vanish on $\partial\Omega_D$. We suppose that f is in $L^2(\Omega)$ and g belongs to

$L^2(\partial\Omega_N)$. The function u^{ε_k} belonging to V is the solution of the weak formulation of the boundary value problem given by

$$\int_\Omega A(\mathbf{x},\mathbf{x}/\varepsilon_k)\nabla u^{\varepsilon_k}\cdot\nabla\varphi\,d\mathbf{x}=\int_\Omega f\varphi\,d\mathbf{x}+\int_{\partial\Omega_N} g\varphi\,ds \tag{2.1}$$

for every φ in V. Here ds is an element of surface area.

In order to express the two-scale weak formulation of (1.11) we introduce the following function spaces. The space of square integrable Q-periodic mean zero functions with square integrable derivatives is denoted by $H^1_{per}(Q)/\mathbf{R}$. The norm of an element v in this space is denoted by $\|v\|_{H^1_{per}(Q)/\mathbf{R}}$. The space of measurable functions h from Ω to $H^1_{per}(Q)/\mathbf{R}$ for which $\int_\Omega \|h(\mathbf{x})\|^2_{H^1_{per}(Q)/\mathbf{R}}\,d\mathbf{x}<\infty$ is denoted by $L^2[\Omega;H^1_{per}(Q)/\mathbf{R}]$. This function space was introduced for the description of the two-scale homogenized problem in [18]. The weak formulation of the two-scale homogenized problem (1.11) is given by the unfolded variational principle [1], [5], [14].

THEOREM 2.1 (unfolded variational principle). *The pair (u^H,u_1) is the unique solution in $V\times L^2[\Omega;H^1_{per}(Q)/\mathbf{R}]$ of*

$$\begin{aligned}&\int_\Omega\int_Q A(\mathbf{x},\mathbf{y})(\nabla u^H(\mathbf{x})+\nabla_{\mathbf{y}}u_1(\mathbf{x},\mathbf{y}))\cdot(\nabla\varphi(\mathbf{x})+\nabla_{\mathbf{y}}\varphi_1(\mathbf{x},\mathbf{y}))\,d\mathbf{y}\,d\mathbf{x}\\&=\int_\Omega f\varphi\,d\mathbf{x}+\int_{\partial\Omega_N} g\varphi\,ds\end{aligned} \tag{2.2}$$

for every (φ,φ_1) in $V\times L^2[\Omega;H^1_{per}(Q)/\mathbf{R}]$. Moreover,

$$\nabla u^H+\nabla_{\mathbf{y}}u_1(\mathbf{x},\mathbf{y})=P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x}). \tag{2.3}$$

In order to establish Theorems 1.2 and 1.3 we recall the function spaces used in the description of two-scale convergence [14]. The space $C_{per}(Q)$ denotes Q-periodic continuous functions defined on $\mathbf{R}^3$. For $1\le r<\infty$, the space $L^r[D;C_{per}(Q)]$ is the space of functions $h:D\to C_{per}(Q)$, which are measurable and satisfy $\int_D\|h(\mathbf{x})\|^r_{C_{per}(Q)}d\mathbf{x}<\infty$. Here $\|h(\mathbf{x})\|_{C_{per}(Q)}=\sup_{\mathbf{y}\in Q}|h(\mathbf{x},\mathbf{y})|$. The intersection of the spaces $L^\infty(D\times Q)$ and $L^r[D;C_{per}(Q)]$ is denoted by V^r. For $1<r<\infty$ we introduce $1<r'<\infty$ such that $\frac{1}{r}+\frac{1}{r'}=1$. We establish Theorems 1.2 and 1.3 with the aid of the following lemmas.

LEMMA 2.1 (localization lemma). *Fix a domain of interest D inside Ω. Let $q(\mathbf{x},\mathbf{y})$ be any test function in V^r; then one can pass to the limit $\varepsilon_k\to 0$ in the sequence of solutions $\{u^{\varepsilon_k}\}_{\varepsilon_k>0}$ of (1.6) to obtain*

$$\lim_{\varepsilon_k\to 0}\int_D q(\mathbf{x},\mathbf{x}/\varepsilon_k)\,|\nabla u^{\varepsilon_k}|^2\,d\mathbf{x}=\int_D\int_Q q(\mathbf{x},\mathbf{y})\,|P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x})|^2\,d\mathbf{y}\,d\mathbf{x}. \tag{2.4}$$

For multiphase composites with coefficients described by (1.2) we restrict our attention to the inside of each phase and state the following lemma.

LEMMA 2.2 (localization lemma in multiphase composites). *Let $q(\mathbf{x},\mathbf{y})$ be any test function in V^r; then one can pass to the limit $\varepsilon_k\to 0$ in the sequence of solutions $\{u^{\varepsilon_k}\}_{\varepsilon_k>0}$ of (1.6) to obtain*

$$\begin{aligned}&\lim_{\varepsilon_k\to 0}\int_D q(\mathbf{x},\mathbf{x}/\varepsilon_k)\,\chi_i^{\varepsilon_k}(\mathbf{x})\,|\nabla u^{\varepsilon_k}|^2\,d\mathbf{x}\\&=\int_D\int_Q q(\mathbf{x},\mathbf{y})\,\chi_i(\mathbf{x},\mathbf{y})\,|P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x})|^2\,d\mathbf{y}\,d\mathbf{x}.\end{aligned} \tag{2.5}$$

The proofs of Lemmas 2.1 and 2.2 are given at the end of this section.

To illustrate the ideas, we use Lemma 2.2 to establish Theorem 1.3, noting that Theorem 1.2 follows from Lemma 2.1 in the same way.

Proof of Theorem 1.3. For each $\varepsilon_k > 0$, we apply Hölder's inequality to the left side of (2.5) to obtain

$$\begin{aligned}&\int_D \int_Q q(\mathbf{x},\mathbf{y})\,\chi_i(\mathbf{x},\mathbf{y})\,|P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x})|^2\,d\mathbf{y}\,d\mathbf{x}\\ &\le \lim_{\varepsilon_k\to 0}\left(\int_D |q(\mathbf{x},\mathbf{x}/\varepsilon_k)|^r\,dx\right)^{1/r} \liminf_{\varepsilon_k\to 0}\left(\int_D \chi_i^{\varepsilon_k}(\mathbf{x})\,|\nabla u^{\varepsilon_k}|^{2r'}\,d\mathbf{x}\right)^{1/r'}. \qquad (2.6)\end{aligned}$$

Noting [14] that

$$\lim_{\varepsilon_k\to 0}\left(\int_D |q(\mathbf{x},\mathbf{x}/\varepsilon_k)|^r\,dx\right)^{1/r} = \left(\int_D\int_Q |q(\mathbf{x},\mathbf{y})|^r\,d\mathbf{y}\,d\mathbf{x}\right)^{1/r} \equiv \|q(\mathbf{x},\mathbf{y})\|_{L^r(D\times Q)}, \qquad (2.7)$$

we obtain

$$\frac{\int_D\int_Q q(\mathbf{x},\mathbf{y})\,\chi_i(\mathbf{x},\mathbf{y})\,|P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x})|^2\,d\mathbf{y}\,d\mathbf{x}}{\|q(\mathbf{x},\mathbf{y})\|_{L^r(D\times Q)}} \le \liminf_{\varepsilon_k\to 0}\left(\int_D \chi_i^{\varepsilon_k}(\mathbf{x})|\nabla u^{\varepsilon_k}|^{2r'}d\mathbf{x}\right)^{1/r'}. \qquad (2.8)$$

Since V^r is dense in $L^r(D\times Q)$, we substitute an approximation of

$$\chi_i(\mathbf{x},\mathbf{y})\,|P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x})|^{2r'-2}$$

for q in (2.8) to find that

$$\left(\int_D\int_Q \chi_i(\mathbf{x},\mathbf{y})\,|P(\mathbf{x},\mathbf{y})\nabla u^H(\mathbf{x})|^{2r'}d\mathbf{y}\,d\mathbf{x}\right)^{1/r'} \le \liminf_{\varepsilon_k\to 0}\left(\int_D \chi_i^{\varepsilon_k}(\mathbf{x})|\nabla u^{\varepsilon_k}|^{2r'}d\mathbf{x}\right)^{1/r'}. \qquad (2.9)$$

Theorem 1.3 follows for $2 < p < \infty$ upon taking the square root on both sides of (2.9). The case $p = 2$ follows immediately upon choosing $q(\mathbf{x},\mathbf{y}) = 1$ in Lemma 2.2.

We conclude by providing the proof of Lemma 2.2; note that the proof of Lemma 2.1 is identical.

Proof of Lemma 2.2. The indicator function of the set of interest D is denoted by $\chi_D(\mathbf{x})$. We choose a test function $q(\mathbf{x},\mathbf{y})$ in V^r and set $p(\mathbf{x},\mathbf{y}) = \chi_D(\mathbf{x})\chi_i(\mathbf{x},\mathbf{y})q(\mathbf{x},\mathbf{y})$. For $\delta\beta > 0$ we form the perturbed conductivity tensor $\tilde{A}_{ij}(\mathbf{x},\mathbf{y}) = A_{ij}(\mathbf{x},\mathbf{y}) + \delta\beta p(\mathbf{x},\mathbf{y})\delta_{ij}$. We choose $\delta\beta$ sufficiently small so that $\tilde{A}(\mathbf{x},\mathbf{y})$ satisfies (1.3). By construction, $\tilde{A}(\mathbf{x},\mathbf{x}/\varepsilon_k)$ is measurable and satisfies (1.4) and (1.5). Consider the associated solution $\tilde{u}^{\varepsilon_k}$ in V of the weak formulation of the boundary value problem given by

$$\int_\Omega \tilde{A}(\mathbf{x},\mathbf{x}/\varepsilon_k)\nabla\tilde{u}^{\varepsilon_k}\cdot\nabla\varphi\,d\mathbf{x} = \int_\Omega f\varphi\,d\mathbf{x} + \int_{\partial\Omega_N} g\varphi\,ds \quad \text{for every } \varphi \text{ in } V. \qquad (2.10)$$

Set $\tilde{u}^{\varepsilon_k} = u^{\varepsilon_k} + \delta u^{\varepsilon_k}$; subtraction of (2.1) from (2.10) gives

$$\int_\Omega \tilde{A}(\mathbf{x},\mathbf{x}/\varepsilon_k)\nabla\delta u^{\varepsilon_k}\cdot\nabla\varphi\,d\mathbf{x} + \int_\Omega \delta\beta\,p(\mathbf{x},\mathbf{x}/\varepsilon_k)\nabla u^{\varepsilon_k}\cdot\nabla\varphi\,d\mathbf{x} = 0. \qquad (2.11)$$

Choosing $\varphi = u^{\varepsilon_k}$ in (2.11) and applying the identity

$$\int_\Omega A(\mathbf{x}, \mathbf{x}/\varepsilon_k)\nabla u^{\varepsilon_k} \cdot \nabla \delta u^{\varepsilon_k}\, d\mathbf{x} = \int_\Omega f \delta u^{\varepsilon_k}\, d\mathbf{x} + \int_{\partial\Omega_N} g\delta u^{\varepsilon_k}\, ds \tag{2.12}$$

gives

$$\delta\beta \times \int_\Omega p(\mathbf{x}, \mathbf{x}/\varepsilon_k)|\nabla u^{\varepsilon_k}|^2\, d\mathbf{x} + T^{\varepsilon_k} = -\int_\Omega f\delta u^{\varepsilon_k}\, d\mathbf{x} - \int_{\partial\Omega_N} g\delta u^{\varepsilon_k}\, ds, \tag{2.13}$$

where

$$T^{\varepsilon_k} = \delta\beta \times \int_\Omega p(\mathbf{x}, \mathbf{x}/\varepsilon_k)(\nabla \delta u^{\varepsilon_k}) \cdot \nabla u^{\varepsilon_k}\, d\mathbf{x}. \tag{2.14}$$

Next set $\varphi = \delta u^{\varepsilon_k}$ in (2.11); it then follows from Cauchy's inequality and (1.3) that

$$\|\nabla \delta u^{\varepsilon_k}\|_{L^2(\Omega)} \leq C\delta\beta, \tag{2.15}$$

where here and throughout C denotes a generic constant independent of ε_k. From this it is evident that

$$|T^{\varepsilon_k}| < C\delta\beta^2. \tag{2.16}$$

Next we pass to the $\varepsilon_k \to 0$ limit and apply Theorems 1.1 and 2.1 to find that the sequence $\{\tilde{u}^{\varepsilon_k}\}_{\varepsilon_k>0}$ converges weakly in $H^1(\Omega)$ to $\tilde{u}^H$, where $(\tilde{u}^H, \tilde{u}_1)$ is the solution in $V \times L^2[\Omega; H^1_{per}(Q)/\mathbf{R}]$ of

$$\begin{aligned} &\int_\Omega \int_Q \tilde{A}(\mathbf{x}, \mathbf{y})(\nabla \tilde{u}^H(\mathbf{x}) + \nabla_{\mathbf{y}}\tilde{u}_1(\mathbf{x}, \mathbf{y})) \cdot (\nabla\varphi(\mathbf{x}) + \nabla_{\mathbf{y}}\varphi_1(\mathbf{x}, \mathbf{y}))\, d\mathbf{y}\, d\mathbf{x} \\ &= \int_\Omega f\varphi\, d\mathbf{x} + \int_{\partial\Omega_N} g\varphi\, ds \end{aligned} \tag{2.17}$$

for every (φ, φ_1) in $V \times L^2[\Omega; H^1_{per}(Q)/\mathbf{R}]$. Set $\tilde{u}^H - u^H = \delta u^H$, $\tilde{u}_1 - u_1 = \delta u_1$; subtraction of (2.2) from (2.17) gives

$$\begin{aligned} &\int_\Omega \int_Q \tilde{A}(\mathbf{x}, \mathbf{y})(\nabla \delta u^H(\mathbf{x}) + \nabla_{\mathbf{y}}\delta u_1(\mathbf{x}, \mathbf{y})) \cdot (\nabla\varphi(\mathbf{x}) + \nabla_{\mathbf{y}}\varphi_1(\mathbf{x}, \mathbf{y}))\, d\mathbf{y}\, d\mathbf{x} \\ &+ \int_\Omega \int_Q \delta\beta p(\mathbf{x}, \mathbf{y})(\nabla u^H(\mathbf{x}) + \nabla_{\mathbf{y}}u_1(\mathbf{x}, \mathbf{y})) \cdot (\nabla\varphi(\mathbf{x}) + \nabla_{\mathbf{y}}\varphi_1(\mathbf{x}, \mathbf{y}))\, d\mathbf{y}\, d\mathbf{x} = 0. \end{aligned} \tag{2.18}$$

Choosing $(\varphi, \varphi_1) = (u^H, u_1)$ in (2.18) together with the identity

$$\begin{aligned} &\int_\Omega \int_Q A(\mathbf{x}, \mathbf{y})(\nabla u^H(\mathbf{x}) + \nabla_{\mathbf{y}}u_1(\mathbf{x}, \mathbf{y})) \cdot (\nabla \delta u^H(\mathbf{x}) + \nabla_{\mathbf{y}}\delta u_1(\mathbf{x}, \mathbf{y}))\, d\mathbf{y}\, d\mathbf{x} \\ &= \int_\Omega f\delta u^H\, d\mathbf{x} + \int_{\partial\Omega_N} g\delta u^H\, ds \end{aligned} \tag{2.19}$$

gives

$$\begin{aligned} &\delta\beta \times \int_\Omega \int_Q p(\mathbf{x}, \mathbf{y})|P(\mathbf{x}, \mathbf{y})\nabla u^H(\mathbf{x})|^2\, d\mathbf{y}\, d\mathbf{x} + \tilde{T} \\ &= -\int_\Omega f\delta u^H\, d\mathbf{x} - \int_{\partial\Omega_N} g\delta u^H\, ds, \end{aligned} \tag{2.20}$$

where

$$\tilde{T} = \delta\beta \times \int_\Omega \int_Q p(\mathbf{x},\mathbf{y})(\nabla \delta u^H + \nabla_{\mathbf{y}} \delta u_1(\mathbf{x},\mathbf{y})) \cdot (\nabla u^H + \nabla_{\mathbf{y}} u_1(\mathbf{x},\mathbf{y}))\, d\mathbf{x}. \tag{2.21}$$

Next set $(\varphi, \varphi_1) = (\delta u^H, \delta u_1)$ in (2.18); it then follows from Cauchy's inequality and (1.3) that

$$\|\nabla \delta u^H + \nabla_y \delta u_1\|_{L^2(\Omega \times Q)} \leq C\delta\beta, \tag{2.22}$$

and it follows easily that

$$|\tilde{T}| < C\delta\beta^2. \tag{2.23}$$

Taking the $\varepsilon_k \to 0$ limit in (2.13), noting that $\lim_{\varepsilon_k \to 0} \delta u^{\varepsilon_k} = \delta u^H$ (weakly in $H^1(\Omega)$), and recalling (2.16) gives

$$\delta\beta \times \lim_{\varepsilon_k \to 0} \int_\Omega p(\mathbf{x}, \mathbf{x}/\varepsilon_k) |\nabla u^{\varepsilon_k}|^2\, d\mathbf{x} + O(\delta\beta^2) = -\int_\Omega f \delta u^H\, d\mathbf{x} - \int_{\partial\Omega_N} g \delta u^H\, ds. \tag{2.24}$$

Lemma 2.2 now follows immediately from (2.20), (2.23), and (2.24) and from identifying like powers of $\delta\beta$.

3. Explicit lower bounds for aggregates of Schulgasser crystallites. In this section we derive the lower bound (1.22) for the microstructure consisting of Schulgasser crystallites embedded within a homogeneous matrix with unit thermal conductivity. The temperature field inside the unit period cell $\Phi^i(\mathbf{y}) = w^i(\mathbf{y}) + \mathbf{y}_i$ is the solution of the local problem

$$\operatorname{div}_{\mathbf{y}} \left(A(\mathbf{y})(\nabla_{\mathbf{y}} w^i(\mathbf{y}) + \mathbf{e}^i) \right) = 0, \tag{3.1}$$

with w^i Q-periodic and $\int_Q w^i(\mathbf{y})\, d\mathbf{y} = 0$. For this microstructure, $A(\mathbf{y})$ is given by (1.20) for $\mathbf{y}$ in $B(\mathbf{y}^\ell, r_\ell)$ and $A(\mathbf{y}) = I$ outside. Here we restrict λ_2 to the interval $(1/2, 1)$ and choose λ_1 so that $\lambda_1 = 1/(2\lambda_2 - 1)$. A calculation shows that the solution $\Phi^i(\mathbf{y})$ is given by

$$\Phi^i = \begin{cases} \mathbf{y}_i, & \mathbf{y} \in Q \setminus \cup_{\ell=1}^N B(\mathbf{y}^\ell, r_\ell), \\ r_\ell^{1-\alpha} |\mathbf{y} - \mathbf{y}^\ell|^{\alpha-1} (\mathbf{y}_i - \mathbf{y}_i^\ell) + \mathbf{y}_i^\ell, & \mathbf{y} \in B(\mathbf{y}^\ell, r_\ell), \end{cases} \tag{3.2}$$

where $\alpha = 2\lambda_2 - 1$. The corrector matrix $P(\mathbf{y})$ is given by

$$P(\mathbf{y}) = \begin{cases} I, & \mathbf{y} \in Q \setminus \cup_{\ell=1}^N B(\mathbf{y}^\ell, r_\ell), \\ r_\ell^{1-\alpha} |\mathbf{y} - \mathbf{y}^\ell|^{\alpha-1} (I + (\alpha - 1)\mathbf{n} \otimes \mathbf{n}), & \mathbf{y} \in B(\mathbf{y}^\ell, r_\ell), \end{cases} \tag{3.3}$$

where $\mathbf{n} = (\mathbf{y} - \mathbf{y}^\ell)/|\mathbf{y} - \mathbf{y}^\ell|$ for $\mathbf{y} \in B(\mathbf{y}^\ell, r_\ell)$. A direct calculation shows that

$$A^E = \int_Q A(\mathbf{y}) P(\mathbf{y})\, d\mathbf{y} = I. \tag{3.4}$$

Next we provide the lower bound for $\int_\Omega \int_Q |P(\mathbf{y}) \nabla u^H(\mathbf{x})|^p\, d\mathbf{y}\, d\mathbf{x}$. Note for any η in $\mathbf{R}^3$ that $P^T(\mathbf{y})P(\mathbf{y})\eta \cdot \eta = |P(\mathbf{y})\eta|^2$. The smallest eigenvalue $\lambda(\mathbf{y})$ of $P^T(\mathbf{y})P(\mathbf{y})$ delivers the lower bound $\lambda(\mathbf{y})|\eta|^2 \leq |P(\mathbf{y})\eta|^2$ and

$$\int_\Omega \int_Q \lambda(\mathbf{y})^{p/2} |\nabla u^H(\mathbf{x})|^p\, d\mathbf{y}\, d\mathbf{x} \leq \int_\Omega \int_Q |P(\mathbf{y}) \nabla u^H(\mathbf{x})|^p\, d\mathbf{y}\, d\mathbf{x}. \tag{3.5}$$

Calculation shows that

$$\lambda(\mathbf{y}) = \alpha^2 r_\ell^{2(1-\alpha)} |\mathbf{y} - \mathbf{y}^\ell|^{2(\alpha-1)} \tag{3.6}$$

for $\mathbf{y} \in B(\mathbf{y}^\ell, r_\ell)$ and $\lambda(\mathbf{y}) = 1$ for $\mathbf{y} \in Q \setminus \cup_\ell^N B(\mathbf{y}^\ell, r_\ell)$. The lower bound (1.22) follows upon substitution of (3.6) into (3.5).

4. Optimality of the lower bounds. Conditions are presented on f and $A(\mathbf{y})$ for which the lower bound (1.14) is attained for a range of exponents $2 < p < \hat{q}$. We suppose, as in Avellaneda and Lin [3], that Ω is a $C^{1,\alpha}$ domain ($0 < \alpha \leq 1$) and suppose for $0 < \gamma \leq 1$, $0 < C$, that $A(\mathbf{y}) \in C^\gamma(\mathbf{R}^n)$ and $\|A(\mathbf{y})\|_{C^\gamma(\mathbf{R}^n)} \leq C$. We set $A^{\varepsilon_k} = A(\mathbf{x}/\varepsilon_k)$. For $\delta > 0$ suppose $2 \leq q \leq n + \delta$ and $f \in L^q$ and set $1/\hat{q} = 1/q - 1/(n+\delta)$. Given these choices, we consider the $W_0^{1,2}(\Omega)$ solutions u^{ε_k} of

$$-\text{div}\,(A^{\varepsilon_k}(\mathbf{x})\nabla u^{\varepsilon_k}) = f \quad \text{in} \quad \Omega. \tag{4.1}$$

Theorem 4 of [3] shows that there exists a constant independent of ε_k for which

$$\|\nabla u^{\varepsilon_k}\|_{L^{\hat{q}}(\Omega)} \leq C \|f\|_{L^q(\Omega)} \tag{4.2}$$

holds for every $\varepsilon_k > 0$. Subject to these hypotheses it will be shown that the lower bound (1.14) is attained for $p < \hat{q}$.

Passing to a subsequence if necessary we start by considering the Young measure ν associated with the sequence $\{P(\mathbf{x}/\varepsilon_k)\nabla u^H(\mathbf{x})\}_{\varepsilon_k > 0}$. Here ν is represented by a family of probability measures $\nu = \{\nu_{\mathbf{x}}\}_{\mathbf{x}\in\Omega}$ depending measurably on $\mathbf{x}$. We denote by $C_0(\mathbf{R}^n)$ the set of continuous functions φ defined on $\mathbf{R}^n$ such that $\lim_{\eta\to\infty} \varphi(\eta) = 0$. Elementary arguments show that

$$\langle \nu_{\mathbf{x}}, \varphi \rangle = \int_{\mathbf{R}^n} \varphi(\eta) d\nu_{\mathbf{x}}(\eta) = \int_Q \varphi(P(\mathbf{z})\nabla u^H(\mathbf{x})) d\mathbf{z} \quad \text{a.e. } \mathbf{x} \in \Omega, \tag{4.3}$$

for every φ in $C_0(\mathbf{R}^n)$. From corrector theory [17] there exists an exponent $r \geq 1$ for which one has the strong convergence

$$\lim_{\varepsilon_k \to 0} \|\nabla u^{\varepsilon_k} - P(\mathbf{x}/\varepsilon_k)\nabla u^H\|_{L^r(\Omega)} = 0. \tag{4.4}$$

The strong convergence (4.4) shows that both sequences

$$\{\nabla u^{\varepsilon_k}\}_{\varepsilon_k > 0} \quad \text{and} \quad \{P(\mathbf{x}/\varepsilon_k)\nabla u^H(\mathbf{x})\}_{\varepsilon_k > 0}$$

share the same Young measure; see, for example, Lemma 6.3 of [20]. From (4.2) it follows, on passage to a subsequence if necessary, that $\{|\nabla u^{\varepsilon_k}|^p\}_{\varepsilon_k}$ is weakly convergent in $L^1(\Omega)$; thus,

$$\lim_{\varepsilon_k \to 0} \int_D |\nabla u^{\varepsilon_k}|^p \, d\mathbf{x} = \int_D \int_{\mathbf{R}^n} |\eta|^p \, d\nu_{\mathbf{x}}(\eta) \, d\mathbf{x} = \int_D \int_Q |P(\mathbf{z})\nabla u^H(\mathbf{x})|^p \, d\mathbf{z} \, d\mathbf{x}, \tag{4.5}$$

and optimality follows. Last, it follows immediately from Proposition 6.5 of [20] that for every Carathéodory function $\psi(\mathbf{x}, \eta)$ satisfying the growth condition (1.25) (on passage to a subsequence if necessary)

$$\lim_{\varepsilon_k \to 0} \int_D \psi(\mathbf{x}, \nabla u^{\varepsilon_k}) \, d\mathbf{x} = \int_D \int_{\mathbf{R}^n} \psi(\mathbf{x}, \eta) \, d\nu_{\mathbf{x}}(\eta) \, d\mathbf{x}, \tag{4.6}$$

and (1.26) follows since (4.3) implies that

$$\int_D \int_{\mathbf{R}^n} \psi(\mathbf{x}, \eta) \, d\nu_{\mathbf{x}}(\eta) \, d\mathbf{x} = \int_D \int_Q \psi(\mathbf{x}, P(\mathbf{z})\nabla u^H(\mathbf{x})) \, d\mathbf{z} \, d\mathbf{x}. \tag{4.7}$$

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] G. Allaire and M. Briane, *Multi-scale convergence and reiterated homogenization*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 297–342.

[3] M. Avellaneda and F. H. Lin, *Compactness methods in the theory of homogenization*, Comm. Pure Appl. Math., 40 (1987), pp. 806–847.

[4] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North-Holland, Amsterdam, 1978.

[5] D. Cioranescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 99–104.

[6] J. H. Gosse and S. Christensen, *Strain Invariant Failure Criteria for Polymers in Composite Materials*, AIAA paper 2001-1184, American Institute of Aeronautics and Astronautics, Reston, VA, 2001.

[7] C. E. Gutierrez and I. Peral, *A harmonic analysis theorem and applications to homogenization*, Indiana Univ. Math. J., 50 (2001), pp. 1651–1674.

[8] A. Kelly and N. H. Macmillan, *Strong Solids*, Monogr. Phys. Chem. Mater., Clarendon Press, Oxford, 1986.

[9] Y. Y. Li and M. Vogelius, *Gradient estimates for solutions to divergence form elliptic equations with discontinuous coefficients*, Arch. Ration. Mech. Anal., 153 (2000), pp. 91–151.

[10] Y. Y. Li and L. Nirenberg, *Estimates for elliptic systems from composite material*, Comm. Pure Appl. Math., 56 (2003), pp. 892–925.

[11] R. Lipton, *Assessment of the local stress state through macroscopic variables*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 361 (2003), pp. 921–946.

[12] R. Lipton, *Homogenization theory and the assessment of extreme field values in composites with random microstructure*, SIAM J. Appl. Math., 65 (2004), pp. 475–493.

[13] R. Lipton, *Stress constrained G closure and relaxation of structural design problems*, Quart. Appl. Math., 62 (2004), pp. 295–321.

[14] D. Luckkassen, G. Nguetseng, and P. Wall, *Two-scale convergence*, Int. J. Pure Appl. Math., 2 (2002), pp. 35–86.

[15] N. Meyers, *An L^p estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.

[16] G. W. Milton, *Modelling the properties of composites by laminates*, in Homogenization and Effective Moduli of Materials and Media, IMA Vol. Math. Appl. 1, J. L. Ericksen, D. Kinderleher, R. Kohn, and J.-L. Lions, eds., Springer-Verlag, New York, 1986.

[17] F. Murat and L. Tartar, *H-convergence*, mimeographed notes, Séminaire d'Analyse Fonctionnelle et Numérique de l'Université d'Alger, 1978 (in French). English translation in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev and R. V. Kohn, eds., Birkhäuser Boston, Boston, MA, 1997, pp. 21–43.

[18] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[19] N. J. Pagano and F. G. Yuan, *On the significance of effective modulus theory (homogenization) in composite laminate mechanics*, Comp. Sci. Tech., 60 (2000), pp. 2471–2488.

[20] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser Verlag, Basel, 1997.

[21] P. Raghaven, S. Moorthy, S. Ghosh, and N. J. Pagano, *Revisiting the composite laminate problem with an adaptive multi-level computational model*, Comp. Sci. Tech., 61 (2001), pp. 1017–1040.

[22] K. Schulgasser, *Sphere assemblage model for polycrystals and symmetric materials*, J. Appl. Phys., 54 (1983), pp. 1380–1382.

[23] B. Schweizer, *Uniform estimates in two periodic problems*, Comm. Pure Appl. Math., 53 (2000), pp. 1153–1176.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1060–1074

ON THE SEMIRELATIVISTIC HARTREE-TYPE EQUATION*

YONGGEUN CHO† AND TOHRU OZAWA†

Abstract. We study the global Cauchy problem and scattering problem for the semirelativistic Hartree-type equation in $\mathbb{R}^n$, $n \geq 1$, with nonlocal nonlinearity $F(u) = \lambda(|x|^{-\gamma} * |u|^2)u$, $0 < \gamma < n$. We prove the existence and uniqueness of global solutions for $0 < \gamma < \frac{2n}{n+1}$, $n \geq 2$ or $\gamma > 2$, $n \geq 3$, and the nonexistence of asymptotically-free solutions for $0 < \gamma \leq 1$, $n \geq 3$. We also specify asymptotic behavior of solutions as the mass tends to zero and infinity.

Key words. semirelativistic Hartree-type equation, global solution, scattering, nonexistence of asymptotically free solutions

AMS subject classifications. 35Q40, 35Q55, 47J35

DOI. 10.1137/060653688

1. Introduction.

In this paper we consider the following Cauchy problem:

$$\begin{cases} i\partial_t u = \sqrt{m^2 - \Delta}u + F(u) \ \text{ in } \ \mathbb{R}^n \times \mathbb{R}, \qquad n \geq 1, \\ u(x,0) = \varphi(x) \ \text{ in } \ \mathbb{R}^n, \end{cases} \tag{1.1}$$

where $m > 0$ denotes the mass of bosons in units $\hbar = c = 1$, $F(u)$ is a nonlinear functional of Hartree type such that $F(u) = (V_\gamma * |u|^2)u$, where $*$ denotes the convolution in $\mathbb{R}^n$, $V_\gamma(x) = \lambda|x|^{-\gamma}$ for some fixed constant $\lambda \in \mathbb{R}$, and $0 < \gamma < n$.

Equation (1.1) is called a semirelativistic Hartree equation which was used to describe Boson stars. See [8, 9, 20] and the references therein.

The purpose of this paper is to establish the local and global existence theory to (1.1) and the scattering theory of the global solutions. In this paper we study the Cauchy problem (1.1) in the form of the integral equation

$$u(t) = U(t)\varphi - i\int_0^t U(t-t')F(u)(t')dt', \tag{1.2}$$

where

$$U(t)\varphi(x) = (e^{-it\sqrt{m^2-\Delta}}\varphi)(x) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{i(x\cdot\xi - t\sqrt{m^2+|\xi|^2})}\widehat{\varphi}(\xi)\,d\xi.$$

Here $\widehat{\varphi}$ denotes the Fourier transform of φ such that $\widehat{\varphi}(\xi) = \int_{\mathbb{R}^n} e^{-ix\cdot\xi}\varphi(x)\,dx$.

One of the key tools for the existence and scattering is the conservation law. If the solution u of (1.1) has sufficient decay at infinity and smoothness, it satisfies two conservation laws:

$$\begin{aligned} \|u(t)\|_{L^2} &= \|\varphi\|_{L^2}, \\ E(u) \equiv K_m(u) + V(u) &= E(\varphi), \end{aligned} \tag{1.3}$$

*Received by the editors March 5, 2006; accepted for publication (in revised form) July 10, 2006; published electronically November 21, 2006.
http://www.siam.org/journals/sima/38-4/65368.html

†Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan (ygcho@math.sci.hokudai.ac.jp, ozawa@math.sci.hokudai.ac.jp). The first author was supported by the Japan Society for the Promotion of Science under the JSPS Postdoctoral Fellowship for Foreign Researchers.

where $K_m(u) = \frac{1}{2}\langle\sqrt{m^2-\Delta}\,u, u\rangle$, $V(u) = \frac{1}{4}\langle F(u), u\rangle$, and $\langle,\rangle$ is the complex inner product in L^2. For actual proof of (1.3) a regularizing method is simply applicable, as in [21] in the case of $0 < \gamma \le 1$. For local solutions constructed by a contraction argument based on the Strichartz estimate stated below, the case of $1 < \gamma \le 2$ is treated by exactly the same method as in [29] without using approximate or a regularizing approach.

In section 2, local existence is shown for $0 < \gamma < n$ and $\varphi \in H^s$ with $s \ge \frac{\gamma}{2}$ by the Plancherel theorem and the standard contraction mapping theorem without resorting to a Strichartz estimate. Then we use the conservation laws to obtain the global existence for $s \ge \frac{1}{2}$, $0 < \gamma \le 1$, $n \ge 2$, and $0 < \gamma < 1$, $n = 1$. This result is an extension of the work of Lenzmann [21] in which global well-posedness is considered for a Coulomb-type potential in three space dimensions. From the energy conservation, we get a bound of a solution, which is uniform in the mass m on any finite time interval, if m is bounded from above, and then get a strong convergence of solutions of (1.1) to a solution of the equation without mass. However, if m is large, then the kinetic energy $K_m(u)$ is no longer bounded globally in time. Instead, we can get a uniform bound of local solutions in H^s, provided $s \ge \frac{\gamma}{2}$. Then after a phase modulation, we prove that the modulated solution is closely approximated by a solution of a Schrödinger equation of Hartree type if m is sufficiently large. This phenomenon can be interpreted as a kind of nonrelativistic limit and eventually as a semiclassical or vanishing dispersion limit. See Proposition 2.5 below.

The second tool is the Strichartz estimate. We consider the following Strichartz estimate for the unitary group $U(t)$ (see [22, 23]):

$$\begin{aligned} &\|U(t)\varphi\|_{L_T^{q_0}H_{r_0}^{s_0-\sigma_0}} \lesssim \|\varphi\|_{H^{s_0}}, \\ &\left\|\int_0^t U(t-t')f(t')\,dt'\right\|_{L_T^{q_1}H_{r_1}^{s_1-\sigma_1}} \lesssim \|f\|_{L_T^1 H^{s_1}}, \end{aligned} \tag{1.4}$$

where (q_i, r_i), $i = 0, 1$, satisfy that for any $\theta \in [0, 1]$

$$\begin{gathered} \frac{2}{q_i} = (n-1+\theta)\left(\frac{1}{2}-\frac{1}{r_i}\right), \quad 2\sigma_i = (n+1+\theta)\left(\frac{1}{2}-\frac{1}{r_i}\right), \\ 2 \le q_i, r_i \le \infty, \quad (q_i, r_i) \ne (2, \infty). \end{gathered} \tag{1.5}$$

We call the pair (q, r, σ) satisfying (1.5) the *admissible pair*. If $\theta = 0$, it is called wave admissible, and if $\theta = 1$, then it is called Schrödinger admissible. Here $H_r^s = (1-\Delta)^{-s/2}L^r$ is the usual Sobolev space, and $H^s = H_2^s$. Hereafter, we denote the space $L_T^q(B)$ by $L^q(0, T; B)$ and its norm by $\|\cdot\|_{L_T^q B}$ for some Banach space B, and also denote $L^q(B)$ with norm $\|\cdot\|_{L^q B}$ by $L^q(0, \infty; B)$, $1 \le q \le \infty$.

In section 3, we consider the global existence and scattering in the case when $0 < \gamma < n$. We first show the local existence for $0 < \gamma < n$, $n \ge 1$, and s slightly less than $\frac{\gamma}{2}$ by the Strichartz estimate of non–endpoint wave admissible pairs. Then we extend the local solution to the global one for $0 < \gamma < \frac{2n}{n+1}$ by the energy conservation and continuation procedure. The gain of upper bound $\frac{2n}{n+1}$ follows from the fact that the Sobolev exponent s can be made smaller than $\frac{\gamma}{2}$, which enables us to use the continuation procedure. Second, we get a small data global existence result and scattering for the case when $2 < \gamma < n$ and $n \ge 3$ by using the endpoint Strichartz estimate for the Schrödinger admissible pair.

In the last section, as the usual case of nonlinearity with long range potential, nonexistence of nontrivial asymptotically free solutions is shown for the cases

$0 < \gamma \le 1$, $n \ge 3$, and $0 < \gamma < \frac{n}{2}$, $n = 1, 2$, by a similar method applied to a large class of dispersive equations. See [2, 6, 14, 15, 24, 36].

Until now, it remains open to show the global existence for $\frac{2n}{n+1} \le \gamma \le 2$ as well as the scattering for $1 < \gamma \le 2$. For the blow-up result, we refer to [10] of Fröhlich and Lenzmann in which the finite time blow-up of radially symmetric solutions was shown in the case when $\gamma = 1$, $\lambda < 0$, $n = 3$, and $E(\varphi) < 0$.

There is a large literature on partial differential equations with Hartree-type nonlinearity. We refer the reader to [5, 11, 12, 16, 17, 18, 28, 26, 27] for Schrödinger-related equations and, to [1, 25, 31, 32, 35, 34, 37] for Klein–Gordon-related equations in both massive and massless cases.

If not specified, throughout this paper the notation $A \lesssim B$ and $A \gtrsim B$ denote $A \le CB$ and $A \ge C^{-1}B$, respectively. Different positive constants possibly depending on n, m, λ, and γ might be denoted by the same letter C. $A \sim B$ means that both $A \lesssim B$ and $A \gtrsim B$ hold.

2. Global existence I. In this section, we study the global existence and the limiting problem as $m \to 0$ or as $m \to \infty$ with $0 < \gamma \le 1$.

Let us first introduce the following local existence result.

PROPOSITION 2.1. *Let $0 < \gamma < n$ and $n \ge 1$. Suppose $\varphi \in H^s(\mathbb{R}^n)$ with $s \ge \frac{\gamma}{2}$. Then there exists a positive time T independent of m such that (1.2) has a unique solution $u \in C([0,T]; H^s)$ with $\|u\|_{L^\infty_T H^s} \le C\|\varphi\|_{H^s}$, where C does not depend on m.*

Proof. Let $(X^s_{T,\rho}, d)$ be a complete metric space with metric d defined by

$$X^s_{T,\rho} = \{u \in L^\infty_T(H^s(\mathbb{R}^n)) : \ \|u\|_{L^\infty_T H^s} \le \rho\}, \ \ d(u,v) = \|u - v\|_{L^\infty_T L^2}.$$

Now we define a mapping $N : u \mapsto N(u)$ on $X^s_{T,\rho}$ by

$$N(u)(t) = U(t)\varphi - i\int_0^t U(t-t')F(u)(t')\,dt'. \tag{2.1}$$

Our strategy is to use the standard contraction mapping argument. To do so, let us introduce a generalized Leibniz rule (see Lemmas A1–A4 in the appendix of [19]).

LEMMA 2.2. *For any $s \ge 0$ we have*

$$\|D^s(uv)\|_{L^r} \lesssim \|D^s u\|_{L^{r_1}}\|v\|_{L^{q_2}} + \|u\|_{L^{q_1}}\|D^s v\|_{L^{r_2}},$$

where $D^s = (-\Delta)^{s/2}$

and $\dfrac{1}{r} = \dfrac{1}{r_1} + \dfrac{1}{q_2} = \dfrac{1}{q_1} + \dfrac{1}{r_2}, \quad r_i \in (1,\infty),\ q_i \in (1,\infty], \quad i = 1,2.$

Then for all $u \in X(T,\rho)$ we have

$$\begin{aligned}
\|N(u)\|_{L^\infty_T H^s} &\le \|\varphi\|_{H^s} + T\|F(u)\|_{L^\infty_T H^s} \\
&\lesssim \|\varphi\|_{H^s} + T\Big(\|I_{n-\gamma}(|u|^2)\|_{L^\infty_T L^\infty}\|u\|_{L^\infty_T H^s} \\
&\qquad\qquad + \|I_{n-\gamma}(|u|^2)\|_{L^\infty_T H^s_{\frac{2n}{\gamma}}}\|u\|_{L^\infty_T L^{\frac{2n}{n-\gamma}}}\Big) \\
&\lesssim \|\varphi\|_{H^s} + T\left(\|u\|^2_{L^\infty_T H^{\frac{\gamma}{2}}}\|u\|_{L^\infty_T H^s} + \|u\|^2_{L^\infty_T L^{\frac{2n}{n-\gamma}}}\|u\|_{L^\infty_T H^s}\right) \\
&\lesssim \|\varphi\|_{H^s} + T\|u\|^2_{L^\infty_T H^{\frac{\gamma}{2}}}\|u\|_{L^\infty_T H^s} \lesssim \|\varphi\|_{H^s} + T\rho^3,
\end{aligned} \tag{2.2}$$

where I_α is the fractional integral operator given by $I_\alpha(v)(x) = \int_{\mathbb{R}^n} |x-y|^{\alpha-n} v(y)\,dy$. It is well known that I_α satisfies the inequality (see [33], for instance)

$$\|I_\alpha(\psi)\|_{L^q} \lesssim \|\psi\|_{L^p},\ \frac{1}{q} = \frac{1}{p} - \frac{\alpha}{n}, 1 < p < q < \infty.$$

For the third inequality we used the fractional integral inequality, generalized Leibniz rule (Lemma 2.2), and the fact that

$$\sup_{x\in\mathbb{R}^n} \left| \int_{\mathbb{R}^n} \frac{|u(x-y)|^2}{|y|^\gamma}\,dy \right| \lesssim \|u\|^2_{\dot{H}^{\frac{\gamma}{2}}}. \tag{2.3}$$

For the last one, we used the Sobolev embedding $H^{\frac{\gamma}{2}} \hookrightarrow L^{\frac{2n}{n-\gamma}}$.

If we choose ρ and T such that $\|\varphi\|_{H^s} \le \rho/2$ and $CT\rho^3 \le \rho/2$, then N maps $X^s_{T,\rho}$ to itself.

Now we have only to show that N is a Lipschitz map for sufficiently small T. Let $u, v \in X^s_{T,\rho}$. Then we have

$$\begin{aligned}
&d(N(u), N(v)) \\
&\quad\lesssim T \left\| I_{n-\gamma}(|u|^2)u - I_{n-\gamma}(|v|^2)v \right\|_{L^\infty_T L^2} \\
&\quad\lesssim T \left(\left\| I_{n-\gamma}(|u|^2)(u-v) \right\|_{L^\infty_T L^2} + \left\| I_{n-\gamma}(|u|^2 - |v|^2)v \right\|_{L^\infty_T L^2} \right) \\
&\quad\lesssim T \left(\|u\|^2_{L^\infty_T H^{\frac{\gamma}{2}}} d(u,v) + \|I_{n-\gamma}(|u|^2 - |v|^2)\|_{L^\infty_T L^{\frac{2n}{\gamma}}} \|v\|_{L^\infty_T L^{\frac{2n}{n-\gamma}}} \right) \\
&\quad\lesssim T \left(\rho^2 d(u,v) + \rho \| |u|^2 - |v|^2 \|_{L^\infty_T L^{\frac{2n}{2n-\gamma}}} \right) \\
&\quad\lesssim T \left(\rho^2 + \rho \left(\|u\|_{L^\infty_T L^{\frac{2n}{n-\gamma}}} + \|v\|_{L^\infty_T L^{\frac{2n}{n-\gamma}}} \right) \right) d(u,v) \\
&\quad\lesssim T\rho^2 d(u,v).
\end{aligned}$$

The above estimate implies that the mapping N is a contraction if T is sufficiently small.

The uniqueness and time continuity follows easily from (1.2) and the contraction argument. This completes the proof of the proposition. □

From the conservation laws (1.3), we get the following global well-posedness.

THEOREM 2.3. *Let $0 < \gamma \le 1$ for $n \ge 2$, $0 < \gamma < 1$ for $n = 1$, and $s \ge \frac{1}{2}$. Let T^* be the maximal existence time of the solution u as in Theorem 2.1. Then if $\lambda \ge 0$, or if $\lambda < 0$ and $\|\varphi\|_{L^2}$ is sufficiently small, then $T^* = \infty$. Moreover, $\|u(t)\|_{H^s} \le C\|\varphi\|_{H^s} e^{C(|E(\varphi)| + \|\varphi\|^2_{L^2})t}$, where C does not depend on m.*

Proof. From the estimate (2.3) and L^2 conservation, we have

$$|V(u)| \lesssim \|u\|^2_{\dot{H}^{\frac{\gamma}{2}}} \|u\|^2_{L^2} \lesssim \|u\|^{2\gamma}_{\dot{H}^{\frac{1}{2}}} \|\varphi\|^{4-2\gamma}_{L^2}. \tag{2.4}$$

Hence if $\lambda \ge 0$ or if $\lambda < 0$ and $\|\varphi\|_{L^2}$ is sufficiently small, then

$$\|u(t)\|^2_{\dot{H}^{\frac{\gamma}{2}}} \le C(|E(u)| + \|\varphi\|^2_{L^2}) = C(|E(\varphi)| + \|\varphi\|^2_{L^2}). \tag{2.5}$$

From (2.5) and a similar estimate to (2.2), we have

$$\begin{aligned}\|u(t)\|_{H^s} &\lesssim \|\varphi\|_{H^s} + \int_0^t \|u\|_{H^{\frac{\gamma}{2}}}^2 \|u\|_{H^s}\, dt' \\ &\lesssim \|\varphi\|_{H^s} + (|E(\varphi)| + \|\varphi\|_{L^2}^2) \int_0^t \|u\|_{H^s}\, dt'.\end{aligned}$$

Gronwall's inequality shows that

$$\|u(t)\|_{H^s} \le C\|\varphi\|_{H^s} \exp(C(|E(\varphi)| + \|\varphi\|_{L^2}^2)t).$$

This completes the proof. □

REMARK 1. *If $\lambda < 0$, $0 < \gamma < 2$, and $n \ge 2$, then*

$$\begin{aligned}|V(u)| &\lesssim \|V_\gamma * |u|^2\|_{L^n} \|u\|_{L^{\frac{2n}{n-1}}}^2 \lesssim \|u\|_{L^{\frac{2n}{n-\gamma+1}}}^2 \|u\|_{L^{\frac{2n}{n-1}}}^2 \\ &\lesssim \|u\|_{L^2}^\theta \|u\|_{L^{\frac{2n}{n-1}}}^{4-\theta} = \|\varphi\|_{L^2}^\theta \|u\|_{\dot H^{\frac12}}^{4-\theta}\end{aligned}$$

for some small positive number $\theta < 2$. Hence

$$E(u) \ge K_m(u) - |V(u)| \ge \frac{m}{2}\|\varphi\|_{L^2}^2 + \frac12 \|u\|_{\dot H^{\frac12}}^2 - C\|\varphi\|_{L^2}^\theta \|u\|_{\dot H^{\frac12}}^{4-\theta}.$$

Thus we can always make $E(u)$ strictly positive, provided $\|\varphi\|_{L^2}$ is sufficiently small, and with smaller $\|\varphi\|_{L^2}$ than in Theorem 2.3, *$|E(\varphi)|$ is changed by $E(\varphi)$. Using* (2.4), *the same argument holds for $n = 1$.*

If m is bounded above, then the energy $E(\varphi)$ is also bounded, and hence the H^s norm of solution u is bounded in a finite time interval uniformly on small m. This enables us to treat a limit problem as $m \to 0$. We have the following. See [1] for related second-order equations.

PROPOSITION 2.4. *If $u_m \in (C \cap L^\infty)(H^s)$ is the global solution of* (1.2) *satisfying the same condition as in Theorem* 2.3, *then for any finite time T, $u_m \to u_0$ in $L_T^\infty(H^s)$ with $s \ge \frac12$ as $m \to 0$, where u_0 is the global solution to the Cauchy problem*

$$i\partial_t u_0 = \sqrt{-\Delta} u_0 + F(u_0), \quad u_0(x,0) = \varphi(x) \in H^s. \tag{2.6}$$

Proof. One can easily show the global existence of (2.6) by the same argument as in the proof of Theorem 2.3. The solution u_0 can be written as

$$u_0(t) = U_0(t)\varphi - i\int_0^t U_0(t-t')F(u_0)(t')\, dt',$$

where $U_0(t) = e^{-it\sqrt{-\Delta}}$.

For any $T > 0$ there exists M such that $\sup_{0<m\le 1}(\|u_m\|_{L_T^\infty H^s} + \|u_0\|_{L_T^\infty H^s}) \le M$. Then observing that for any $t \in [0,T]$

$$\begin{aligned}\|(U(t) - U_0(t))\varphi\|_{H^s} &\le \||1 - e^{it(\sqrt{m^2+|\xi|^2}-|\xi|)}|(1+|\xi|^2)^{\frac{s}{2}}|\widehat{\varphi}|\|_{L^2} \\ &\le t\||\sqrt{m^2+|\xi|^2} - |\xi||(1+|\xi|^2)^{\frac{s}{2}}|\widehat{\varphi}|\|_{L^2} \\ &\le tm\|\varphi\|_{H^s},\end{aligned}$$

we have that

$$
\begin{aligned}
&\|u_m(t)-u_0(t)\|_{H^s}\\
&\le \|(U(t)-U_0(t))\varphi\|_{H^s}\\
&\qquad + \int_0^t \left(\|F(u_m)-F(u_0)\|_{H^s} + \|(U(t-t')-U_0(t-t'))F(u_0)\|_{H^s}\right)\,dt'\\
&\lesssim Tm\|\varphi\|_{H^s} + \int_0^t \|I_{n-\gamma}(|u_m|^2-|u_0|^2)u_m\|_{H^s}\,dt'\\
&\qquad + \int_0^t \|I_{n-\gamma}(|u_0|^2)(u_m-u_0)\|_{H^s})\,dt' + mT\int_0^t \|F(u_0)\|_{H^s}\,dt'.
\end{aligned}
\tag{2.7}
$$

From Lemma 2.2, fractional integration, and the estimate (2.3), it follows that

$$
\begin{aligned}
&\|I_{n-\gamma}(|u_m|^2-|u_0|^2)u_m\|_{H^s}\\
&\lesssim \|I_{n-\gamma}(|u_m|^2-|u_0|^2)\|_{L^\infty}\|u_m\|_{H^s} + \|I_{n-\gamma}(|u_m|^2-|u_0|^2)\|_{H^s_{\frac{2n}{\gamma}}}\|u_m\|_{L^{\frac{2n}{n-\gamma}}}\\
&\lesssim \|u_m-u_0\|_{H^{\frac{\gamma}{2}}}(\|u_m\|_{H^{\frac{\gamma}{2}}}+\|u_0\|_{H^{\frac{\gamma}{2}}})\|u_m\|_{H^s}\\
&\qquad + \||u_m|^2-|u_0|^2\|_{L^{\frac{2n}{2n-\gamma}}}\|u_m\|_{H^{\frac{\gamma}{2}}}\\
&\lesssim M^2\|u_m-u_0\|_{H^s}
\end{aligned}
$$

and similarly that

$$
\begin{aligned}
&\|I_{n-\gamma}(|u_0|^2)(u_m-u_0)\|_{H^s}\\
&\lesssim \|u_0\|^2_{H^{\frac{\gamma}{2}}}\|u_m-u_0\|_{H^s} + \|I_{n-\gamma}(|u_0|^2)\|_{H^s_{\frac{2n}{\gamma}}}\|u_m-u_0\|_{L^{\frac{2n}{n-\gamma}}}\\
&\lesssim M^2\|u_m-u_0\|_{H^s}.
\end{aligned}
$$

Substituting these into (2.7), we have for any $t\in[0,T]$ that

$$
\|u_m(t)-u_0(t)\|_{H^s} \lesssim MTm + M^3mT^2 + M^2\int_0^t \|u_m(t')-u_0(t')\|_{H^s}\,dt'.
$$

Then Gronwall's inequality implies the strong convergence $u_m\to u_0$ in $L^\infty_T(H^s)$. ∎

In the case of large mass, the situation is different. Since $E(u)=E(\varphi)=\frac{1}{2}\langle\sqrt{m^2-\Delta}\varphi,\varphi\rangle+V(\varphi)$ diverges as $m\to\infty$, it is difficult to obtain the uniform bound for $\|u\|_{H^{\frac{1}{2}}}$ from the energy conservation law. However, from Proposition 2.1 we see that the local existence time T and the constant C can be chosen independently of the mass m if $s\ge\frac{\gamma}{2}$. To be more specific, we have $\|u_m(t)\|_{H^s}\le C\|\varphi\|_{H^s}$, where u_m is the solution of the equation with mass m. Now using the phase modulation $v_m=e^{imt}u_m$, the function v_m satisfies the equation

$$
i\partial_t v_m = (\sqrt{m^2-\Delta}-m)v_m + F(v_m),\quad v_m(0)=\varphi,
$$

and equivalently

$$
v_m(t) = U_m(t)\varphi - i\int_0^t U_m(t-t')F(v_m)(t')\,dt',
\tag{2.8}
$$

where $U_m(t)=e^{-it(\sqrt{m^2-\Delta}-m)}$. Let $\widetilde{U}_m$ be the unitary group $e^{it\frac{1}{2m}\Delta}$. As was first observed by Segal [32] at a formal level, we expect that the linear solutions $U_m\varphi$

and $\widetilde{U}_m\varphi$ become very close in the $L^\infty_T(H^s)$ norm if T is finite and $\varphi \in H^s$. That observation is in fact justified by

$$\begin{aligned}
&\|(U_m(t) - \widetilde{U}_m(t))\varphi\|^2_{L^\infty_T H^s} \\
&\quad \le \sup_{0<t<T} \int_{|\xi|\le\sqrt{m}} \left|1 - e^{it(\sqrt{m^2+|\xi|^2}-m-\frac{|\xi|^2}{2m})}\right|^2 (1+|\xi|)^{2s}|\widehat{\varphi}|^2\,d\xi \\
&\qquad + 2\int_{|\xi|\ge\sqrt{m}} (1+|\xi|)^{2s}|\widehat{\varphi}|^2\,d\xi \\
&\quad \le T\int \left|2m/(\sqrt{m^2+|\xi|^2}+m) - 1\right|^2 (1+|\xi|)^{2s}|\widehat{\varphi}|^2\,d\xi \\
&\qquad + 2\int_{|\xi|\ge\sqrt{m}} (1+|\xi|)^{2s}|\widehat{\varphi}|^2\,d\xi \\
&\quad \to 0 \ \text{ as } \ m\to\infty.
\end{aligned}$$

Hence we can expect that v_m is very close to a function w_m in $L^\infty_T(H^s)$, where w_m is a solution of the nonlinear Schrödinger equation:

$$i\partial_t w_m = -\frac{1}{2m}\Delta w_m + F(w_m), \ \ w_m(0) = \varphi. \tag{2.9}$$

Of course, by the same argument as the proof of Proposition 2.1, we find T and C independently of m and a unique solution $w_m \in C([0,T];H^s)$ of (2.9) for $s \ge \frac{\gamma}{2}$ such that $\|w_m\|_{L^\infty_T H^s} \le C\|\varphi\|_{H^s}$.

Now let $T^*_{v_m}$ and $T^*_{w_m}$ be the maximal existence times of the solutions u_m and w_m, respectively. Then from the local existence result (Proposition 2.1) we deduce that $T^* \equiv \inf_{m>1}\min(T^*_{v_m}, T^*_{w_m})$ is strictly positive and has the following.

PROPOSITION 2.5. *If $s \ge \frac{\gamma}{2}$ and $T < T^*$, then $v_m - w_m \to 0$ in $L^\infty_T(H^s)$ as $m\to\infty$.*

Proof. First we consider the integral equation

$$u_\infty = \varphi - i\int_0^t F(u_\infty)\,dt',$$

which is equivalent to the ordinary differential equation $i\partial_t u_\infty = F(u_\infty)$, $u_\infty(x,0) = \varphi$. This equation has an exact solution $u_\infty(x,t) = \varphi(x)e^{-i\lambda t(|\cdot|^{-\gamma} * |\varphi|^2)(x)}$ for any $t \ge 0$. If $s \ge \frac{\gamma}{2}$, then the uniqueness of u_∞ is guaranteed.

To prove $v_m - w_m \to 0$ in $L^\infty_T(H^s)$, we have only to prove that $v_m - u_\infty \to 0$ in $L^\infty_T(H^s)$ and $w_m - u_\infty \to 0$ in $L^\infty_T(H^s)$. At first we have

$$\begin{aligned}
&\|v_m(t) - u_\infty(t)\|_{H^s} \\
&\le \|\,(U_m(t)-1)\,\varphi\|_{H^s} + \int_0^t \|(U_m(t-t')-1)F(u_\infty)\|_{H^s}\,dt' \\
&\qquad\qquad + \int_0^t \|F(v_m) - F(u_\infty)\|_{L^\infty}\,dt'
\end{aligned}$$

and

$$\begin{aligned}
&\|\,(U_m(t)-1)\,\varphi\|_{H^s}^2\\
&=\int\left|e^{-it(\sqrt{m^2+|\xi|^2}-m)}-1\right|^2|\widehat{\varphi}(\xi)|^2\,d\xi\\
&=\int_{|\xi|\le m^{\frac14}}+\int_{|\xi|>m^{\frac14}}\\
&\le\int_{|\xi|\le m^{\frac14}}\frac{t^2|\xi|^4}{(\sqrt{m^2+|\xi|^2}+m)^2}(1+|\xi|)^{2s}|\widehat{\varphi}(\xi)|^2\,d\xi+4\int_{|\xi|>m^{\frac14}}(1+|\xi|)^{2s}|\widehat{\varphi}(\xi)|^2\,d\xi\\
&=\frac{T^2}{4m}\|\varphi\|_{H^s}^2+4\int_{|\xi|>m^{\frac14}}(1+|\xi|)^{2s}|\widehat{\varphi}(\xi)|^2\,d\xi\\
&\to 0\ \text{ as }\ m\to\infty.
\end{aligned}$$

We take $M=M(T)$ such that $\sup_{m\ge1}(\|v_m\|_{L_T^\infty H^s}+\|w_m\|_{L_T^\infty H^s}+\|u_\infty\|_{L_T^\infty H^s})\le M$. Then since $F(u_\infty)\in L_T^\infty(H^s)$, we have

$$\int_0^T\|(U_m(t-t')-1)F(u_\infty)\|_{H^s}\,dt'\to0\ \text{ as }\ m\to\infty.$$

We also have

$$\|F(v_m)-F(u_\infty)\|_{H^s}\le CM^2\|v_m-u_\infty\|_{H^s}.$$

Thus

$$\|v_m(t)-u_\infty(t)\|_{H^s}\le o(1)+CM^2\int_0^t\|v_m-u_\infty\|_{H^s}\,dt',\tag{2.10}$$

and as for w_m, by the same argument as that of v_m

$$\|w_m(t)-u_\infty(t)\|_{H^s}\le o(1)+CM^2\int_0^t\|w_m-u_\infty\|_{H^s}\,dt'.\tag{2.11}$$

Therefore Gronwall's inequality yields the claim. □

3. Global existence II. In this section, we re-examine the existence result and get a slightly lower regularity by using the Strichartz estimate. The first result is the following local existence for $0<\gamma<n$.

PROPOSITION 3.1. *Let $0<\gamma<n$ and $n\ge2$. Then there is a number α with $0<\alpha<\min(\gamma,\frac{2n}{n+1})$ satisfying that given $s>\frac{\gamma}{2}-\frac{(n-1)\alpha}{4n}$ and $\varphi\in H^s$ there exists a positive time T such that (1.2) has a unique solution $u\in C([0,T];H^s)\cap L_T^q(H_r^{s-\sigma})$, where $q=\frac{4n}{(n-1)\alpha}$, $r=\frac{2n}{n-\alpha}$, and $\sigma=\frac{(n+1)\alpha}{4n}$.*

Proof. Given n and γ, choose a number α with $0<\alpha<\min(\gamma,\frac{2n}{n-1})$ and fix $s>\frac{\gamma}{2}-\frac{(n-1)\alpha}{4n}$. Then for some positive number T to be chosen later, let us define a complete metric space $(Y_{T,\rho}^s,d_T)$ with metric d_T by

$$\begin{aligned}
&Y_{T,\rho}^s=\left\{v\in L_T^\infty(H^s)\cap L_T^q(H_r^{s-\sigma}):\|v\|_{L_T^\infty H^s}+\|v\|_{L_T^qH_r^{s-\sigma}}\le\rho\right\},\\
&d_T(u,v)=\|u-v\|_{L_T^\infty H^s\cap L_T^qH_r^{s-\sigma}},
\end{aligned}$$

where q, r, σ are the same indices as in Proposition 3.1.

From now on, we will prove that the nonlinear mapping N defined as (2.1) is a contraction on $Y^s_{T,\rho}$, provided T is sufficiently small. We will use the following lemma instead of (2.3), which follows by estimating the (fractional) integral inside and outside the ball with radius $R > 0$ separately by Hölder's inequality and by minimizing the resulting estimates with respect to R.

LEMMA 3.2. *Let $0 < \gamma < n$. Then for any $0 < \varepsilon < n - \gamma$ we have*

$$\big\|I_{n-\gamma}(|u|^2)\big\|_{L^\infty} \lesssim \|u\|_{L^{\frac{2n}{n-\gamma-\varepsilon}}} \|u\|_{L^{\frac{2n}{n-\gamma+\varepsilon}}}.$$

If we take $\theta = 0$ in the Strichartz estimate (1.4), then the pair $(q, r, \sigma) = (\frac{4n}{(n-1)\alpha}, \frac{2n}{n-\alpha}, \frac{(n+1)\alpha}{4n})$ becomes an admissible one. Hence the Strichartz estimate together with the Plancherel theorem, Lemma 3.2, and generalized Leibniz rules (Lemma 2.2), enables us to deduce that for sufficiently small ε

$$\begin{aligned}
&\|N(u)\|_{L^\infty_T H^s \cap L^q_T H^{s-\sigma}_r} \\
&\quad\lesssim \|\varphi\|_{H^s} + \|F(u)\|_{L^1_T H^s} \\
&\quad\lesssim \|\varphi\|_{H^s} + \|I_{n-\gamma}(|u|^2)\|_{L^1_T L^\infty} \|u\|_{L^\infty_T H^s} \\
&\qquad\qquad + \int_0^T \|I_{n-\gamma}(|u|^2)\|_{H^s_{\frac{2n}{\gamma+\varepsilon}}} \|u\|_{L^{\frac{2n}{n-(\gamma+\varepsilon)}}} \, dt \\
&\quad\lesssim \|\varphi\|_{H^s} + \|u\|_{L^2_T L^{\frac{2n}{n-(\gamma+\varepsilon)}}} \|u\|_{L^2_T L^{\frac{2n}{n-(\gamma-\varepsilon)}}} \|u\|_{L^\infty_T H^s} \\
&\qquad\qquad + \int_0^T \||u|^2\|_{H^s_{\frac{2n}{2n-(\gamma-\varepsilon)}}} \|u\|_{L^{\frac{2n}{n-(\gamma+\varepsilon)}}} \, dt \\
&\quad\lesssim \|\varphi\|_{H^s} + \|u\|_{L^2_T L^{\frac{2n}{n-(\gamma+\varepsilon)}}} \|u\|_{L^2_T L^{\frac{2n}{n-(\gamma-\varepsilon)}}} \|u\|_{L^\infty_T H^s}.
\end{aligned} \tag{3.1}$$

Using Hölder's inequality for time integral, we have

$$\begin{aligned}
&\|N(u)\|_{L^\infty_T H^s \cap L^q_T H^{s-\sigma}_r} \\
&\quad\lesssim \|\varphi\|_{H^s} + T^{1-\frac{2}{q}} \|u\|_{L^q_T L^{\frac{2n}{n-(\gamma+\varepsilon)}}} \|u\|_{L^q_T L^{\frac{2n}{n-(\gamma-\varepsilon)}}} \|u\|_{L^\infty_T H^s}.
\end{aligned} \tag{3.2}$$

Now if we choose $\varepsilon > 0$ so small that $\varepsilon < \min(\gamma - \alpha, 2(s + \frac{(n-1)\alpha}{4n}) - \gamma)$, then since

$$\frac{2n}{n-\alpha} \le \frac{2n}{n-(\gamma-\varepsilon)} < \frac{2n}{n-(\gamma+\varepsilon)} \le \frac{2n}{n-\alpha-2(s-\sigma)},$$

we have from (3.2) and the Sobolev embedding $H^{s-\sigma}_r \hookrightarrow L^r \cap L^{\frac{2n}{n-\alpha-2(s-\sigma)}}$ that

$$\begin{aligned}
\|N(u)\|_{L^\infty_T H^s \cap L^q_T H^{s-\sigma}_r} &\le C(\|\varphi\|_{H^s} + T^{1-\frac{2}{q}} \|u\|_{L^\infty_T H^s} \|u\|^2_{L^q_T H^{s-\sigma}_r}) \\
&\le C(\|\varphi\|_{H^s} + T^{1-\frac{2}{q}} \rho^3)
\end{aligned}$$

for some constant C. Here we used the conventional embedding that if $2(s-\sigma) \ge n-\alpha$, then $H^{s-\sigma}_r \hookrightarrow L^{r_1}$ for any $r_1 \ge r$. Thus if we choose ρ and T so that $C\|\varphi\|_{H^s} \le \frac{\rho}{2}$ and $CT^{1-\frac{2}{q}}\rho^3 \le \frac{\rho}{2}$, then we conclude that N maps from $Y^s_{T,\rho}$ to itself.

For any $u, v \in Y^s_{T,\rho}$, we have

$$\begin{aligned}
d_T(N(u), N(v)) &\lesssim \|F(u) - F(v)\|_{L^1_T H^s} \\
&\lesssim \|I_{n-\gamma}(|u|^2 - |v|^2)u\|_{L^1_T H^s} + \|I_{n-\gamma}(|v|^2)(u-v)\|_{L^1_T H^s}.
\end{aligned} \tag{3.3}$$

By Lemma 3.2 and Hölder's inequality, we have for sufficiently small $\varepsilon > 0$

$$
\begin{aligned}
&\|I_{n-\gamma}(|u|^2-|v|^2)u\|_{L_T^1 H^s} \\
&\quad\lesssim \|I_{n-\gamma}(|u|^2-|v|^2)\|_{L_T^2 L^\infty}\|u\|_{L_T^\infty H^s} \\
&\qquad\quad + \|I_{n-\gamma}(|u|^2-|v|^2)\|_{L_T^2 H^s_{\frac{2n}{\gamma+\varepsilon}}}\|u\|_{L_T^2 L^{\frac{2n}{n-(\gamma+\varepsilon)}}} \\
&\quad\lesssim \rho\||u|^2-|v|^2\|^{\frac12}_{L_T^1 L^{\frac{n}{n-(\gamma+\varepsilon)}}}\||u|^2-|v|^2\|^{\frac12}_{L_T^1 L^{\frac{n}{n-(\gamma-\varepsilon)}}} \\
&\qquad + \rho\|u-v\|_{L_T^\infty H^s}\left(\|u\|_{L_T^2 L^{\frac{2n}{n-(\gamma-\varepsilon)}}}+\|v\|_{L_T^2 L^{\frac{2n}{n-(\gamma-\varepsilon)}}}\right) \\
&\qquad + \rho\|u-v\|_{L_T^2 L^{\frac{2n}{n-(\gamma-\varepsilon)}}}(\|u\|_{L_T^\infty H^s}+\|v\|_{L_T^\infty H^s}).
\end{aligned}
\tag{3.4}
$$

Now by another Hölder's inequality with respect to the time variable, we have

$$
\|I_{n-\gamma}(|u|^2-|v|^2)u\|_{L_T^1 H^s} \lesssim T^{1-\frac{2}{q}}\rho^2 d_T(u,v).
$$

Similarly,

$$
\begin{aligned}
&\|I_{n-\gamma}(|v|^2)(u-v)\|_{L_T^1 H^s} \\
&\lesssim \|I_{n-\gamma}(|v|^2)\|_{L_T^1 L^\infty}\|u-v\|_{L_T^\infty H^s} + \|I_{n-\gamma}(|v|^2)\|_{L_T^2 L^{\frac{2n}{\gamma+\varepsilon}}}\|u-v\|_{L_T^2 L^{\frac{2n}{n-(\gamma+\varepsilon)}}} \\
&\lesssim \|v\|_{L_T^2 L^{\frac{2n}{n-(\gamma-\varepsilon)}}}\|v\|_{L_T^2 L^{\frac{2n}{n-(\gamma+\varepsilon)}}} d_T(u,v) \\
&\quad + \|v\|_{L_T^\infty H^s}\|v\|_{L_T^2 L^{\frac{2n}{n-(\gamma-\varepsilon)}}}\|u-v\|_{L_T^2 L^{\frac{2n}{n-(\gamma+\varepsilon)}}}.
\end{aligned}
\tag{3.5}
$$

Hence we get

$$
\|I_{n-\gamma}(|v|^2)(u-v)\|_{L_T^1 H^s} \lesssim T^{1-\frac{2}{q}}\rho^2 d_T(u,v).
$$

Substituting these two estimates into (3.3) and then using the fact that $CT^{1-\frac{2}{q}}\rho^2 \le \frac12$ for small T, we conclude that N is a contraction mapping. ∎

REMARK 2. *If we follow the proof above with the Schrödinger admissible pairs, we conclude that Proposition* 3.1 *holds for* $n \ge 3$, $0 < \alpha < \gamma$, $\alpha \le 2$, $s > \frac{\gamma}{2} - \frac{n-2}{4n}\alpha$, $q = \frac{4}{\alpha}$, $r = \frac{2n}{n-\alpha}$, *and* $\sigma = \frac{(n+2)\alpha}{4n}$.

Now we show that the local solutions can be extended globally in time by using the energy conservation law.

THEOREM 3.3. *Let* $0 < \gamma < \frac{2n}{n+1}$, $n \ge 2$. *Then there exists an* α *with* $0 < \alpha < \gamma$ *such that if* $\varphi \in H^{\frac12}$ *and if* $\lambda > 0$, *or* $\lambda < 0$ *but* $\|\varphi\|_{L^2}$ *is sufficiently small, then* (1.2) *has a unique solution* $u \in C([0,\infty); H^{\frac12}) \cap L^q_{loc}(H_r^{\frac12-\sigma})$, *where* $q = \frac{4n}{(n-1)\alpha}$, $r = \frac{2n}{n-\alpha}$, *and* $\sigma = \frac{(n+1)\alpha}{4n}$.

Proof. Let T^* be the maximal existence time. We will prove that T^* is infinite by contradiction. Suppose that $T^* < \infty$. Then the local theory shows that $\|u\|_{L^q_{T^*} H_r^{\frac12-\sigma}} = \infty$. Since $\gamma < 2$, from the local existence proposition, Proposition 3.1, we see that the energy conservation law (1.3) holds. Thus at any $t < T^*$, the solution u satisfies that for $\lambda > 0$

$$
\frac12\|u(t)\|^2_{H^{\frac12}} \le \frac12\|u(t)\|^2_{L^2} + E(u) = \frac12\|\varphi\|^2_{L^2} + E(\varphi)
$$

and for $\lambda < 0$

$$\begin{aligned}\frac{1}{2}\|u(t)\|^2_{H^{\frac{1}{2}}} &\le \frac{1}{2}\|u(t)\|^2_{L^2} + |E(u)| + |V(u)| \\ &\le \frac{1}{2}\|\varphi\|^2_{L^2} + |E(\varphi)| + C\|u\|^2_{L^{\frac{2n}{n-\gamma+1}}}\|u\|^2_{H^{\frac{1}{2}}} \\ &\le \frac{1}{2}\|\varphi\|^2_{L^2} + |E(\varphi)| + C\|u\|^{2-\gamma}_{L^2}\|u\|^{1+\gamma}_{H^{\frac{1}{2}}} \\ &= \frac{1}{2}\|\varphi\|^2_{L^2} + |E(\varphi)| + C\|\varphi\|^{2-\gamma}_{L^2}\|u\|^{1+\gamma}_{H^{\frac{1}{2}}}.\end{aligned}$$

Hence by Young's inequality and the smallness of $\|\varphi\|_{L^2}$

$$\|u(t)\|^2_{H^{\frac{1}{2}}} \le C(\|\varphi\|^2_{L^2} + |E(\varphi)|). \tag{3.6}$$

From the estimates (3.6) and (3.2), which are used with $s = \frac{1}{2}$, we have

$$\|u\|_{L^q_T H^{\frac{1}{2}-\sigma}_r} \lesssim \|\varphi\|^2_{L^2} + |E(\varphi)| + T^{1-\frac{2}{q}}(\|\varphi\|^2_{L^2} + |E(\varphi)|)^{\frac{1}{2}}\|u\|^2_{L^q_T H^{\frac{1}{2}-\sigma}_r}.$$

Thus for sufficiently small T depending on $\|\varphi\|^2_{L^2} + |E(\varphi)|$,

$$\|u\|_{L^q(T_{j-1},T_j;H^{\frac{1}{2}-\sigma}_r)} \le C(\|\varphi\|^2_{L^2} + |E(\varphi)|),$$

where $T_j - T_{j-1} = T$ for $j \le k-1$ and $T_k = T^*$. This means that

$$\|u\|^q_{L^q(0,T^*;H^{\frac{1}{2}-\sigma}_r)} \le \sum_{1\le j\le k}\|u\|^q_{L^q(T_{j-1},T_j;H^{\frac{1}{2}-\sigma}_r)} \le (kC(\|\varphi\|^2_{L^2} + |E(\varphi)|))^q < \infty.$$

This is the contradiction to the hypothesis $T^* < \infty$.

The condition $\gamma < \frac{2n}{n+1}$ is necessary for the existence of α satisfying $s = \frac{1}{2} > \frac{\gamma}{2} - \frac{(n-1)\alpha}{4n}$, and $\alpha < \gamma$. This completes the proof. □

REMARK 3. *If we choose $\theta = 1$, then we deduce the same result as in Theorem* 3.3 *with $0 < \gamma < \frac{2n}{n+2}$, $q = \frac{4}{\alpha}$, $r = \frac{2n}{n-\alpha}$, and $\sigma = \frac{(n+2)\alpha}{4n}$.*

Now we consider the small data global existence and scattering for $2 < \gamma < n$.

THEOREM 3.4. *Let $2 < \gamma < n$, $n \ge 3$, and $s > \frac{\gamma}{2} - \frac{n-2}{2n}$. Then there exists $\rho > 0$ such that for any $\varphi \in H^s$ with $\|\varphi\|_{H^s} \le \rho$,* (1.2) *has a unique solution $u \in (C \cap L^\infty)(H^s) \cap L^2(H^{s-\frac{n+2}{2n}}_{\frac{2n}{n-2}})$. Moreover, there is $\varphi^+ \in H^s$ such that*

$$\|u(t) - U(t)\varphi^+\|_{H^s} \to 0 \ \text{ as } \ t \to \infty.$$

Proof. We will use the Strichartz estimate (1.4) with $\theta = 1$ and endpoint admissible pair $(q, r, \sigma) = (2, \frac{2n}{n-2}, \frac{n+2}{2n})$ (see Remark 2).

Let us define a complete metric space (Y^s_ρ, d) with metric d by

$$\begin{aligned}&Y^s_\rho = \left\{v \in L^\infty(H^s) \cap L^2(H^{s-\sigma}_r) : \|v\|_{L^\infty H^s \cap L^2 H^{s-\sigma}_r} \le \rho\right\}, \\ &d(u,v) = \|u - v\|_{L^\infty H^s \cap L^2 H^{s-\sigma}_r},\end{aligned}$$

Then from the estimate (3.2), we have

$$\|N(u)\|_{L^\infty H^s \cap L^2 H^{s-\sigma}_r} \le C\|\varphi\|_{H^s} + C\|u\|^2_{L^2 H^{s-\sigma}_r}\|u\|_{L^\infty H^s}.$$

If we choose sufficiently small ρ such that $C\|\varphi\|_{H^s} \le \frac{\rho}{2}$ and $C\rho^3 \le \frac{\rho}{2}$, then N maps Y^s_ρ to itself. Similarly, from (3.3)–(3.5), one can show that $d(N(u), N(v)) \le \frac{1}{2}d(u,v)$. This proves the existence part.

To prove the scattering, let us define a function φ^+ by

$$\varphi^+ = \varphi - i\int_0^\infty U(-t')F(u)(t')\,dt'.$$

Then since the solution u is in Y^s_ρ, $\varphi^+ \in H^s$, therefore it holds that

$$\begin{aligned}\|u(t) - u^+(t)\|_{H^s} &\lesssim \int_t^\infty \|F(u)\|_{H^s}\,dt' \\ &\lesssim \|u\|_{L^\infty H^s}\int_t^\infty \|u\|^2_{H^{s-\sigma}_r}\,dt' \to 0 \ \text{ as } \ t\to\infty. \qquad \square\end{aligned}$$

4. Nonexistence of scattering. We prove the nonexistence of nontrivial asymptotically free solution.

THEOREM 4.1. *Assume that $0 < \gamma \le 1$ for $n \ge 3$ and $0 < \gamma < \frac{n}{2}$ for $n = 1, 2$. Suppose that u is a smooth global solution in $C(0,\infty; H^{\frac12}) \cap C^1(0,\infty; H^{-\frac12})$ to (1.1) and there exists a smooth function $\varphi^+ \in H^{\frac12} \cap B^{\frac{n+2}{2}}_{1,1}$ such that*

$$\|u(t) - u^+(t)\|_{L^2} \to 0 \quad as \quad t\to\infty,$$

where $u^+(t) = U(t)\varphi^+$. Then $u = u^+ = 0$. Here $B^s_{1,1}$ is the standard inhomogeneous Besov space.

Proof. Let us define a function of $H(t)$ by

$$H(t) = sgn(\lambda)\mathrm{Re}\langle u(t), u^+(t)\rangle.$$

Then from the condition of u and u^+, $H(t)$ is uniformly bounded on t and

$$\frac{d}{dt}H(t) = |\lambda|\mathrm{Im}\langle I_{n-\gamma}(|u|^2)u, u^+\rangle. \tag{4.1}$$

Suppose $\varphi^+ \ne 0$. Then we derive a contradiction to the uniform boundedness of H on t.

The integration in (4.1) is rewritten as

$$\langle I_{n-\gamma}(|u|^2)u, u^+\rangle = J_1 + J_2 + J_3,$$

where

$$\begin{aligned}J_1 &= \langle I_{n-\gamma}(|u^+|^2)u^+, u^+\rangle, \\ J_2 &= \langle I_{n-\gamma}(|u|^2 - |u^+|^2)u^+, u^+\rangle, \\ J_3 &= \langle I_{n-\gamma}(|u|^2)(u-u^+), u^+\rangle.\end{aligned}$$

To estimate each J_i, we need the following time decay estimate; for its proof see [4, 13, 30].

LEMMA 4.2. *If $\varphi^+ \in B^{\frac{n+2}{2}}_{1,1}$, then*

$$\|U(t)\varphi^+\|_{L^\infty} \lesssim t^{-\frac{n}{2}}\|\varphi^+\|_{B^{\frac{n+2}{2}}_{1,1}}.$$

As for J_2, from Lemma 3.2, we have

$$\begin{aligned}
&|J_2(t)| \\
&\quad = |\langle (|u|^2 - |u^+|^2), I_{n-\gamma}(|u^+|^2)\rangle| \\
&\quad \le \|u-u^+\|_{L^2}(\|u\|_{L^2}+\|u^+\|_{L^2})\|I_{n-\gamma}(|u^+|^2)\|_{L^\infty} \\
&\quad \lesssim \|u-u^+\|_{L^2}(\|u\|_{L^2}+\|u^+\|_{L^2})\|u^+\|_{L^{\frac{2n}{n-\gamma-\varepsilon}}}\|u^+\|_{L^{\frac{2n}{n-\gamma+\varepsilon}}} \\
&\quad \lesssim \|u-u^+\|_{L^2}(\|u\|_{L^2}+\|u^+\|_{L^2})\|u^+\|_{L^2}^{\frac{n-\gamma-\varepsilon}{n}+\frac{n-\gamma+\varepsilon}{n}}\|u^+\|_{L^\infty}^{2-\frac{n-\gamma-\varepsilon}{n}+\frac{n-\gamma+\varepsilon}{n}} \\
&\quad \lesssim \|u-u^+\|_{L^2}(\|u\|_{L^2}+\|u^+\|_{L^2})\|u^+\|_{L^2}^{\frac{2(n-\gamma)}{n}}\|u^+\|_{L^\infty}^{\frac{2\gamma}{n}}.
\end{aligned} \tag{4.2}$$

For the fourth inequality we used Hölder's inequality

$$\|u\|_{L^r} \le \|u\|_{L^2}^{\frac{2}{r}}\|u\|_{L^\infty}^{1-\frac{2}{r}}. \tag{4.3}$$

Now from Lemma 4.2 we get

$$|J_2(t)| = o(t^{-\gamma}). \tag{4.4}$$

Since $\gamma \le 1$ for $n \ge 3$ and $\gamma < \frac{n}{2}$ for $n = 1, 2$, we can take $\varepsilon > 0$ such that $\gamma + \varepsilon < \frac{n}{2}$. Hence by the same argument for J_2 we have for J_3 that

$$\begin{aligned}
|J_3(t)| &= |\langle |u|^2, I_{n-\gamma}((\overline{u}-\overline{u^+})u^+)\rangle| \\
&\le \|u\|_{L^2}^2\|I_{n-\gamma}((\overline{u}-\overline{u^+})u^+)\|_{L^\infty} \\
&\lesssim \|u\|_{L^2}^2\left\||(\overline{u}-\overline{u^+})u^+|^{\frac12}\right\|_{L^{\frac{2n}{n-\gamma-\varepsilon}}}\left\||(\overline{u}-\overline{u^+})u^+|^{\frac12}\right\|_{L^{\frac{2n}{n-\gamma+\varepsilon}}} \\
&\lesssim \|u\|_{L^2}^2\left\|(\overline{u}-\overline{u^+})u^+\right\|_{L^{\frac{n}{n-\gamma-\varepsilon}}}^{\frac12}\left\|(\overline{u}-\overline{u^+})u^+\right\|_{L^{\frac{n}{n-\gamma+\varepsilon}}}^{\frac12} \\
&\lesssim \|u\|_{L^2}^2\|u-u^+\|_{L^2}\|u^+\|_{L^{\frac{2n}{n-2(\gamma+\varepsilon)}}}^{\frac12}\|u^+\|_{L^{\frac{2n}{n-2(\gamma-\varepsilon)}}}^{\frac12} \\
&\lesssim \|u\|_{L^2}^2\|u-u^+\|_{L^2}\|u^+\|_{L^2}^{\frac{n-2\gamma}{n}}\left\|u^+\right\|_{L^\infty}^{1-\frac{n-2\gamma}{n}} \\
&= o(t^{-\gamma}).
\end{aligned} \tag{4.5}$$

As for J_1, if $|x| \le At$ for some $A > 1$, then for any $t > 0$

$$\begin{aligned}
I_{n-\gamma}(|u^+|^2)(x) &\ge \int_{|y|\le At}|x-y|^{-\gamma}|u^+(y)|^2\,dy \\
&\ge \frac{1}{(2At)^\gamma}\int_{|y|\le At}|u^+(y)|^2\,dy.
\end{aligned}$$

Now we prove

$$\int_{|y|\le At}|u^+(y)|^2\,dy \gtrsim \|\varphi^+\|_{L^2}^2 \tag{4.6}$$

for large t, provided φ^+ is sufficiently smooth. Choose a large R such that $\|\eta_R\varphi^+\|_{L^2}^2 \ge \frac23\|\varphi^+\|_{L^2}^2$, where η_R is a smooth cut-off function supported in the ball of radius $2R$ with center at the origin. Then

$$\|u^+\|_{L^2(|x|\le At)}^2 \ge \left|\|U(t)(\eta_R\varphi^+)\|_{L^2(|x|\le At)}^2 - \|\varphi^+\|_{L^2(|x|>R)}^2\right|.$$

Since the linear solution u^+ has the finite propagation speed (actually speed 1), one can easily show that $|U(t)(\eta_R\varphi^+)(x)| \lesssim |x|^{-N}\|\varphi^+\|_{L^2}$ for any N, provided $|x| > 1+2R+t$. Hence we deduce that if $N > \frac{n}{2}$ and t is large enough so that $At > 1+3R+t$, then

$$\begin{aligned}\|U(t)(\eta_R\varphi^+)\|^2_{L^2(|x|\le At)} &= \|U(t)(\eta_R\varphi^+)\|^2_{L^2} - \int_{|x|>At} |U(t)\eta_R\varphi^+|^2\,dx \\ &\ge \|\eta_R\varphi^+\|_{L^2} - C\int_{|x|>At} |x|^{-2N}\,dx\|\varphi^+\|^2_{L^2} \\ &\ge \frac{2}{3}\|\varphi^+\|^2_{L^2} - C(At)^{n-2N}\|\varphi^+\|^2_{L^2}.\end{aligned}$$

Therefore for t large enough,

$$\|u^+\|^2_{L^2(|x|\le At)} \ge \frac{1}{3}\|\varphi^+\|^2_{L^2}$$

and hence

$$J_1(t) \gtrsim t^{-\gamma}. \tag{4.7}$$

Now combining (4.7) with (4.4) and (4.5), we deduce that for t sufficiently large

$$\frac{d}{dt}H(t) \gtrsim t^{-\gamma} \ge t^{-1}.$$

This is a contradiction to the uniform boundedness of $H(t)$ on t. ∎

Acknowledgments. The authors would like to thank Professor Kenji Nakanishi for enlightening discussions and the referees for their valuable comments which improved the presentation of this paper.

REFERENCES

[1] A. Bachelot, *Convergence dans $L^p(\mathbb{R}^{n+1})$ de la solution de l'équation de Klein–Gordon vers celle de l'équation des ondes*, Ann. Fac. Sci. Toulouse Math., 5 (1986/87), pp. 37–60.

[2] J. E. Barab, *Nonexistence of asymptotically free solutions for a nonlinear Schrödinger equation*, J. Math. Phys., 25 (1984), pp. 3270–3274.

[3] J. Bergh and J. Löfström, *Interpolation Spaces*, Springer-Verlag, New York, 1976.

[4] P. Brenner, *On scattering and everywhere defined scattering operators for nonlinear Klein–Gordon equations*, J. Differential Equations, 56 (1985), pp. 310–344.

[5] T. Cazenave, *Semilinear Schrödinger Equations*, Courant Lecture Notes in Math. 10, New York University, Courant Institute of Mathematical Sciences, New York, American Mathematical Society, Providence, RI, 2003.

[6] Y. Cho and T. Ozawa, *Remarks on modified improved Boussinesq equations in one space dimension*, Proc. Roy. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 462 (2006), pp. 1949–1963.

[7] F. M. Christ and M. I. Weinstein, *Dispersion of small amplitude solution of the generalized Korteweg-de Vries equation*, J. Funct. Anal., 100 (1991), pp. 87–109.

[8] A. Elgart and B. Schlein, *Mean field dynamics of boson stars*, Comm. Pure Appl. Math., to appear. Available online at http://arxiv.org/abs/math-ph/0504051.

[9] J. Fröhlich and E. Lenzmann, *Mean-field limit of quantum Bose gases and nonlinear Hartree equation*, in Séminaire: Équations aux Dérivées Partielles., 2003–2004, Exp. No. xix, École Polytech., Pabsieu, 2004, pp. 1–26.

[10] J. Fröhlich and E. Lenzmann, *Blow-up for nonlinear wave equations describing Boson stars*, Comm. Pure Appl. Math., to appear. Available online at http://arxiv.org/abs/math-ph/0511003.

[11] J. Ginibre and T. Ozawa, *Long range scattering for nonlinear Schrödinger and Hartree equations in space dimension $n \geq 2$*, Comm. Math. Phys., 151 (1993), pp. 619–645.
[12] J. Ginibre and G. Velo, *On a class of nonlinear Schrödinger equations with nonlocal interaction*, Math. Z., 170 (1980), pp. 109–136.
[13] J. Ginibre and G. Velo, *Time decay of finite energy solutions of the nonlinear Klein–Gordon and Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 43 (1985), pp. 399–442.
[14] R. T. Glassey, *On the asymptotic behavior of nonlinear wave equations*, Trans. Amer. Math. Soc., 182 (1973), pp. 187–200.
[15] R. T. Glassey, *Asymptotic behavior of solutions to certain nonlinear Schrödinger–Hartree equations*, Comm. Math. Phys., 53 (1977), pp. 9–18.
[16] N. Hayashi and T. Ozawa, *Smoothing effect for some Schrödinger equations*, J. Funct. Anal., 85 (1989), pp. 307–348.
[17] N. Hayashi and T. Ozawa, *Scattering theory in the weighted $L^2(\mathbb{R}^n)$ spaces for some Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 48 (1988), pp. 17–37.
[18] N. Hayashi and Y. Tsutsumi, *Scattering theory for Hartree type equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987), pp. 187–213.
[19] T. Kato, *On nonlinear Schrödinger equations* II. *H^s-solutions and unconditional well-posedness*, J. Anal. Math., 67 (1995), pp. 281–306.
[20] E. H. Lieb and H.-T. Yau, *The Chandrasekhar theory of stellar collapse as the limit of quantum mechanics*, Comm. Math. Phys., 112 (1987), pp. 147–174.
[21] E. Lenzmann, *Well-Posedness For Semirelativistic Hartree Equations of Critical Type*, http://arxiv.org/abs/math.AP/0505456 (2005).
[22] S. Machihara, K. Nakanishi, and T. Ozawa, *Nonrelativistic limit in the energy space for nonlinear Klein–Gordon equations*, Math. Ann., 322 (2002), pp. 603–621.
[23] S. Machihara, K. Nakanishi, and T. Ozawa, *Small global solutions and the nonrelativistic limit for the nonlinear Dirac equation*, Rev. Mat. Iberoamericana, 19 (2003), pp. 179–194.
[24] A. Matsumura, *On the asymptotic behavior of solutions of semi-linear wave equations*, Publ. Res. Inst. Math. Sci., 12 (1976/77), pp. 169–189.
[25] G. P. Menzala and W. A. Strauss, *On a wave equation with a cubic convolution*, J. Differential Equations, 43 (1982), pp. 93–105.
[26] K. Nakanishi, *Modified wave operators for the Hartree equation with data, image and convergence in the same space*, Comm. Pure Appl. Anal., 1 (2002), pp. 237–252.
[27] K. Nakanishi, *Modified wave operators for the Hartree equation with data, image and convergence in the same space* II, Ann. Inst. H. Poincaré, 3 (2002), pp. 503–535.
[28] H. Nawa and T. Ozawa, *Nonlinear scattering with nonlocal interaction*, Comm. Math. Phys., 146 (1992), pp. 259–275.
[29] T. Ozawa, *Remarks on proofs of conservation laws for nonlinear Schrödinger equations*, Calc. Var. Partial Differential Equations, 25 (2006), pp. 403–408.
[30] H. Pecher, *Nonlinear small data scattering for the wave and Klein–Gordon equation*, Math. Z., 185 (1984), pp. 261–270.
[31] H. Sasaki, *Small data scattering for the Klein–Gordon equation with cubic convolution nonlinearity*, Discrete Contin. Dynam. Systems, 15 (2006), pp. 973–981.
[32] I. Segal, *Space-time decay for solutions of wave equations*, Adv. in Math., 22 (1976), pp. 305–311.
[33] E. M. Stein, *Harmonic Analysis*, Princeton University Press, Princeton, NJ, 1993.
[34] W. A. Strauss, *Nonlinear scattering theory at low energy*, J. Funct. Anal., 41 (1981), pp. 110–133.
[35] W. A. Strauss, *Nonlinear scattering theory at low energy: Sequel*, J. Funct. Anal., 43 (1981), pp. 281–293.
[36] W. A. Strauss, *Nonlinear Wave Equations*, CBMS Reg. Conf. Ser. Math. 73, AMS, Providence, RI, 1989.
[37] K. Tsutaya, *Existence and blow up for a wave equation with a cubic convolution*, in Advances in Analysis, World Scientific Publishing, River Edge, NJ, 2005, pp. 321–325.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1075–1089

A SOLIDIFICATION PHENOMENON IN RANDOM PACKINGS*

L. BOWEN†, R. LYONS†, C. RADIN‡, AND P. WINKLER§

Abstract. We prove that uniformly random packings of copies of a certain simply connected figure in the plane exhibit global connectedness at all sufficiently high densities, but not at low densities.

Key words. sphere packing, solidification

AMS subject classifications. 52C25, 52C17

DOI. 10.1137/050647785

1. Introduction.

The densest way to cover a large area with nonoverlapping unit disks is as in Figure 1, in which the disk centers form the vertices of a triangular lattice.

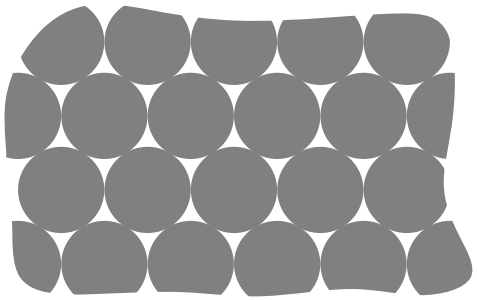


FIG. 1. *The densest packing of unit disks in the plane.*

A *packing* is a collection of congruent copies of a subset with pairwise disjoint interiors. See [5] for a proof that the above packing is indeed the densest possible for unit disks.

It is an old unsolved problem to understand whether densest packings of spheres, simplices, or other shapes, in a Euclidean or hyperbolic space of any dimension, exhibit crystallographic symmetry, such as that of Figure 1. This is the spirit, for instance, of Hilbert's eighteenth problem; see [5, 10] for background.

Using physics models of two- and three-dimensional matter as a guide, we are tempted to try to gain insight into densest packings by considering packings at densities below the maximum. (For an example concerning spheres in $\mathbb{R}^3$, see [9].) In

*Received by the editors December 16, 2005; accepted for publication (in revised form) July 19, 2006; published electronically November 21, 2006. This research was partially supported by the National Science Foundation under grants DMS-0406017 and DMS-0352999. The authors also thank the Banff International Research Station for support at a workshop where this research was begun.
http://www.siam.org/journals/sima/38-4/64778.html

†Department of Mathematics, Indiana University, Bloomington, IN 47405 (lpbowen@indiana.edu, rdlyons@indiana.edu).

‡Department of Mathematics, University of Texas, Austin, TX 78712 (radin@math.utexas.edu).

§Department of Mathematics, Dartmouth College, Hanover, NH 03755 (peter.winkler@dartmouth.edu).

effect, we are emphasizing not so much that densest packings and low density packings differ by their *symmetry*, but rather that they differ in some fundamental geometric fashion. Indeed, it is commonly suggested in the physics literature (see, for instance, [1]) that two-dimensional models of matter do not exhibit crystallographic symmetry, and it is sometimes said by mathematicians that in high-dimensional Euclidean space, densest packings of spheres may not have crystallographic symmetry. So perhaps it is appropriate to re-examine the precise manner in which densest packings differ fundamentally from low density packings, and to use packings at less than optimum density as a guide.

In this work we replace round disks with deformed disks, which are copies of a "zipper" tile; see Figure 2. This tile can cover the plane completely, in which case the packing has density 1, and is completely connected in any sense. What we show is that even at somewhat lower densities, the uniform random packing still has rigid structure; in particular it has a form of connectedness associated with site percolation [7]. What this means for packing large but finite boxes (with torus boundary conditions) is that the necessary gross irregularities of most packings at such high densities disconnect the packings, if at all, along fault lines whose density tends to 0 as the size of the box tends to infinity. Although we define "uniform random packing" of the plane by limits of measures on packings of finite boxes, the key to our proof is to examine isometry-invariant probability measures on packings of the whole plane and to show that the ones that maximize "degrees of freedom per tile" are unique for high densities.

We show that at high density in our model there is a nonzero probability of an infinite linked component, and that this probability is zero at low density. Thus, there are different "phases" of the packings [2]. (This is closely related to continuum percolation, where one looks at overlapping disks with random independent centers, but our methods are quite different.)

Although we believe that such a result also holds for packings of disks or of spheres—pairs of which would be called "linked" if sufficiently close—we are able to prove the result only for our tiles, which are shaped to allow three well-defined levels of pairwise separation. (We discuss the appropriate notion of linking for collections of disks in the last section of the paper.) It is generally understood that true crystalline symmetry is not seen below optimal density in two dimensions—see [11]—so the form of connectedness we use may be useful in understanding the role of geometry in Hilbert's problem.

2. Description of the tile. We consider packings by a deformed disk denoted by t, referred to as "the tile" and depicted in Figure 2. In this section, we define it precisely.

Let H be a regular hexagon of area 1. Let r be the radius of the in-circle of H. Let D be a disk concentric with the in-circle and of radius $r + \rho$, where $0 < \rho \ll 1$ is a number we shall choose more precisely later. We shall construct the shape t by modifying H as follows; D will be called the *shadow disk* of t.

As shown in Figure 2, the tile t equals H with each side modified by a "fringe" and each corner modified by a hook and inlet, where a hook is about half an element of the fringe. As shown in Figure 3, the fringe height is 2ρ. The elements of the fringe have two different size "necks," one of size ρ^2 and one of size $2\rho^2$, allowing neighboring tiles to be linked in either of two well-defined modes, "tightly linked" and "loosely linked"; the former is illustrated in Figure 4 and the latter in Figure 5. We say that two tiles t are *linked* (tightly or loosely) if, when one is held fixed, the other can be moved continuously only by a bounded amount (without overlapping the first). A

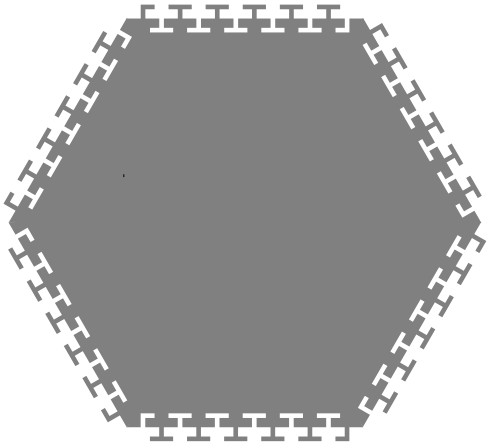


FIG. 2. *The zipper tile.*

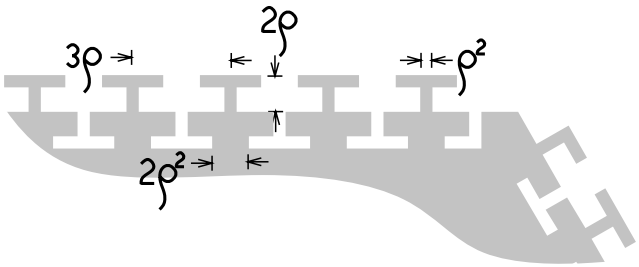


FIG. 3. *Close-up view of the fringe.*

tight link is one that permits no movement of one tile while fixing the other, while a link that is not tight is called *loose*. A key feature of our model is that when two tiles are tightly linked, any motion of one would necessitate a corresponding motion of the other. As we shall explain, the uniform probability distribution on packings of the plane at given density is a limit of such a distribution on packings of larger and larger tori. In our model, these distributions on packings of finite tori are concentrated on packings with the largest number of degrees of freedom, and therefore, roughly speaking, the fewest tiles bound by tight links. This gives us useful control on the packings in the support of our distributions.

A tile is called *fully linked on one side* if it is linked with another tile on that side in such a way that either they are tightly linked, and the line joining their centers goes through the midpoint of the sides of the corresponding hexagons, or they are not tightly linked but can be moved continuously so that their shadow disks touch each other. A tile is *fully linked* if it is fully linked on all sides. We note that the fully tightly linked packing (Figure 4) corresponds to a tiling by the original hexagon and has density 1, and that the tile has area 1 by construction.

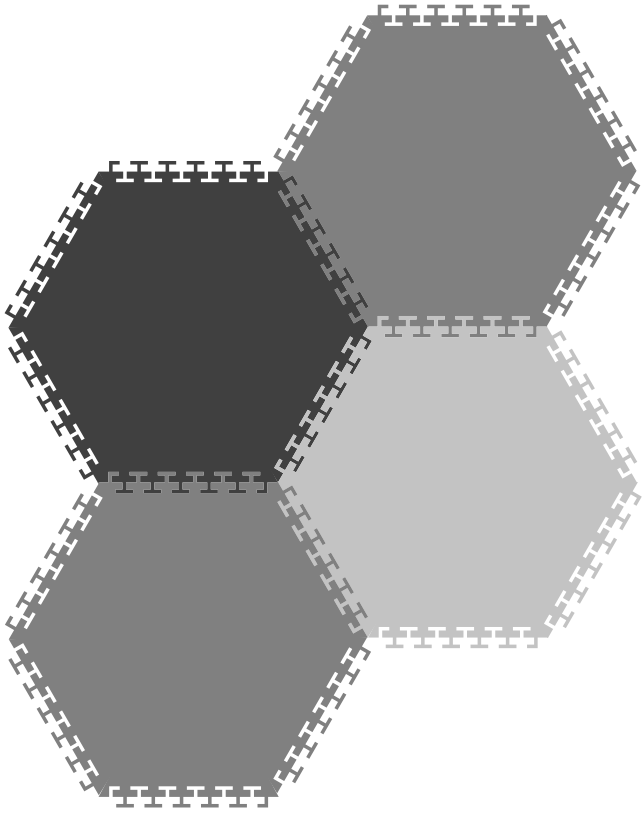


FIG. 4. *Tightly linked tiles.*

3. Statement of results. To state our results we need some notation. Let X be the space of all packings of the plane by the tile t. Given a compact subset K of the plane and two packings of the plane, we consider the distance between the two packings with respect to K to be the Hausdorff distance between the unions of the tiles in the respective packings intersected with K. Then X is endowed with the topology of Hausdorff convergence on compact subsets; X is compact. Intuitively, two packings are close in X if they are close in the Hausdorff sense in a large ball centered at the origin. We shall define a probability measure on X that is "uniform" on the set of all packings of a fixed density. For this, we shall need the space X_n of all packings by the tile of the $n \times n$ torus $\mathbb{R}^2/(n\mathbb{Z})^2$.

For any integer m, let $X_{n,m} \subset X_n$ consist of those packings which contain exactly m tiles ($X_{n,m}$ is empty if m is large enough). To each tile, we assign the set of six unit vectors based on its center and pointing toward the center of each of its edges. Through this assignment, we can view $X_{n,m}$ as a subset of T_n^m/Σ_m, where T_n is the unit tangent bundle of the $n \times n$ torus modulo a $2\pi/6$ rotation and the symmetric group Σ_m acts by permuting the factors.

When m/n^2 is small, $X_{n,m}$ is $(3m)$-dimensional. However, when m/n^2 is sufficiently large, the dimension of $X_{n,m}$ inside T_n^m/Σ_m is less than $3m$. This is because

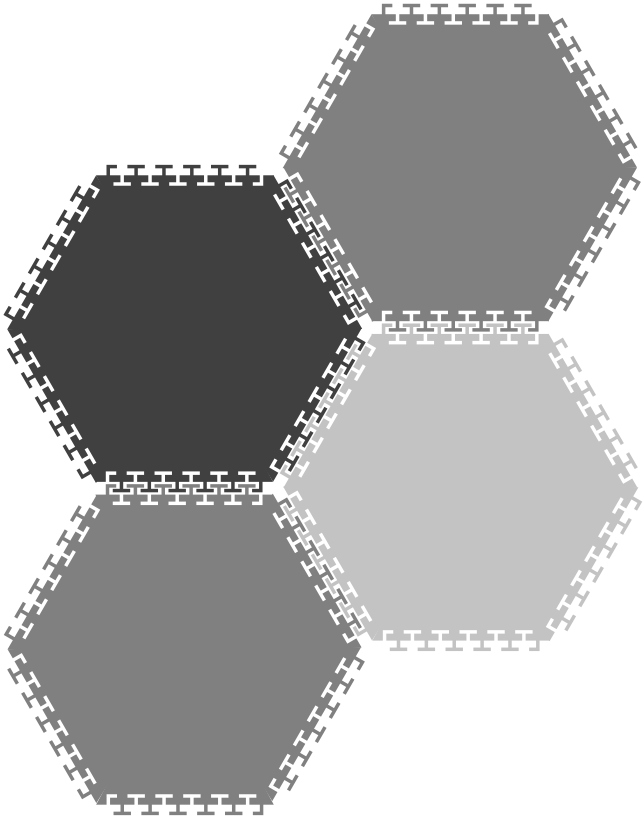


FIG. 5. *Loosely linked tiles.*

at least two tiles in any packing of $X_{n,m}$ will have to be tightly linked, so that it is impossible to move one continuously without moving the other. Thus, it is useful to decompose $X_{n,m}$ into a (finite) disjoint union of sets $X_{n,m,k}$ of packings containing exactly k tight links. Generically, the dimension of $X_{n,m,k}$ is $3(m-k)$. The dimension can be strictly less than this if the packings are jammed in the sense of [4], although this fact will not be important for us. The *top dimension* of $X_{n,m}$ means the maximum dimension of all $X_{n,m,k}$. Let $\mu_{n,m}$ be the probability measure on $X_{n,m}$ obtained by normalizing the Hausdorff measure on $X_{n,m}$ in the top dimension of $X_{n,m}$ with respect to the natural metric inherited from T_n^m/Σ_m. We interpret $\mu_{n,m}$ as being a uniform measure. The fact that $\mu_{n,m}$ is supported on those packings with the fewest tight links will be crucial in the analysis to follow.

Let $\widetilde{X}_n$ be the space of all $(n \times n)$-periodic packings of the plane. In other words, $\widetilde{X}_n$ consists of those packings that are preserved under translations by $n\mathbb{Z}\times n\mathbb{Z}$. Under the quotient map, this space is naturally identified with X_n. Therefore, we can view the measures $\mu_{n,m}$ as living on $\widetilde{X}_n \subset X$.

For a fixed density $d \in [0,1]$, let $\mu^{(d)}$ be any measure obtained as the weak* limit of measures of the form $\mu_{n,m}$ such that $n \to \infty$ and $m/n^2 \to d$. (Note that m/n^2 is the density of every packing in the support of $\mu_{n,m}$ and d is the average density

of a packing chosen with respect to $\mu^{(d)}$; see Lemma 5.1.) A priori, $\mu^{(d)}$ may not be unique, although we shall prove that it is for large enough d.

A *linked component* of a packing is a maximal subpacking in which for every two tiles t, t', there is a sequence $t = t_1, t_2, \ldots, t_n = t'$ such that t_i is linked to t_{i+1} $(i = 1, \ldots, n-1)$. A *tightly linked component* is defined similarly, except that we require t_i to be tightly linked to t_{i+1}.

We say that a measure on the space X of packings is invariant if it is preserved under the full isometry group of the plane. All the measures we consider are probability measures unless stated otherwise.

Let λ_0 be the unique invariant measure on tilings (packings that cover $\mathbb{R}^2$) by our tile. Let λ_1 be the unique invariant measure on packings by t such that all tiles are fully loosely linked, are as close as possible to each other, and the packing has hexagonal symmetry. Write $\lambda_s := s\lambda_1 + (1-s)\lambda_0$.

Our main results are the following.

THEOREM 3.1. *There exists $0 < d_1 < 1$ such that if $d \geq d_1$, $\mu^{(d)}$ is unique and equals λ_s, where $s := (1-d)/(1-d_1)$.*

COROLLARY 3.2. *The $\mu^{(d)}$-probability that the origin is inside a tile belonging to an infinite linked component is nonzero for $d \geq d_1$.*

PROPOSITION 3.3. *For some $d_2 > 0$, the probability (with respect to any $\mu^{(d)}$ for any $d < d_2$) that the origin is inside an infinite linked component is zero.*

4. Tile properties.

LEMMA 4.1. *For small ρ, if tiles t_1 and t_2 are not tightly linked and do not overlap, then the distance between their centers is at least $2r + 2\rho$.*

Proof. Consider the line segment from the center of t_1 to the center of t_2. If this segment traverses near a corner of t_1 or t_2, then it must be longer than $2r + 2\rho$ for small enough ρ. Suppose it crosses a fringe of t_1 and of t_2. If the tiles are not linked, then the claim is obvious. If they are linked, then to minimize the distance, it must be that their fringes match up (so they are fully linked on one side). Thus, they can come closest to each other when pushed flat up against each other so that their shadow disks touch. In this case, the distance between the centers is exactly $2r + \rho$. □

We shall say that two tiles are *densely loosely linked* if they are loosely linked and their shadow disks touch. There is a unique invariant measure on maximally dense packings by congruent disks [3]. Hence the probability measure λ_1 that we defined earlier is the unique invariant measure on packings by t such that all tiles are fully and densely loosely linked. Let d_1 be the density of such a packing.

Given a tile t in a packing P, we denote by $V(t)$ the Voronoi cell of the center of t with respect to the centers of the other tiles; that is, $V(t)$ is the open set of points closer to the center of t than to the center of any other tile. We denote the area of a region A of the plane by $|A|$.

LEMMA 4.2. *The following holds for small enough $\rho > 0$. For any packing P, if $t \in P$ is a tile that has no tight links, then the area of $V(t)$ is at least $1/d_1$. Moreover, equality holds iff the configuration of tiles determining $V(t)$ is congruent to a corresponding configuration of a packing in the support of λ_1.*

Proof. For a tile t, let $H(t)$ denote the hexagon from which t is created. For $x > 0$, let $H_x(t)$ denote the homothetic copy $\frac{r+x}{r} H(t)$ about the center of $H(t)$.

Suppose t is a tile of P without any tight links. Consider the rays $R_1, \ldots, R_6$ from the center of the hexagon $H(t)$ through each of its six vertices. These rays divide the plane into six sectors, $S_1, \ldots, S_6$.

By construction, if t and t_1 are loosely linked, then $|H_\rho(t) \cap H_\rho(t_1)| = \mathrm{O}(\rho^2)$: The hexagon interiors do not intersect if they are parallel, while if they are not parallel, they can intersect only very slightly at a corner. The openings at which a corner can enter have area $\mathrm{O}(\rho^2)$ as $\rho \to 0$.

Thus, we have proved that whenever t is loosely linked in the sector S_i, then $|V(t) \cap S_i| \geq |H_\rho(t)|/6 - \delta_1$, with $\delta_1 = \mathrm{O}(\rho^2)$ as $\rho \to 0$.

Similarly, if t and t_2 are not linked at all, then $|H_{2\rho}(t) \cap H_{2\rho}(t_2)| = \mathrm{O}(\rho^2)$: Again, their interiors do not intersect if they are parallel, while if they are not parallel, they can intersect only very slightly at a corner. So there exists $\delta_2 > 0$ such that whenever t is not linked in the sector S_i, we have $|V(t) \cap S_i| \geq |H_{2\rho}(t)|/6 - \delta_2$, with $\delta_2 = \mathrm{O}(\rho^2)$ as $\rho \to 0$.

Therefore, if t has no tight links but is not fully linked, then

$$|V(t)| \geq j\Big(\frac{|H_\rho(t)|}{6} - \delta_1\Big) + (6-j)\Big(\frac{|H_{2\rho}(t)|}{6} - \delta_2\Big)$$

for some j with $0 \leq j \leq 5$. Given that δ_1, δ_2 are of order ρ^2 while $|H_{2\rho}(t)| - |H_\rho(t)|$ is of order ρ, for ρ small enough we may conclude that $|V(t)| > |H_\rho(t)|$ in this case.

On the other hand, the geometry of a tile is such that for small ρ, if t_1 and t_2 are two tiles loosely linked to t, then t_1 cannot be tightly linked to t_2. Now suppose that t is fully loosely linked. Then the Voronoi cell of the center of t is determined by six tiles $t_1, \ldots, t_6$ all loosely linked to t and all with the property that their shadow disks $D, D_1, \ldots, D_6$ do not overlap (by the previous lemma). It follows [6] that $|V(t)| \geq |H_\rho(t)|$, with equality iff each of the disks $D_1, \ldots, D_6$ touches D. But there is only one way in which this can occur (up to isometry). So $V(t) = H_\rho(t)$ in this case. This implies that the configuration $t, t_1, \ldots, t_6$ is congruent to a corresponding configuration of a packing in the support of λ_1. □

It is easy to see that given $\rho > 0$, there exists $\varepsilon > 0$ such that for any finite component c of tightly linked tiles in any packing, the union V_c of the Voronoi cells of the centers of the tiles of c has area at least $j_c + \varepsilon \mathrm{Per}_c$. Here j_c is the number of tiles in c and Per_c is the perimeter of the union of hexagons corresponding to c. Let ε be the largest such constant. Let $\delta > 0$ be such that the area of the Voronoi cell in the fully densely loosely linked packing equals $1 + \varepsilon \mathrm{Per}_1 + \delta$, where Per_1 is the perimeter of the hexagon of a single tile. Since $\varepsilon = \rho + \mathrm{O}(\rho^2)$ and $\delta = \mathrm{O}(\rho^2)$, we have the following.

LEMMA 4.3. *For sufficiently small ρ, there are $\varepsilon, \delta > 0$ such that for any finite tightly linked component c,*

$$d_1 = \frac{1}{1 + \mathrm{Per}_1 \varepsilon + \delta},$$
$$|V_c| \geq j_c + \varepsilon \mathrm{Per}_c,$$

and

$$\delta \leq \varepsilon/100.$$

5. High density. Recall that X is the compact space of all packings of the plane by the tile (with the topology of Hausdorff-metric convergence on compact subsets). Let $\widetilde{M}$ be the space of isometry-invariant Borel probability measures on X. For any $\mu \in \widetilde{M}$, we denote by $|\mu| := \mu(A_0)$ the *density* of μ, where A_0 is the set of all packings $P \in X$, one of whose tiles contains the origin. Since a tile is the closure of its interior, A_0 is a closed set.

LEMMA 5.1. *If $\mu_i \in \widetilde{M}$ converges to μ in the weak* topology, then $|\mu_i|$ converges to $|\mu|$.*

Proof. Let $\widehat{P}$ denote the union of tiles in a packing, P. For any invariant probability measure ν and any $z \in \mathbb{R}^2$, we have

$$|\nu| = \int \mathbf{1}_{\{0 \in \widehat{P}\}} \, d\nu(P) = \int \mathbf{1}_{\{z \in \widehat{P}\}} \, d\nu(P).$$

Integrating over z in a unit-area disk, D, with respect to the Lebesgue measure and using Fubini's theorem gives the identity $|\nu| = \int |\widehat{P} \cap D| \, d\nu(P)$. Since the function $P \mapsto |\widehat{P} \cap D|$ is continuous on X, the lemma follows. □

Recall that λ_0 is the unique invariant measure on tilings by our tile, so that $|\lambda_0| = 1$. Recalling that d_1 is the density of a fully densely loosely linked tiling, fix a density d with $d_1 \le d \le 1$. Let μ_N be the uniform measure on configuration of tiles at density d_N in an $N \times N$ torus, where $d_N \to d$ as $N \to \infty$. To prove Theorem 3.1, we shall show that the weak* limit of μ_N exists and equals λ_s, where $s := (1-d)/(1-d_1)$.

We shall use several lemmas that depend on the following notation. Given a packing $P \in X$, let

- t_P be the tile of P such that the origin belongs to $V(t_P)$ (this exists as long as the origin is not on the boundary of a Voronoi cell),
- K_P be the tightly linked component containing t_P,
- j_P be the number of tiles in K_P, and
- $f(P) := 3/j_P$ if j_P is finite and t_P contains the origin, and 0 otherwise.

Thus $f(P)$, in a sense, measures the number of degrees of freedom per tile near the origin.

LEMMA 5.2. *If ν is any measure in $\widetilde{M}$, then $\int f \, d\nu(f) \le 3|\nu|$, with equality iff t_P has no tight links for ν-almost every packing P.*

The proof is immediate.

LEMMA 5.3. *If a sequence $\langle \nu_n \rangle \subset \widetilde{M}$ converges to ν in the weak* topology, then $\int f \, d\nu_n$ converges to $\int f \, d\nu$.*

Lemma 5.3 is proved in a manner similar to Lemma 5.1.

Given a finite tightly linked component c, let the congruence class of c be C and let $X_C \subset X$ be the space of all packings P for which t_P exists and K_P is in C. Let X'' be the space of all packings P with density 1, where "density" refers to the usual concept of the limit of the proportion of the area of P inside a large disk centered at the origin as the radius tends to infinity. Let $X' \subset X$ be the space of all packings P such that K_P is infinite and either the density of P is less than 1, the density is not defined, or t_P does not exist. Thus, X is the disjoint union of X', X'' and the collection of X_C for all C.

Let ν be any invariant probability measure with density d. Let ν_C be ν conditioned on X_C, ν' be ν conditioned on X', and ν'' be ν conditioned on X''. Since λ_0 is the only invariant probability measure with support in X'', we have $\nu'' = \lambda_0$. Thus,

$$\nu = \nu(X')\nu' + \nu(X'')\lambda_0 + \sum_C \nu(X_C)\nu_C.$$

Define the density $|\omega| := \omega(A_0)$ as before, but for any (invariant or noninvariant) probability measure ω on X. We have

$$d = |\nu| = \nu(X') \, |\nu'| + \nu(X'') + \sum_C \nu(X_C) \, |\nu_C|$$

and

$$\int f\,d\nu = \sum_C \nu(X_C) \int f\,d\nu_C = \sum_C \nu(X_C)\frac{3}{j_C}|\nu_C|,$$

where j_C is the number of tiles in C.

LEMMA 5.4. *Let $\nu \in \widetilde{M}$ and C be a finite-component class. Suppose that $0 \le s \le 1$ is such that $|\lambda_s| = |\nu_C|$. Then $\int f\,d\lambda_s \ge \int f\,d\nu_C$. Moreover, equality holds only if $j_C = 1$.*

Proof. As in the proof of Lemma 5.1, we have that $|\nu_C| = \int j_C/|V(t_P)|\,d\nu_C(P)$.

First suppose that $j_C = 1$. Then $|\nu_C| = \int 1/|V(t_P)|\,d\nu_C(P) \le d_1$ by Lemma 4.2. This means that $s = 1$ and $\int f\,d\nu_C = 3\,|\nu_C| = 3\,|\lambda_s| = \int f\,d\lambda_s$.

Now assume that $j = j_C > 1$ and put $p := \mathrm{Per}_C$. By definition,

$$\int f\,d\lambda_s = s\int f\,d\lambda_1 + (1-s)\int f\,d\lambda_0 = s\int f\,d\lambda_1 = 3sd_1.$$

Since $\nu_C(f) = 3|\nu_C|/j_C = 3|\lambda_s|/j_C = 3(sd_1 + 1 - s)/j$, it suffices to show that

$$sd_1 > \frac{sd_1 + (1-s)}{j},$$

which is equivalent to

$$s(jd_1 - d_1 + 1) > 1.$$

Now $sd_1 + (1-s) = |\nu_C| \le \frac{j}{j+\varepsilon p}$, where ε is from Lemma 4.3. Solving for s gives

$$s \ge \frac{1 - \frac{j}{j+\varepsilon p}}{1 - d_1},$$

whence it is enough to show that

$$(jd_1 - d_1 + 1)\frac{1 - \frac{j}{j+\varepsilon p}}{1 - d_1} > 1.$$

This boils down to

$$d_1(p\varepsilon + 1) > 1.$$

Now, $j > 1$ implies that $p \ge (7/6)\mathrm{Per}_1$, where Per_1 is the perimeter of a single tile. Since $\varepsilon/100 > \delta$ (by Lemma 4.3), this implies that $p\varepsilon + 1 > 1 + \varepsilon\mathrm{Per}_1 + \delta = 1/d_1$, proving the last inequality. □

LEMMA 5.5. *We have $\int f\,d\nu \le \int f\,d\lambda_s$ for all $\nu \in \widetilde{M}$ with $|\nu| = |\lambda_s|$. Equality holds only if*

- *$\nu(X_C) = 0$ for every component class C with $j_C > 1$, and*
- *whenever $\nu(X_C) > 0$ and $j_C = 1$, we have $|\nu_C| = d_1$.*

Proof. Recall that

$$\int f\,d\nu = \sum_C \nu(X_C) \int f\,d\nu_C.$$

For each component class C, let s_C be defined as follows:

- If there exists $s \in [0,1]$ such that $|\nu_C| = sd_1 + (1-s)$, then set $s_C := s$;
- otherwise, set $s_C := 1$.

Let $\omega_C := s_C\lambda_1 + (1-s_C)\lambda_0$ and

$$\sigma := \big(\nu(X') + \nu(X'')\big)\lambda_0 + \sum_C \nu(X_C)\omega_C.$$

From the previous lemma, if $|\nu_C| \geq d_1$, then $\int f\,d\nu_C \leq \int f\,d\omega_C$, with equality only if $j_C = 1$. If $|\nu_C| < d_1$, then $s_C = 1$ and

$$\int f\,d\nu_C = \frac{3\,|\nu_C|}{j_C} < 3d_1 = \int f\,d\omega_C.$$

Summing up, we obtain

$$\begin{aligned}\int f\,d\sigma &= \sum_C \nu(X_C)\int f\,d\omega_C \\ &\geq \sum_C \nu(X_C)\int f\,d\nu_C = \int f\,d\nu.\end{aligned}$$

Moreover, equality holds only if $\nu(X_C) = 0$ for every component C with $j_C > 1$ and $|\nu_C| = d_1$ whenever $j_C = 1$. Since $|\omega_C| \geq |\nu_C|$, we have

$$\begin{aligned}|\sigma| &= \nu(X') + \nu(X'') + \sum_C \nu(X_C)\,|\omega_C| \\ &\geq \nu(X')\,|\nu'| + \nu(X'') + \sum_C \nu(X_C)\,|\nu_C| \\ &= |\nu| \\ &= |\lambda_s|.\end{aligned}$$

Since σ and λ_s are both convex combinations of λ_0 and λ_1, this implies that $\int f\,d\sigma \leq \int f\,d\lambda_s$ with equality iff $\sigma = \lambda_s$. Thus, $\int f\,d\nu \leq \int f\,d\lambda_s$. In the equality case we must have $\int f\,d\nu = \int f\,d\sigma = \int f\,d\lambda_s$ and $\sigma = \lambda_s$. This implies that $\nu(X_C) = 0$ if $j_C > 1$ and $|\nu_C| = d_1$ if $j_C = 1$. □

LEMMA 5.6. *Let $\nu \in \widetilde{M}$. If $|\nu| = |\lambda_s|$, then $\int f\,d\nu \leq \int f\,d\lambda_s$. Equality holds iff $\nu = \lambda_s$.*

(Informally, λ_s uniquely maximizes the number of degrees of freedom per tile for invariant measures of a fixed density.)

Proof. The previous lemma implies $\int f\,d\nu \leq \int f\,d\lambda_s$. Assume $\int f\,d\nu = \int f\,d\lambda_s$; then

$$\nu = \nu(X')\nu' + \nu(X'')\lambda_0 + \nu(X_C)\nu_C,$$

where C is the component of size 1 and $|\nu_C| = d_1$. This gives $\int f\,d\nu = \nu(X_C)3d_1 = \int f\,d\lambda_s = 3sd_1$. Hence $\nu(X_C) = s$. Since ν' has density strictly less than $1 = |\lambda_0|$ but $|\nu| = |\lambda_s|$, we must have $\nu(X') = 0$. That is,

$$\nu = \nu(X'')\lambda_0 + \nu(X_C)\nu_C.$$

Since ν and λ_0 are isometry invariant, ν_C must also be isometry invariant. By Lemma 4.2, λ_1 is the unique isometry-invariant measure with support in X_C and with density d_1. Hence $\nu_C = \lambda_1$. This implies $\nu = \lambda_s$, and the proof is finished. □

Proof of Theorem 3.1. It is easy to see that one can pack the $N \times N$ torus in such a way that there is a large region of tightly linked tiles and a large region of densely loosely linked tiles, and in such a way that the interface between the two regions has a density which approaches zero as N tends to infinity, and the density d_N of the packing P_N tends to d. Let ω_N be the invariant measure supported on isometric copies of $\tilde{P}_N$ (a pull-back of P_N to the plane). Then ω_N tends to λ_s in the weak* topology. By Lemma 5.3, this implies that $\int f\, d\omega_N \to \int f\, d\lambda_s$.

Now μ_N, the uniform measure of density d_N on the $N \times N$ torus, satisfies $\int f\, d\mu_N \geq \int f\, d\omega_N$. This is because μ_N is by definition supported on packings with the maximal number of degrees of freedom for the given density d_N. Hence $\liminf_N \int f\, d\mu_N \geq \liminf_N \int f\, d\omega_N = \int f\, d\lambda_s$.

Therefore, if μ_∞ is any weak* subsequential limit of $\langle \mu_N \rangle_N$, then $\int f\, d\mu_\infty \geq \int f\, d\lambda_s$. But $d_N \to d$, so $|\mu_\infty| = |\lambda_s|$ by Lemma 5.1. The previous lemma now implies that $\mu_\infty = \lambda_s$. □

Returning to the discussion of the introduction, we note that from simulations of hard disks, one would expect the corollary to hold even for a range of densities below d_1, but we do not know how to prove this.

Remark on higher dimensions. The basic features of our argument can be generalized to dimension 3 or higher, except for our use in Lemma 4.2 of [6] on the minimal Voronoi region in disk packings in the plane. It would be of interest if this part of our proof could be replaced with an argument insensitive to dimension, since the Voronoi regions of, say, the spheres in a face-centered cubic lattice do not minimize volume per site [8].

6. Low density. In this final section, we confirm the intuition that at low densities there will be no infinite loosely linked component. It is obvious that there is no infinite *tightly* linked component at densities smaller than d_1.

We begin with a lemma that holds for any tile shape (in fact, for any *collection* of shapes and sizes, as long as each can be fit into a disk of some fixed radius s, and "density" is interpreted as the number of tiles per unit area).

LEMMA 6.1. *For small enough density d, if a packing P is drawn from $\mu^{(d)}$, then the probability that the disk B_R of radius R about the origin contains more than $9R^2d$ tile centers goes to zero as $R \to \infty$.*

Proof. Let s be the radius of the smallest disk containing the tile (in our case, s is about $2^{1/2} \cdot 3^{-3/4} \cdot (1+2\rho)$) and choose

$$0 < d < \frac{.05}{13\pi s^2} \ ;$$

for our zipper tiles with small enough ρ, $d \leq .003$ suffices. Let T be the set of tiles whose centers fall in B_R, $k := \lceil \pi R^2 d \rceil$, and $\ell > 9R^2 d$. Letting $\mu^{(d)}(\cdot)$ denote the probability of an event with respect to the measure $\mu^{(d)}$, we shall show that

$$\frac{\mu^{(d)}(|T| = \ell)}{\mu^{(d)}(|T| = k)} \leq \gamma^{\ell - k}$$

for some constant $\gamma < 1$. It then follows that

$$\mu^{(d)}(|T| > 9R^2 d) \leq \mu^{(d)}(|T| = k) \sum_{\ell = \lceil 9R^2 d \rceil}^{\infty} \gamma^{\ell - k} \leq \mu^{(d)}(|T| = k) \frac{\gamma^{\lfloor (9-\pi) R^2 d \rfloor}}{1 - \gamma} \to 0$$

as $R \to \infty$, as desired.

The measure $\mu^{(d)}$ is the limit of uniform distributions of configurations on the $N \times N$ torus $\mathbf{T}_N$, in turn obtainable by choosing a sequence of $n = \lfloor N^2 d \rfloor$ points from the Lebesgue distribution λ on $\mathbf{T}_N^n$ as the centers of the tiles, orienting each tile independently and uniformly at random, and finally conditioning on no overlap. We denote by $\lambda(|T|=j)$ the a priori probability that exactly j points fall inside B_R (which we take to be some fixed disk in the torus).

Let Φ be the event that there is no overlap among the tiles whose centers lie in B_R, and Ψ be the event that there is no overlap involving any tile whose center falls *outside* B_R. Then

$$\frac{\mu^{(d)}(|T|=\ell)}{\mu^{(d)}(|T|=k)} = \frac{\lambda(|T|=\ell)}{\lambda(|T|=k)} \cdot \frac{\lambda(\Phi \mid |T|=\ell)}{\lambda(\Phi \mid |T|=k)} \cdot \frac{\lambda(\Psi \mid |T|=\ell \wedge \Phi)}{\lambda(\Psi \mid |T|=k \wedge \Phi)},$$

and our job is to bound the three fractions on the right.

For the first, we note that $|T|$ is binomially distributed in the measure λ; hence

$$\frac{\lambda(|T|=\ell)}{\lambda(|T|=k)} = \frac{\binom{n}{\ell}\left(\frac{\pi R^2}{N^2}\right)^{\ell}\left(1-\frac{\pi R^2}{N^2}\right)^{n-\ell}}{\binom{n}{k}\left(\frac{\pi R^2}{N^2}\right)^{k}\left(1-\frac{\pi R^2}{N^2}\right)^{n-k}} \le \frac{(n-k)!/(n-\ell)!}{\ell!/k!}\left(\frac{\frac{k}{n}}{1-\frac{k}{n}}\right)^{\ell-k}$$
$$< \frac{(n-k)^{\ell-k}}{(\ell/e)^{\ell-k}} \cdot \frac{k^{\ell-k}}{(n-k)^{\ell-k}} = \left(\frac{ek}{\ell}\right)^{\ell-k} \le 0.95^{\ell-k}$$

for large R.

The next fraction is easy: Since we may throw the first k centers into B_R, then for the remaining $\ell-k$, we have that

$$\frac{\lambda(\Phi \mid |T|=\ell)}{\lambda(\Phi \mid |T|=k)}$$

is the probability that the additional $\ell-k$ centers do not cause a collision, which is at most 1.

For the (inverse of) the third fraction, we throw $n-\ell$ centers into the region outside B_R, then throw the remaining $\ell-k$. A new point, if it lands at a distance greater than $2s$ from any previous point or from the disk B_R, causes no new overlap, and at each stage there are fewer than $n-k$ points already placed. Hence

$$\frac{\lambda(\Psi \mid |T|=k \wedge \Phi)}{\lambda(\Psi \mid |T|=\ell \wedge \Phi)} > \left(\frac{N^2 - \pi(R+2s)^2 - (n-k)4\pi s^2}{N^2 - \pi R^2}\right)^{\ell-k}$$

$$\ge \left(1 - 4\pi s^2 d - \frac{4\pi sR + 4\pi s^2}{N^2 - \pi R^2}\right)^{\ell-k} > \left(1 - 13 s^2 d\right)^{\ell-k}$$

for $N \gg R$.

Putting the inequalities together, we have

$$\frac{\mu^{(d)}(|T|=\ell)}{\mu^{(d)}(|T|=k)} \le \left(\frac{.95}{1-13s^2 d}\right)^{\ell-k} = \gamma^{\ell-k},$$

where $\gamma := .95/(1-13s^2 d) < 1$ by choice of d. □

PROPOSITION 6.2. *For some $d_2 > 0$, the $\mu^{(d)}$-probability that the origin is inside an infinite connected component of loosely linked tiles is zero for $d < d_2$.*

Proof. Let $d \in (0, .003)$ be a density to be chosen later. Let P be a packing drawn from $\mu^{(d)}$; we aim to show that the probability that the origin is connected by a loosely linked chain of tiles of P to some point at distance R approaches zero as $R \to \infty$.

We again choose some large radius R and let T be the set of tiles of P whose centers fall inside the disk B_R.

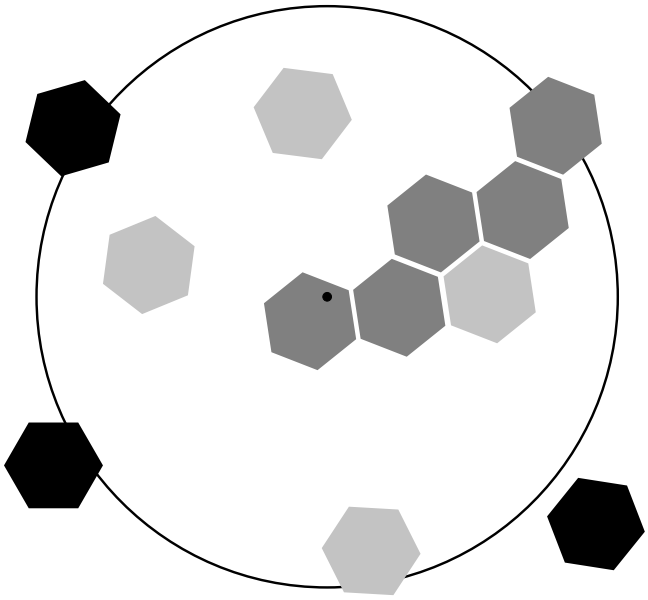


FIG. 6. *An unlikely configuration of tiles in and around B_R.*

Fix the positions of the tiles of $P \setminus T$ (the black tiles of Figure 6) and consider the space of packings having these tiles plus n tiles whose centers fall in B_R. We think of this space as being a subset of $T_1(B_R)^n/\Sigma_n$, where $T_1(B_R)$ is the unit tangent bundle of B_R (modulo a $2\pi/6$ rotation to take into account the symmetries of the tile) and the symmetric group acts by permuting the factors.

If α_n is the volume (in $T_1(B_R)^n/\Sigma_n$-space) of this space and $m < n$, then by packing $n-m$ tiles into B_R and then the remaining m in the leftover space, we have

$$\alpha_n \geq \frac{1}{\binom{n}{n-m}} \alpha_{n-m} \frac{1}{m!} \left[\pi(R-2s)^2 - n\pi(2s)^2\right]^m,$$

where s is, as before, the radius of the circle circumscribing a tile. This takes into account possible intrusion of tiles in $P - T$ into B_R, and the fact that a tile center at point x can exclude nearby centers, but only within distance $2s$ of x.

Let β denote the "wiggle room" of a tile t loosely linked to a stationary tile t', that is, the three-dimensional volume of the space of positions of t; then $\beta = \mathrm{O}(\rho^3)$ (but we use only that β is bounded by a constant). If a packing "percolates," that is, contains a chain of loosely linked tiles connecting the center to the boundary of

B_R, let $t_1, \ldots, t_m$ be a shortest such chain (the dark grey tiles of Figure 6). Note that $m \geq R/(2s)$. For each $i > 2$, the tile t_i is linked to one of the three sides of t_{i-1} farthest from the side of t_{i-1} linked to t_{i-2}, and has wiggle room at most β with respect to t_{i-1}. Accounting for the orientation of t_1 and allowing the remaining $n-m$ tile centers to fall anywhere in B_R, we have that the $3n$-dimensional volume of the set of percolating packings is bounded by $(\pi/3) \cdot 6 \cdot 3^{m-2} \cdot \beta^{m-1} \cdot \alpha_{n-m} < 3^m \beta^{m-1} \alpha_{n-m}$.

Comparing with the lower bound for α_n, we find that given $|T| = n \leq 9dR^2$, the probability of percolation is less than

$$\frac{3^m \beta^{m-1} \alpha_{n-m}}{\alpha_n} \leq \frac{3^m \beta^{m-1}\, n!/(n-m)!}{\left[\pi(R-2s)^2 - n\pi(2s)^2\right]^m} < \left(\frac{27\beta d R^2}{\pi\left[(R-2s)^2 - 36 d R^2 s^2\right]} \right)^m \Big/ \beta,$$

which goes to zero as R (thus also m) increases, for suitably chosen d. Since we know from Lemma 6.1 that $\mu^{(d)}(|T| \leq 9dR^2)$ approaches 1 as $R \to \infty$, the proposition follows. □

A more careful argument would prove Proposition 6.2 for any density below $1/\big(4\pi(2/3\sqrt{3})\big) = .2067^+$ for sufficiently small ρ, but clearly the probability of percolation will remain 0 for much higher densities than that.

7. A conjecture. We have shown that high-density random packings of zipper tiles in the plane contain an infinite loosely linked component with positive probability, while low-density random packings do not. What happens in the case of ordinary disks, where there is no apparent linking mechanism? We believe, but cannot prove, the following.

CONJECTURE. *Suppose $\mu^{(d)}$ is defined as above, but for geometric disks of radius 1. Join two centers by an edge if their distance is at most $2 + \varepsilon$ for some fixed $\varepsilon \ll 1$. Then for sufficiently high-density d below the maximum, the graph resulting from a configuration drawn from $\mu^{(d)}$ will contain an infinite connected component a.s.*

This connectedness property can in fact be proven by a standard Peierls-type argument for large ε. This may be known already, though we do not know a reference; it is a straightforward extension of the traditional percolation proof to a situation with a new length scale given by the size of the disks. In general, there is some parameter set of $(\varepsilon, d) \subset (0, \infty) \times (0, 1)$ for which there is an infinite component. For small d or for large ε, the problem is quite similar to continuum percolation, where one connects by an edge two points of a Poisson point process if their distance is at most r. Because of homotheties, one may fix the intensity of the point process to be 1. Then there is a phase transition in r. Our situation is quite different in that we really have two parameters, due to the size of the disks, but our conjecture is that there is a phase transition in d for every ε nevertheless.

There is a fundamental difference between the connectedness property for small ε and for large ε. An infinite set of disks connected or linked in the sense of small ε would resist shearing in a sense not true for a set linked only in the sense of large ε. We prove the connectedness property for the zipper model for small ε, using in an essential way special features of the nonconvex zipper tiles.

Acknowledgment. We gratefully acknowledge learning of [11] from an anonymous referee.

REFERENCES

[1] B. J. Alder and W. G. Hoover, *Numerical Statistical Mechanics*, in Physics of Simple Liquids, H. N. V. Temperley, J. S. Rowlinson, and G. S. Rushbrooke, eds., John Wiley, New York, 1968, pp. 79–113.

[2] L. Bowen, R. Lyons, C. Radin, and P. Winkler, *Fluid/solid transition in a hard-core system*, Phys. Rev. Lett., 96 (2006), paper 025701.

[3] L. Bowen, C. Holton, C. Radin, and L. Sadun, *Uniqueness and symmetry in problems of optimally dense packings*, Math. Phys. Electron. J., 11 (2005), paper 1.

[4] A. Donev, S. Torquato, F. H. Stillinger, and R. Connelly, *Jamming in hard sphere and disk packings*, J. Appl. Phys., 95 (2004), pp. 989–999.

[5] L. Fejes Tóth, *Regular Figures*, Macmillan, New York, 1964.

[6] L. Fejes Tóth, *Lagerungen in der Ebene auf der Kugel und im Raum*, 2nd ed., Springer-Verlag, Berlin, 1972, pp. 62–63.

[7] G. Grimmett, *Percolation*, 2nd ed., Springer-Verlag, Berlin, 1999.

[8] T. C. Hales and S. McLaughlin, *A Proof of the Dodecahedral Conjecture*, http://www.arxiv.org/abs/math?papernum=9811079 (2002).

[9] H. Koch, L. Sadun, and C. Radin, *Most stable structure for hard spheres*, Phys. Rev. E, 72 (2005), paper 016708.

[10] C. Radin, *Orbits of orbs: Sphere packing meets Penrose tilings*, Amer. Math. Monthly, 111 (2004), pp. 137–149.

[11] T. Richthammer, *Translation-Invariance of Two-Dimensional Gibbsian Systems of Particles with Internal Degrees of Freedom*, preprint, 2006. Available online from http://www.arxiv.org/abs/math.PR/0603140.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1090–1125

TIME PERIODIC SOLUTIONS FOR THE NONLINEAR WAVE EQUATION WITH LONG MINIMAL PERIOD*

LUCA BIASCO† AND LAURA DI GREGORIO†

Abstract. We prove existence and multiplicity of small amplitude periodic solutions for the wave equation with small "mass" and odd nonlinearity. Such solutions bifurcate from resonant finite dimensional invariant tori of the fourth order Birkhoff normal form of the associated Hamiltonian system. The number of geometrically distinct solutions and their minimal periods go to infinity when the "mass" goes to zero. This is the first result about long minimal period for the autonomous wave equation.

Key words. nonlinear wave equation, infinite dimensional Hamiltonian systems, periodic solutions, Birkhoff normal form

AMS subject classifications. 34C25, 35L05, 37K50

DOI. 10.1137/050638606

1. Introduction and main results.

Let us consider the nonlinear wave equation on the interval $[0,\pi]$ with Dirichlet boundary conditions

$$\begin{cases} u_{tt} - u_{xx} + \mu u + f(u) = 0, \\ u(t,0) = u(t,\pi) = 0, \end{cases} \tag{1}$$

where $\mu > 0$ and $f(0) = f'(0) = 0$.

Equation (1) can be studied as an infinite dimensional Hamiltonian system. Denoting $v = u_t$, the Hamiltonian is

$$H(v,u) = \int_0^\pi \left(\frac{v^2}{2} + \frac{u_x^2}{2} + \mu \frac{u^2}{2} + g(u) \right) dx,$$

where $g = \int_0^u f(s)ds$. The Hamiltonian equations are

$$u_t = \frac{\partial H}{\partial v} = v, \qquad v_t = -\frac{\partial H}{\partial u} = u_{xx} - \mu u - f(u).$$

Introducing coordinates $q = (q_1, q_2, \ldots)$, $p = (p_1, p_2, \ldots)$ through the relations

$$v(x) = \sum_{i\geq 1} \sqrt{\omega_i} p_i \chi_i(x), \qquad u(x) = \sum_{i\geq 1} \frac{q_i}{\sqrt{\omega_i}} \chi_i(x), \tag{2}$$

where $\chi_i(x) := \sqrt{2/\pi} \sin ix$ and $\omega_i := \sqrt{i^2 + \mu}$, the Hamiltonian takes the form

$$H = \frac{1}{2} \sum_{i\geq 1} \omega_i (q_i^2 + p_i^2) + \text{higher order terms}. \tag{3}$$

*Received by the editors August 22, 2005; accepted for publication (in revised form) May 25, 2006; published electronically December 1, 2006. This research was supported by M.U.R.S.T. Variational Methods and Nonlinear Differential Equations.

http://www.siam.org/journals/sima/38-4/63860.html

†Dipartimento di Matematica, Università "Roma Tre," Largo S. L. Murialdo 1, 00146 Roma, Italy (biasco@mat.uniroma3.it, digregorio@mat.uniroma3.it).

Since we are interested in small amplitude periodic solutions, the higher order terms in (3) may be neglected in first approximation, and we can consider only the Hamiltonian $\Lambda := \sum_{i\geq 1} \omega_i (q_i^2 + p_i^2)/2$. The origin is an elliptic equilibrium point for Λ, and the Λ-orbits are the superpositions of the harmonic oscillations $q_i(t) = A_i \cos(\omega_i t + \varphi_i)$ of the basic modes χ_i, where $A_i \geq 0$, $\varphi_i \in \mathbb{R}$, and ω_i are, respectively, the amplitude, the phase, and the frequency of the ith harmonic oscillator $\omega_i(q_i^2 + p_i^2)/2$, $i \geq 1$.

Analogously, by neglecting the term $f(u)$ in (1), we see that every solution of the linear equation $u_{tt} - u_{xx} + \mu u = 0$ is of the form

$$u(t,x) = \sum_{i\geq 1} a_i \cos(\omega_i t + \varphi_i) \sin ix, \tag{4}$$

with $a_i = A_i \sqrt{2/\pi\omega_i} \geq 0$. These solutions, in general, are periodic if only one basic mode is excited, namely, if $a_i = 0$ for any $i \neq i_0$, for a suitable $i_0 \geq 1$, while $a_{i_0} > 0$. If two basic modes are excited, the situation changes: Except a countable set of $\mu > 0$, for any $\mathcal{I} := \{i_1, \dots, i_N\} \subset \mathbb{N}^+$, $N \geq 2$, the vector $\omega := (\omega_{i_1}, \dots, \omega_{i_N})$ is rationally independent,[1] and, therefore, any solutions of the form $\sum_{i\in\mathcal{I}} a_i \cos(\omega_i t + \varphi_i) \sin ix$ is quasi periodic. Consequently, if at least two amplitudes in (4) are different from zero, the solution $u(t,x)$ cannot be periodic. One can conclude that, except a countable set of $\mu > 0$, the only periodic solutions of $u_{tt} - u_{xx} + \mu u = 0$ are of the form $u(t,x) = a_{i_0} \cos(\omega_{i_0} t + \varphi_{i_0}) \sin i_0 x$, for $i_0 \geq 1$.

In light of these considerations, a natural way to find periodic solutions of (1) (see, for example, [K87], [W90], [K93], [CW93], [Bou99], [B00]) is to extend the Lyapunov center theorem for finite dimensional Hamiltonian systems in a neighborhood of an elliptic equilibrium. Namely, for any fixed $i_0 \geq 1$, one constructs a family of small amplitude periodic orbits of the Hamiltonian H bifurcating from the i_0th basic mode. This can be done since, for μ far away from zero, the linear frequency ω_{i_0} is not resonant with the other ones. The frequencies $\tilde{\omega}$ of the solutions will be close to the linear frequency $\omega := \omega_{i_0}$, and the corresponding periods $2\pi/\tilde{\omega}$ will be close to the linear period $2\pi/\omega$.

Here we look for solutions having *large* minimal period. Such solutions are interesting as examples of the complexity of the dynamics and because they come up only as nonlinear phenomena.

A classical way to find long-period orbits, close to an elliptic equilibrium point[2] in finite dimensional systems, was carried out by Birkhoff and Lewis in [BL34] (see also [L34], [Mo77]). Their procedure consists of putting the system into fourth order Birkhoff normal form: the truncated Hamiltonian obtained by neglecting the five or higher order terms is integrable. If the so-called "twist" condition on the action-to-frequency map holds, there exist infinitely many resonant tori on which the motion of the truncated Hamiltonian is periodic. By the implicit function theorem and topological arguments, Birkhoff and Lewis showed the existence of a sequence of resonant tori accumulating at the origin with the property that at least two periodic orbits bifurcate from each of them.

In our paper we apply the Birkhoff–Lewis procedure to the nonlinear wave equation seen as an infinite dimensional Hamiltonian system. Such an extension to the Hamiltonian PDEs was recently carried out in [BaB] for the beam equation and the nonlinear Schrödinger (NLS) equation (see Remark 3.9 below for comparison).

[1] See Lemma A.1 in the appendix for a proof.

[2] Actually [BL34] considers a neighborhood of an elliptic, nonconstant, periodic orbit, but the scheme is essentially the same for elliptic equilibria.

We point out that, in the infinite dimensional case, one meets two difficulties that do not appear in the finite dimensional case: the generalization of the Birkhoff normal form, and a small divisors problem. Concerning the first difficulty, we consider only a "seminormal form" following [P96]. We suppose that f is a real analytic odd function $f = \sum_{m\geq 3} f_m u^m$ with $f_3 \neq 0$ and fix a *finite* subset $\mathcal{I} \subset \mathbb{N}$. Then we put the Hamiltonian $\bar{H}$ in (3) into the form

$$H = \Lambda + \bar{G} + \hat{G} + K,$$

where $\Lambda = \sum_{i\geq 1} \omega_i (p_i^2 + q_i^2)/2$, $\bar{G} + \hat{G}$ is the fourth order term, with $\bar{G}$ depending only on the "actions" $\mathcal{J}_i := (p_i^2 + q_i^2)/2$, $i \in \mathbb{N}^+$, and $\hat{G}$ depending only on p_i, q_i, $i \notin \mathcal{I}$, and K is the sixth order term. However, to put the Hamiltonian into normal form, the linear frequencies ω_i must satisfy a suitable nonresonance condition (see Lemma 2.6), which deteriorates for μ going to zero.

The truncated Hamiltonian $\Lambda + \bar{G} + \hat{G}$ possesses the $2N$-dimensional invariant manifold $\{p_i = q_i = 0,\ i \notin \mathcal{I}\}$, which is foliated by N-dimensional invariant tori. Due to the "twist" property of $\bar{G}$, which follows from $f_3 \neq 0$, the linear frequencies of such tori are an open set of $\mathbb{R}^N$. We focus on completely resonant frequencies $\tilde{\omega} := (\tilde{\omega}_{i_1}, \dots, \tilde{\omega}_{i_N})$, namely, $\tilde{\omega}$'s such that there exist T's with $\tilde{\omega} T/2\pi \in \mathbb{Z}^N$. Then the $(\Lambda + \bar{G} + \hat{G})$-flow on the associated N-dimensional tori is periodic with periods T. Such lower dimensional tori are highly degenerate. Hence, in order to show the persistence of periodic orbits for the whole Hamiltonian H, we have to impose some nondegeneracy conditions to avoid resonances between the torus frequencies and the frequencies of the normal oscillations. This is the point in which the small divisors problem appears.

The estimate on the small divisors is the crucial step. To overcome this problem, we have decided to impose a strong condition on the small divisors, avoiding KAM analysis (see Remark 3.12). For such a (technical) reason we can consider only periods T which are multiples of 2π (as in the classical variational approach of Rabinowitz, Brezis, Nirenberg, etc.), and we also need

$$\left| \mu \frac{T}{2\pi} - 2m \right| \geq \frac{1}{2} \qquad \forall\, m \in \mathbb{N}$$

and

$$\mu^2 T \ll 1.$$

We will denote by $\mathcal{T}$ the set of periods verifying the above properties.

Therefore we consider the "mass" $\mu > 0$ as a *small* parameter and we make our analysis perturbative with respect to it. On the other hand, for $\mu \to 0$ the frequencies ω_i tend to the *completely resonant* frequencies i, and the above described normal form *degenerates* in the sense that its domain of definition shrinks to zero while the remainder term K blows up.

We note that the set $\mathcal{T}$ of admissible periods is *finite*, but its cardinality goes to infinity, when μ goes to zero. μ^2-close to $\omega = (\omega_{i_1}, \dots, \omega_{i_N})$, we construct a set of completely resonant frequencies $\tilde{\omega} = \tilde{\omega}(T) \in \mathbb{R}^N$, parametrized by the periods $T \in \mathcal{T}$, with $\tilde{\omega} T/2\pi \in \mathbb{Z}^N$. At the same time, we construct a set of actions $\tilde{\mathcal{J}}_i = \tilde{\mathcal{J}}_i(T) \approx \mu^2$, $i \in \mathcal{I}$, parametrized by $T \in \mathcal{T}$, such that $\tilde{\omega}_i$ is the image of $\tilde{\mathcal{J}}_i$ through the action-to-frequency map $\partial_{\mathcal{J}_i} H$ on $\{p_i = q_i = 0,\ i \notin \mathcal{I}\}$. We will prove the existence

of T-periodic solutions of H, μ^2-close to the T-periodic solutions of the truncated Hamiltonian $\Lambda + \bar{G} + \hat{G}$ defined as

$$\begin{cases} p_i(t) = \sqrt{2\tilde{\mathcal{J}}_i}\sin(\tilde{\omega}_i t + \varphi_i), \quad q_i(t) = \sqrt{2\tilde{\mathcal{J}}_i}\cos(\tilde{\omega}_i t + \varphi_i) & \text{for } i \in \mathcal{I}, \\ p_i(t) = q_i(t) = 0 & \text{for } i \in \mathcal{I}^c, \end{cases} \tag{5}$$

where the angles φ_i, $i \in \mathcal{I}$, have to be determined.

We perform a Lyapunov–Schmidt reduction as in [BBiV], [BaB], splitting the problem into two equations: the kernel (or bifurcation) equation on the N-dimensional torus $\{\varphi_{i_1}, \dots, \varphi_{i_N}\}$ and the range equation on its orthogonal space. We first solve the infinite dimensional range equation by the contraction mapping theorem, using the above estimate on the small divisors and controlling the blow-up of the remainder term K for μ going to zero. Critical points of the action functional restricted to the solutions of the range equation satisfy the kernel equation. Since the restricted action functional is defined on an N-dimensional torus, existence of critical points and, therefore, of solutions of H, follows.

Before stating our main result, we return to the PDE formulation. Recalling (2) and (5), for every $T \in \mathcal{T}$, we find T-periodic solutions of (1), μ^2-close to the T-periodic "pseudosolution"

$$\tilde{u}(t,x) := \mu \sum_{i \in \mathcal{I}} a_i \cos(\tilde{\omega}_i t + \varphi_i) \sin ix,$$

where $\mu a_i := \sqrt{\tilde{\mathcal{J}}_i/\pi\omega_i}$, $a_i \approx 1$. We note that the minimal period of $\tilde{u}$ is $\tilde{T}^{\min} = T/\gcd(k_{i_1}, \dots, k_{i_N})$, where $k_i := \tilde{\omega}_i T/2\pi \in \mathbb{N}$, for $i \in \mathcal{I}$. Since $\tilde{\omega}$ is μ^2-close to the rationally independent[3] frequency ω, $\tilde{T}^{\min}$ will be large for μ small. Because $u - \tilde{u} \approx \mu^2$, we obtain an analogous estimate on the minimal period $T^{\min}$ of u. Not all the solutions of (1), corresponding to different T's belonging to $\mathcal{T}$, have to be geometrically distinct. However, by the very precise estimate $u - \tilde{u} \approx \mu^2$, we prove that the total number of geometrically distinct solutions found here is also large for μ small.

We now state our main result.

THEOREM 1.1. *Let f be a real analytic, odd function of the form $f(u) = \sum_{m \geq 3} f_m u^m$, $f_3 \neq 0$. Let $N \geq 2$ and $\mathcal{I} := \{i_1, \dots, i_N\} \subset \mathbb{N}^+$. Then there exists a constant $0 < c = c(\mathcal{I}) < 1$ such that, if $0 < \mu \leq c$, there exist at least c/μ geometrically distinct smooth periodic solutions $u(t,x)$ of (1), verifying*

$$\sup_{t \in \mathbb{R},\ x \in [0,\pi]} \left| u(t,x) - \mu \sum_{i \in \mathcal{I}} a_i \cos(\tilde{\omega}_i t + \varphi_i) \sin ix \right| \leq c^{-1}\mu^2, \tag{6}$$

for suitable $a_i \geq c$, $\varphi_i \in \mathbb{R}$, and $\tilde{\omega}_i \in \mathbb{R}$, verifying

$$|\tilde{\omega}_i - \omega_i| \leq c^{-1}\mu^2. \tag{7}$$

The minimal period $T^{\min}$ of any solution belongs to $\pi\mathbb{Q}$ and satisfies

$$\frac{c}{\mu} \leq T^{\min} \leq \frac{c}{\mu^2}. \tag{8}$$

Remark 1.1. The solutions $u(t,x)$ found in Theorem 1.1 are infinitely differentiable. Actually, they are analytic in the spatial variable. Estimate (6) can be improved and, in particular, one can obtain analogous estimates on the derivative of u

[3]Except a countable set of μ's.

of any order $k \geq 1$. However, in this case, the constant c will depend on k. See also Remark 3.19 for further details.

We stress that this is the first result about periodic solutions of *large minimal period* for the autonomous nonlinear wave equation (for a different type of results on large minimal period, in the forced case, see [T]).

We also mention the following substantial difference between the periodic solutions found by analogy of the Lyapunov center theorem and the theorems we construct. The "Lyapunov-type" orbits are obtained as the continuation of one linear mode to the nonlinear system; they involve only one of the linear harmonic oscillators, and the amplitudes on the other modes are much smaller (except in the resonant case $\mu = 0$, discussed in [BP01], [BBo03], [BBo04], [BBo06]). On the other hand, the periodic solutions constructed here involve $N \geq 2$ harmonic oscillators, oscillating with the same order of magnitude, and are a *truly nonlinear phenomenon*, as they do not have any analogue in the linear case, where all periodic orbits are the oscillation of only one basic mode and do not have long minimal period (see also [BBiV], [BaB]).

The results proved here have been announced in [BDG].

2. Hamiltonian setting and Birkhoff normal form. We study (1) as an infinite dimensional Hamiltonian system with coordinates u and $v = u_t$. Denoting $g = \int_0^u f(s)ds$, the Hamiltonian is

$$H(v,u) = \int_0^\pi \left(\frac{v^2}{2} + \frac{u_x^2}{2} + \mu \frac{u^2}{2} + g(u) \right) dx.$$

The equations of motion are

$$u_t = \frac{\partial H}{\partial v} = v, \qquad v_t = -\frac{\partial H}{\partial u} = u_{xx} - \mu u - f(u).$$

Let us rewrite the Hamiltonian in infinite coordinates $(p,q) \in \ell^{a,s} \times \ell^{a,s}$, where

$$\ell^{a,s} = \ell^{a,s}(\mathbb{R}) := \left\{ q = (q_1, \ldots),\ q_i \in \mathbb{R},\ i \geq 1 \ \text{ s.t. } \|q\|_{a,s}^2 = \sum_{i \geq 1} |q_i|^2 i^{2s} e^{2ai} < \infty \right\},$$

by the transformation

$$v = \mathcal{S}' p = \sum_{i \geq 1} \sqrt{\omega_i} p_i \chi_i, \quad u = \mathcal{S} q = \sum_{i \geq 1} \frac{q_i}{\sqrt{\omega_i}} \chi_i, \tag{9}$$

with $\omega_i = \sqrt{i^2 + \mu}$ and $\chi_i = \sqrt{2/\pi} \sin ix$. We get

$$H = \Lambda + G = \frac{1}{2} \sum_{i \geq 1} \omega_i (q_i^2 + p_i^2) + \int_0^\pi g(\mathcal{S} q)\, dx, \tag{10}$$

where we denote with Λ the integrable part and with G the nonintegrable part. The equations of motion are

$$\dot{p_i} = -\frac{\partial H}{\partial q_i} = -\omega_i q_i - \frac{\partial G}{\partial q_i}, \qquad \dot{q_i} = \frac{\partial H}{\partial p_i} = \omega_i p_i \tag{11}$$

with respect to the symplectic structure $\sum dp_i \wedge dq_i$ on $\ell^{a,s} \times \ell^{a,s}$. Here, " $\dot{}$ " denotes the time derivative.

For $k \in \mathbb{N}$, we consider the space $C^k(\mathbb{R}, \ell^{a,s})$ of all the functions $\mathbb{R} \ni t \mapsto q(t) \in \ell^{a,s}$ with finite norm

$$\|q\|_{C^k(\mathbb{R},\ell^{a,s})} := \sum_{h=0}^{k} \sup_{t\in\mathbb{R}} \|\partial_t^h q(t)\|_{a,s}. \tag{12}$$

Similarly, for a C^k function $\mathbb{R} \ni t \mapsto \big(p(t), q(t)\big) \in \ell^{a,s} \times \ell^{a,s}$ we consider the norm

$$\|(p,q)\|_{C^k(\mathbb{R},\ell^{a,s}\times\ell^{a,s})} := \|p\|_{C^k(\mathbb{R},\ell^{a,s})} + \|q\|_{C^k(\mathbb{R},\ell^{a,s})}.$$

LEMMA 2.1. *Let us assume that $a > 0$ and that s is arbitrary. Let be $\mathbb{R} \ni t \mapsto \big(p(t), q(t)\big) \in \ell^{a,s} \times \ell^{a,s}$ a solution of (11) of class C^k, $2 \le k \le \infty$; then*

$$u(t,x) := \sum_{i\ge 1} \frac{q_i(t)}{\sqrt{\omega_i}}\, \chi_i(x) \tag{13}$$

is a classical solution of (1) of class C^k.

Proof. See the appendix. □

Let us note that χ_i, for $i \ge 1$, is a complete orthonormal basis for the L^2 functions on $[0,\pi]$, but not for all analytic functions on $[0,\pi]$. Indeed, only the analytic functions on $[0,\pi]$, whose odd extension on $[-\pi,\pi]$ is still analytic, can be expanded in sine-series converging in the analytic norm $\|\cdot\|_{a,s}$.

We now discuss the regularity of the Hamiltonian vectorfield generated by G in (10). Let us consider ℓ^2_b and L^2, which are, respectively, the Hilbert spaces of all *bi*-infinite, square-integrable sequences with complex coefficients and all square-integrable complex valued functions on $[-\pi,\pi]$. To identify the two spaces we can consider the inverse discrete Fourier transform,

$$\mathcal{F} :\, \ell^2_b \to L^2, \qquad q \mapsto \big[\mathcal{F}q\big](x) := \frac{1}{\sqrt{2\pi}} \sum_{i\in\mathbb{Z}} q_i e^{\mathrm{i}ix},$$

which defines an isometry between the two spaces. Let $a \ge 0$ and $s \ge 0$. The subspaces $\ell^{a,s}_b \subset \ell^2_b$ contain all bi-infinite sequences, whose norm is defined by

$$\|q\|^2_{a,s} := |q_0|^2 + \sum_{i\in\mathbb{Z}} |q_i|^2 |i|^{2s} e^{2a|i|}.$$

In this way we obtain, through the Fourier transform $\mathcal{F}$, the subspaces $W^{a,s} \subset L^2$ endowed with the norm

$$\|\mathcal{F}q\|_{a,s} = \|q\|_{a,s}.$$

For $a > 0$, the subspaces $W^{a,s}$ consist of all 2π-periodic functions which are analytic and bounded in the complex strip $|\mathrm{Im} z| < a$ with trace functions on $|\mathrm{Im} z| = a$ belonging to the standard Sobolev space H^s. In this way, we obtain an orthonormal basis for all analytic functions on $[0,\pi]$.

The following two results were proved in [P96].

LEMMA 2.2. *For $a \ge 0$ and $s > 1/2$, the space $\ell^{a,s}_b$ is a Hilbert algebra with respect to convolution of the sequences*

$$\|q * p\,\|_{a,s} \le \mathrm{const}\, \|q\|_{a,s}\, \|p\,\|_{a,s}$$

with a constant depending only on s.

LEMMA 2.3. *For $a \geq 0$ and $s > 0$, the gradient $G_q := (\frac{\partial G}{\partial q_1}, \frac{\partial G}{\partial q_2}, \ldots)$ is a real analytic map from a neighborhood of the origin in $\ell^{a,s}$ into $\ell^{a,s+1}$. Moreover,*

$$\|G_q\|_{a,s+1} = O(\|q^3\|_{a,s}).$$

In light of these considerations, we have the real analytic Hamiltonian H in (10), defined in some neighborhood of the origin in the Hilbert space $\ell^{a,s} \times \ell^{a,s}$ with standard symplectic structure $\sum_{i\geq 1} dp_i \wedge dq_i$. The parameters a and s may be fixed arbitrarily; in particular we take $a > 0$ and $s > 1$. The term G is independent of p, so the associated Hamiltonian vectorfield,

$$X_G := \sum_{i\geq 1} \left(\frac{\partial G}{\partial p_i} \frac{\partial}{\partial q_i} - \frac{\partial G}{\partial q_i} \frac{\partial}{\partial p_i} \right),$$

is *smoothing* of order 1; that is, it defines a real analytic map from $\ell^{a,s} \times \ell^{a,s}$ into $\ell^{a,s+1} \times \ell^{a,s+1}$. In particular, for the nonlinearity u^3 one finds

$$G = \frac{1}{4} \int_0^\pi |u(x)|^4 \, dx = \frac{1}{4} \sum_{i,j,k,l} G_{ijkl} q_i q_j q_k q_l$$

with

$$G_{ijkl} = \frac{1}{\sqrt{\omega_i \omega_j \omega_k \omega_l}} \int_0^\pi \chi_i \chi_j \chi_k \chi_l \, dx.$$

In [P96] it is proved that $G_{ijkl} = 0$, unless $i \pm j \pm k \pm l = 0$, for some combination of plus and minus signs. In particular,

$$G_{iijj} = \frac{1}{2\pi} \frac{2 + \delta_{ij}}{\omega_i \omega_j}. \tag{14}$$

From now on, we focus our attention on the nonlinearity $f(u) = u^3$, since terms of order five or more do not make any difference.

2.1. Partial Birkhoff normal form. For the rest of this paper we introduce the complex coordinates

$$\mathrm{z}_i = \frac{1}{\sqrt{2}}(q_i + \mathrm{i}p_i), \quad \bar{\mathrm{z}}_i = \frac{1}{\sqrt{2}}(q_i - \mathrm{i}p_i) \tag{15}$$

that live in the now *complex* Hilbert space

$$\ell^{a,s} = \ell^{a,s}(\mathbb{C}) := \left\{ \mathrm{z} = (\mathrm{z}_1, \ldots),\ \mathrm{z}_i \in \mathbb{C},\ i \geq 1 \text{ s.t. } \|\mathrm{z}\|_{a,s}^2 = \sum_{i\geq 1} |\mathrm{z}_i|^2 i^{2s} e^{2ai} < \infty \right\},$$

with symplectic structure $-\mathrm{i}\sum_{i\geq 1} d\mathrm{z}_i \wedge d\bar{\mathrm{z}}_i = \sum_{i\geq 1} dp_i \wedge dq_i$. The Hamiltonian becomes

$$H = \Lambda + G = \sum_{i\geq 1} \omega_i |\mathrm{z}_i|^2 + G(\mathrm{z}, \bar{\mathrm{z}}), \tag{16}$$

where, with abuse of notation, we have still denoted by G the function $G(\mathrm{z}, \bar{\mathrm{z}}) = G(p, q)$. The Hamiltonian equations write $\dot{\mathrm{z}} = -\mathrm{i}\partial_{\bar{\mathrm{z}}} H$, $\dot{\bar{\mathrm{z}}} = \mathrm{i}\partial_{\mathrm{z}} H$. The Hamiltonian H

is real analytic. Real analytic means that H is a function of z and $\bar{\mathrm{z}}$, real analytic in the real and imaginary parts of z; we denote by $A(\ell^{a,s}, \ell^{a,s+1})$ the class of all real analytic maps from some neighborhood of the origin in $\ell^{a,s}$ into $\ell^{a,s+1}$.

Notation. Given a finite subset $\mathcal{I} = \{i_1, \dots, i_N\}$ of $\mathbb{N}^+$, we will denote by $\hat{\mathrm{z}}$ the infinite vector obtained by excising $\mathrm{z} = (\mathrm{z}_1, \mathrm{z}_2, \dots)$ of its $\mathcal{I}$-components, namely, $\hat{\mathrm{z}} := (\dots, \mathrm{z}_{i_1-1}, \mathrm{z}_{i_1+1}, \dots, \mathrm{z}_{i_j-1}, \mathrm{z}_{i_j+1}, \dots, \mathrm{z}_{i_N-1}, \mathrm{z}_{i_N+1}, \dots) = (\mathrm{z}_i)_{i\in\mathcal{I}^c}$, where $\mathcal{I}^c := \mathbb{N}^+ \setminus \mathcal{I}$. The symbol "$\vee$" will mean "or" in the sense of the Latin "vel"; for example, $i \vee j \in \mathcal{I}$ means that one of the following three cases holds: (1) $i \in \mathcal{I}$, $j \notin \mathcal{I}$; (2) $i \notin \mathcal{I}$, $j \in \mathcal{I}$; (3) $i, j \in \mathcal{I}$. Fix $a > 0$, $s > 1$. We will denote by const > 0 and $0 < c_i < 1$, $i = 1, 2, \dots$, suitable constants depending only on $\mathcal{I}, a, s$; moreover, $y = O(x)$ means that $|y| \le \text{const}\, x$. In the following, we will often omit the explicit expressions for $\bar{\mathrm{z}}$, since they can be derived by analogous expressions for z.

Next, following [P96], we transform the Hamiltonian H in (16) into some partial Birkhoff normal form of order four so that, in a small neighborhood of the origin, it appears as a perturbation of a nonlinear integrable system. However, as we have just said in the introduction, the normal form degenerates when μ is close to zero, in the sense that its domain shrinks to zero and the remainder blows up. Then, we need a quantitative version of the Birkhoff normal form, in which we explicitly investigate the dependence on μ, for μ small. Such an analysis is not available in literature.

PROPOSITION 2.4 (Birkhoff normal form). *Let be $0 < \mu < 1$, $\mathcal{I} \subset \mathbb{N}^+$. There exists a real analytic, close to the identity, symplectic change of coordinates $\mathrm{z} := \Gamma(\mathrm{z}_*)$ defined in $B_r \subset \ell^{a,s}$ into $B_{2r} \subset \ell^{a,s}$ with*

$$r := c_1\sqrt{\mu}, \tag{17}$$

verifying

$$\|\mathrm{z} - \mathrm{z}_*\|_{a,s+1} = O\left(\frac{\|\mathrm{z}_*\|^3_{a,s}}{\mu}\right), \tag{18}$$

transforming the Hamiltonian $H = \Lambda + G$ in (16) *into seminormal form up to order six. That is,*

$$H \circ \Gamma = \Lambda + \bar{G} + \hat{G} + K,$$

where

$$X_{\bar{G}}, X_{\hat{G}}, X_K \in A(\ell^{a,s}, \ell^{a,s+1}), \tag{19}$$

$$\bar{G} = \frac{1}{2} \sum_{i\vee j\,\in\mathcal{I}} \bar{G}_{ij} |\mathrm{z}_{*i}|^2 |\mathrm{z}_{*j}|^2,$$

with uniquely determined coefficients $\bar{G}_{ij} = (3/8\pi)\,(4 - \delta_{ij}/\omega_i\omega_j)$, and

$$|\hat{G}| = O(\|\hat{\mathrm{z}}_*\|^4_{a,s}), \qquad |K| = O\left(\frac{\|\mathrm{z}_*\|^6_{a,s}}{\mu}\right).$$

Remark 2.5. It is worth pointing out that the Hamiltonian $\Lambda + \bar{G}$ is integrable with integrals $|\mathrm{z}_{*i}|^2, i = 1, 2, \dots$. Moreover, although the fourth order term $\hat{G}$ is not

integrable, it depends only on $\hat{z}_* := (z_{*i})_{i\in\mathcal{I}^c}$; namely, it is independent of the $\mathcal{I}$-modes.

Proof. Let us introduce another set of coordinates $(\dots, w_{-2}, w_{-1}, w_1, w_2, \dots)$ in $\ell^{a,s}_b$ defined by $z_{*i} = w_i$ and $\bar{z}_{*i} = w_{-i}$. The Hamiltonian becomes

$$\begin{aligned} H &= \Lambda + G \\ &= \sum_{i\geq 1} \omega_i\, z_{*i}\bar{z}_{*i} + \frac{1}{16}\sum_{i,j,k,\ell} G_{ijk\ell}(z_{*i}+\bar{z}_{*i})\cdots(z_{*\ell}+\bar{z}_{*\ell}) \\ &= \sum_{i\geq 1} \omega_i\, w_i w_{-i} + \frac{1}{16}\sum_{i,j,k,\ell}{}' G_{ijk\ell}\, w_i w_j w_k w_\ell, \end{aligned}$$

where the prime means that the summation is over all nonzero integers. The coefficients are defined for arbitrary integers by setting $G_{ijk\ell} = G_{|i|,|j|,|k|,|\ell|}$. We notice that $G_{ijkl} = 0$, unless $i \pm j \pm k \pm l = 0$, for some combination of plus and minus signs. The transformation Γ is obtained as the time-1-map of the flow of the Hamiltonian vectorfield X_F given by a Hamiltonian

$$F = \sum_{i,j,k,\ell}{}' F_{ijk\ell}\, w_i w_j w_k w_\ell, \tag{20}$$

with coefficients

$$\mathrm{i}F_{ijk\ell} = \begin{cases} \dfrac{G_{ijk\ell}}{16(\omega'_i+\omega'_j+\omega'_k+\omega'_\ell)} & \text{for } \dfrac{(i,j,k,\ell)\ \in\ \mathcal{L}_\mathcal{I}}{\mathcal{N}_\mathcal{I}}, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Here $\omega'_i = \operatorname{sign} i \cdot \omega_{|i|}$,

$$\mathcal{L}_\mathcal{I} = \big\{\, (i,j,k,\ell)\in\mathbb{Z}^4 \quad \text{s.t.} \quad |i|\vee|j|\vee|k|\vee|\ell| \in \mathcal{I}\,\big\},$$

and $\mathcal{N}_\mathcal{I} \subset \mathcal{L}_\mathcal{I}$ is the subset of all $(i,j,k,\ell) = (p,-p,q,-q)$ or some permutation of it. For these indices the denominator $\omega'_i+\omega'_j+\omega'_k+\omega'_\ell$ vanishes identically in μ. In [P96] the following is proved.

LEMMA 2.6. *Let i,j,k,ℓ be nonzero integers such that $i\pm j\pm k\pm l = 0$, but $(i,j,k,\ell)\neq(p,-p,q,-q)$; then*

$$|\omega'_i+\omega'_j+\omega'_k+\omega'_\ell| \geq \frac{\operatorname{const}\mu}{(M^2+\mu)^{3/2}}, \quad M = \min(|i|,\dots,|\ell|),$$

with some absolute const > 0.

Moreover, we note that F is real. Indeed,

$$\begin{aligned} \bar{F} &= \sum_{i,j,k,\ell}{}' \overline{F}_{ijk\ell}\, \bar{w}_i\bar{w}_j\bar{w}_k\bar{w}_\ell \\ &= \sum_{i,j,k,\ell}{}' \mathrm{i}\frac{G_{ijk\ell}}{\omega'_i+\omega'_j+\omega'_k+\omega'_\ell}\, w_{-i}w_{-j}w_{-k}w_{-\ell} \\ &= -\sum_{i,j,k,\ell}{}' \mathrm{i}\frac{G_{ijk\ell}}{\omega'_i+\omega'_j+\omega'_k+\omega'_\ell}\, w_i w_j w_k w_\ell = F, \end{aligned}$$

where we used that $G_{ijk\ell} = G_{-i,-j,-k,-\ell} = G_{|i|,|j|,|k|,|\ell|}$ and $\omega'_{-i} = -\omega'_i$. Expanding at $t = 0$ with the Taylor's formula we obtain

$$\begin{aligned} H \circ \Gamma &= H \circ X_F^t\big|_{t=1} \\ &= H + \{H, F\} + \int_0^1 (1-t)\big\{\{H, F\}, F\big\} \circ X_F^t \, dt \\ &= \Lambda + G + \{\Lambda, F\} + \{G, F\} + \int_0^1 (1-t)\big\{\{H, F\}, F\big\} \circ X_F^t \, dt, \end{aligned}$$

where $\{\cdot, \cdot\}$ denote the Poisson brackets. We can compute

$$\{\Lambda, F\} = -\mathrm{i} \sum_{i,j,k,\ell}{}' (\omega'_i + \omega'_j + \omega'_k + \omega'_\ell) F_{ijk\ell} \, w_i w_j w_k w_\ell;$$

thus

$$G + \{\Lambda, F\} = \frac{1}{16} \left(\sum_{(i,j,k,\ell) \in \mathcal{N}_\mathcal{I}} + \sum_{(i,j,k,\ell) \notin \mathcal{L}_\mathcal{I}} G_{ijk\ell} \, w_i w_j w_k w_\ell \right) = \bar{G} + \hat{G},$$

where $\hat{G}$ is independent of the $\mathcal{I}$-coordinates.

In the variables z_*,$\bar{\mathrm{z}}_*$ we find, from (14) and counting the multiplicities, that

$$\bar{G} = \frac{1}{2} \sum_{i \vee j \in \mathcal{I}} \bar{G}_{ij} |\mathrm{z}_{*i}|^2 |\mathrm{z}_{*j}|^2,$$

with uniquely determined coefficients

$$\bar{G}_{ij} = \begin{cases} 24 G_{iijj} = \dfrac{3}{2\pi} \dfrac{1}{\omega_i \omega_j} & \text{for } i \neq j, \\ 12 G_{iiii} = \dfrac{9}{8\pi} \dfrac{1}{\omega_i \omega_j} & \text{for } i = j. \end{cases} \tag{22}$$

Hence, we have $H \circ \Gamma = \Lambda + \bar{G} + \hat{G} + K$, where

$$K = \{G, F\} + \int_0^1 (1-t)\big\{\{H, F\}, F\big\} \circ X_F^t \, dt \tag{23}$$

is composed by all the terms of order six or more.

Claim. The vectorfield of the Hamiltonian F is analytic, that is

$$X_F \in \mathcal{A}(\ell_b^{a,s}, \ell_b^{a,s+1}). \tag{24}$$

In fact, from Lemma 2.6, it results in

$$\begin{aligned} \left| \frac{\partial F}{\partial w_\ell} \right| &\leq \sum\nolimits'_{\pm i \pm j \pm k = l} |F_{ijk\ell}| \, |w_i w_j w_k| \\ &\leq \frac{\text{const}}{\mu \sqrt{\ell}} \sum\nolimits'_{\pm i \pm j \pm k = l} \frac{|w_i w_j w_k|}{\sqrt{|ijk|}} \\ &\leq \frac{\text{const}}{\mu \sqrt{\ell}} \sum\nolimits'_{i+j+k=l} \tilde{w}_i \tilde{w}_j \tilde{w}_k = \frac{\text{const}}{\mu \sqrt{\ell}} (\tilde{w} * \tilde{w} * \tilde{w})_\ell, \end{aligned}$$

where $\tilde{w}_i = \frac{|w_i|+|w_{-i}|}{\sqrt{|i|}}$. If $w \in \ell^{a,s}_b$, then $\tilde{w} \in \ell^{a,s+1/2}_b$, which is a Hilbert algebra for $s > 0$ by Lemma 2.2; thus $\tilde{w} * \tilde{w} * \tilde{w}$ also belongs to $\ell^{a,s+1/2}_b$. Therefore $F_w \in \ell^{a,s+1}_b$, with

$$\|F_w\|_{a,s+1} \leq \frac{\text{const}}{\mu}\|\tilde{w} * \tilde{w} * \tilde{w}\|_{a,s+1/2} \leq \frac{\text{const}}{\mu}\|\tilde{w}\|^3_{a,s+1/2} \leq \frac{\text{const}}{\mu}\|w\|^3_{a,s}. \tag{25}$$

The analyticity of F_w follows from the analyticity of each component function and its local boundedness, proving (24). From (24) and (25) it follows that the Hamiltonian flow X^t_F is well defined, in a sufficiently small neighborhood of the origin in $\ell^{a,s}$ for all $0 \leq t \leq 1$; in particular, by (25), for $\frac{\|w\|_{a,s}}{2} = \|\mathrm{z}_*\|_{a,s} \leq r$ the map $\Gamma := X^1_F$ verifies

$$\|\Gamma(\mathrm{z}_*) - \mathrm{z}_*\|_{a,s+1} \leq \frac{\text{const}}{\mu}\|\mathrm{z}_*\|^3_{a,s} \leq \text{const}\, c_1^3\sqrt{\mu} \leq c_1\sqrt{\mu} = r, \tag{26}$$

taking c_1 small enough in (17). In the same way (taking c_1 small enough),

$$\|D\Gamma - \mathbb{I}\|^{\text{op}}_{a,s+1,s} \leq \frac{\text{const}}{\mu}\, r^2 = \text{const}\, c_1^2 \leq \frac{1}{2},$$

where the operator norm $\|\cdot\|^{\text{op}}_{a,r,s}$, is defined by

$$\|\cdot\|^{\text{op}}_{a,r,s} = \sup_{w\neq 0}\frac{\|Aw\|_{a,r}}{\|w\|_{a,s}}.$$

Accordingly, $\Gamma : \ell^{a,s} \supset B_r \to B_{2r} \subset \ell^{a,s}$ is a real analytic, symplectic change of coordinates, and (18) follows from (26); moreover, since $\|D\Gamma - \mathbb{I}\|^{\text{op}}_{a,s+1,s+1} \leq \|D\Gamma - \mathbb{I}\|^{\text{op}}_{a,s+1,s}$, $D\Gamma$ defines an isomorphism of $B_r \subset \ell^{a,s+1}$. It follows that with $X_H \in \mathcal{A}(\ell^{a,s}, \ell^{a,s+1})$, we also have

$$D\Gamma^{-1} X_H \circ \Gamma = X_{H\circ\Gamma} \in \mathcal{A}(\ell^{a,s}, \ell^{a,s+1}).$$

The same holds for the Lie bracket: the boundedness of $\|DX_F\|^{\text{op}}_{a,s+1,s}$ implies that

$$[X_F, X_H] = X_{\{H,F\}} \in \mathcal{A}(\ell^{a,s}, \ell^{a,s+1}).$$

These two facts show that $X_K \in \mathcal{A}(\ell^{a,s}, \ell^{a,s+1})$. The analogue claims for $X_{\bar{G}}$ and $X_{\hat{G}}$ are obvious.

Finally, recalling (23), we can write

$$K = \{G, F\} + \int_0^1 (1-t)\big[\big\{\{\Lambda, F\}, F\big\} + \big\{\{G, F\}, F\big\}\big] \circ X^t_F\, dt. \tag{27}$$

It results in

$$\{G, F\} = O\left(\frac{\|w\|^6_{a,s}}{\mu}\right)$$

and

$$\big\{\{\Lambda, F\}, F\big\} = O\left(\frac{\|w\|^6_{a,s}}{\mu}\right), \tag{28}$$

since $\{\Lambda, F\} = \bar{G} + \hat{G} - G = O(\|w\|^4_{a,s})$. Moreover,

$$\big\{\{G, F\}, F\big\} = O\left(\frac{\|w\|^8_{a,s}}{\mu^2}\right); \tag{29}$$

hence, by (27)–(29),

$$|K| = O\left(\frac{\|w\|^6_{a,s}}{\mu}\right) \qquad \text{for} \quad \|w\|_{a,s} \le \text{const}\,\sqrt{\mu}.$$

Since we are looking for small amplitude solutions it is convenient to introduce the small parameter $0 < \eta < 1$ and perform the following rescaling:

$$\mathrm{z}_* =: \eta z, \quad \bar{\mathrm{z}}_* =: \eta \bar{z}, \qquad H \circ \Gamma \longrightarrow \eta^{-2}\big(H \circ \Gamma\big) =: \mathcal{H}, \tag{30}$$

by which the Hamiltonian is written

$$\mathcal{H}(z, \bar{z}; \eta) = \Lambda + \eta^2(\bar{G} + \hat{G}) + \eta^4 \widetilde{K}(z, \bar{z}; \eta), \qquad \|z\|_{a,s}, \|\bar{z}\|_{a,s} \le c_1 \frac{\sqrt{\mu}}{\eta}, \tag{31}$$

where

$$\widetilde{K}(z, \bar{z}; \eta) := \eta^{-2} K(\eta z, \eta \bar{z}), \qquad |\widetilde{K}| = O\left(\frac{\|z\|^6_{a,s}}{\mu}\right). \tag{32}$$

Remark 2.7. We note that the rescaling in (30) does not introduce a rescaling of time. In fact if $\big(z(t), \bar{z}(t)\big)$ is a solution of the Hamiltonian equations for $\mathcal{H}$,

$$\begin{cases} \dot{z}(t) = -\mathrm{i}\partial_{\bar{z}} \mathcal{H}\big(z(t), \bar{z}(t)\big) = -\mathrm{i}\eta^{-1} \partial_{\bar{\mathrm{z}}_*} H\big(\eta z(t), \eta \bar{z}(t)\big), \\ \dot{\bar{z}}(t) = \mathrm{i}\partial_z \mathcal{H}\big(z(t), \bar{z}(t)\big) = \mathrm{i}\eta^{-1} \partial_{\mathrm{z}_*} H\big(\eta z(t), \eta \bar{z}(t)\big), \end{cases}$$

then $(\mathrm{z}_*(t), \bar{\mathrm{z}}_*(t)) = (\eta z(t), \eta \bar{z}(t))$ is a solution of the Hamiltonian equations for H.

We now introduce action-angle variables $(I, \phi) \in \mathbb{R}^N_+ \times \mathbb{T}^N$ on the $\mathcal{I}$-modes by the following symplectic change of variables:

$$z_i := \sqrt{I_i}(\cos\phi_i - \mathrm{i}\sin\phi_i), \qquad \bar{z}_i := \sqrt{I_i}(\cos\phi_i + \mathrm{i}\sin\phi_i), \qquad i \in \mathcal{I}. \tag{33}$$

The action $I := (I_i)_{i \in \mathcal{I}}$, $I_i := z_i \bar{z}_i$, is defined for

$$|I| \le c_2 \frac{\mu}{\eta^2}. \tag{34}$$

We note that $\sum_{i \in \mathcal{I}} dI_i \wedge d\phi_i = -\mathrm{i} \sum_{i \in \mathcal{I}} dz_i \wedge d\bar{z}_i = \sum_{i \in \mathcal{I}} dp_i \wedge dq_i$, and the *phase space* is[4]

$$\mathcal{P}_{a,s} := \mathbb{R}^N_+ \times \mathbb{T}^N \times \ell^{a,s} \ni (I, \phi, \hat{z}). \tag{35}$$

In these variables the Hamiltonian becomes

$$\begin{aligned} \widetilde{\mathcal{H}}(I, \phi, \hat{z}, \bar{\hat{z}}; \eta) = \omega \cdot I + \Omega \cdot \hat{z}\bar{\hat{z}} + \eta^2 \left[\frac{1}{2}(AI, I) + (BI, \hat{z}\bar{\hat{z}}) + \hat{G}(\hat{z}, \bar{\hat{z}})\right] \\ + \eta^4 \widetilde{K}(I, \phi, \hat{z}, \bar{\hat{z}}; \eta), \end{aligned} \tag{36}$$

[4]Clearly here $\ell^{a,s} = \ell^{a,s}(\mathbb{C})$.

where

$$\omega := (\omega_{i_1}, \dots, \omega_{i_N}),$$
$$\Omega := (\dots, \omega_{i_1-1}, \omega_{i_1+1}, \dots, \omega_{i_j-1}, \omega_{i_j+1}, \dots, \omega_{i_N-1}, \omega_{i_N+1}, \dots);$$

$\Omega \cdot \hat{z}\bar{\hat{z}}$ is short for $\sum_{i\in\mathcal{I}^c} \omega_i \hat{z}_i \bar{\hat{z}}_i$, A is the $N \times N$ matrix

$$A = A_{\mathcal{I}} := \left(\bar{G}_{ij}\right)_{i,j\in\mathcal{I}},$$

and B is the $\infty \times N$ matrix

$$B = B_{\mathcal{I}} := \left(\bar{G}_{ij}\right)_{i\in\mathcal{I}^c,\, j\in\mathcal{I}} .$$

Moreover, $(\cdot,\cdot)$ denotes the standard scalar product, and we have denoted again by $\widetilde{K}$ the function $\widetilde{K}(I,\phi,\hat{z},\bar{\hat{z}};\eta) = \widetilde{K}(z,\bar{z};\eta)$. Recalling (22), we have

$$A = \frac{3}{8\pi}\begin{pmatrix} \frac{3}{\omega_{i_1}^2} & \frac{4}{\omega_{i_1}\omega_{i_2}} & \cdots & \frac{4}{\omega_{i_1}\omega_{i_N}} \\ \frac{4}{\omega_{i_2}\omega_{i_1}} & \frac{3}{\omega_{i_2}^2} & \cdots & \frac{4}{\omega_{i_2}\omega_{i_N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{4}{\omega_{i_N}\omega_{i_1}} & \frac{4}{\omega_{i_N}\omega_{i_2}} & \cdots & \frac{3}{\omega_{i_N}^2} \end{pmatrix},$$

$$B = \frac{3}{8\pi}\begin{pmatrix} \vdots & \vdots & \vdots \\ \frac{4}{\omega_{i_j-1}\omega_{i_1}} & \cdots & \frac{4}{\omega_{i_j-1}\omega_{i_N}} \\ \frac{4}{\omega_{i_j+1}\omega_{i_1}} & \cdots & \frac{4}{\omega_{i_j+1}\omega_{i_N}} \\ \vdots & \vdots & \vdots \end{pmatrix}.$$

Defining the matrices

$$D = D_{\mathcal{I}} := \operatorname{diag}[\omega] \in \mathrm{Mat}_{N\times N}, \qquad E = E_{\mathcal{I}} := \operatorname{diag}[\Omega] \in \mathrm{Mat}_{\infty\times\infty}, \tag{37}$$

we can rewrite A and B as

$$A = \frac{3}{8\pi} D^{-1}\tilde{A}D^{-1}, \qquad B = \frac{3}{2\pi} E^{-1}\tilde{B}D^{-1}, \tag{38}$$

where

$$\tilde{A} := \begin{pmatrix} 3 & 4 & \dots & 4 \\ 4 & 3 & \dots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 4 & 4 & \dots & 3 \end{pmatrix} \in \mathrm{Mat}_{N\times N}, \qquad \tilde{B} = \begin{pmatrix} 1 & \dots & 1 \\ 1 & \dots & 1 \\ \vdots & \vdots & \vdots \end{pmatrix} \in \mathrm{Mat}_{\infty\times\infty}. \tag{39}$$

We note that the matrix A is invertible, since

$$\det\tilde{A} =: d_N = 3d_{N-1} + (-1)^N(1-N)4^2 = (-1)^N(1-4N) \neq 0 \qquad (d_N \text{ is odd}). \tag{40}$$

Moreover,

$$\tilde{A}^{-1} = \frac{1}{4N-1}\begin{pmatrix} 5-4N & 4 & \dots & 4 \\ 4 & 5-4N & \dots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 4 & 4 & \dots & 5-4N \end{pmatrix}. \tag{41}$$

3. Long-period orbits. We will find periodic solutions of the Hamiltonian $\widetilde{\mathcal{H}}$ in (36) close to the ones of the integrable Hamiltonian

$$\widetilde{\mathcal{H}}_{int} = \omega \cdot I + \Omega \cdot \hat{z}\bar{\hat{z}} + \eta^2 \left[\frac{1}{2}(AI, I) + (BI, \hat{z}\bar{\hat{z}}) \right], \tag{42}$$

in which $\hat{G}$ and $\widetilde{K}$ have been neglected. The equations of motion for $\widetilde{\mathcal{H}}_{int}$,

$$\begin{cases} \dot{I} = 0, \\ \dot{\phi} = \omega + \eta^2 AI + \eta^2 B^t \hat{z}\bar{\hat{z}}, \\ \dot{\hat{z}}_i = -\mathrm{i}(\Omega + \eta^2 BI)_i \, \hat{z}_i, \qquad i \in \mathcal{I}^c, \end{cases} \tag{43}$$

can be easily integrated:

$$\begin{cases} I(t) = I_0 \\ \phi(t) = \phi_0 + \tilde{\omega} t + \eta^2 B^t \hat{z}_0 \bar{\hat{z}}_0 t \\ \hat{z}_i(t) = e^{-\mathrm{i}\tilde{\Omega}_i t} \, (\hat{z}_0)_i, \qquad i \in \mathcal{I}^c, \end{cases} \tag{44}$$

where

$$\tilde{\omega} := \tilde{\omega}(I_0, \eta) = \omega + \eta^2 A I_0 \tag{45}$$

is the vector of the *shifted linear frequencies*, and

$$\tilde{\Omega}_i := \tilde{\Omega}_i(I_0, \eta) = \Omega_i + \eta^2 (BI_0)_i = \omega_i + \eta^2 (BI_0)_i, \qquad i \in \mathcal{I}^c, \tag{46}$$

are the *shifted elliptic frequencies*. Consequently, for (42) $\{\hat{z} = 0\}$ is an invariant manifold which is completely foliated by the N-dimensional invariant tori

$$\mathcal{T}(I_0) := \big\{ I = I_0, \, \phi \in \mathbb{T}^N, \, \hat{z} = 0 \big\}.$$

On $\mathcal{T}(I_0)$ the flow

$$t \mapsto (I_0, \phi_0 + \tilde{\omega} t, 0)$$

is T-periodic, $T > 0$, if and only if

$$\tilde{\omega}(I_0, \eta)\tau =: k \in \mathbb{Z}^N, \tag{47}$$

where

$$\tau := \frac{T}{2\pi}$$

is the rescaled period. Hence, if (47) holds, the torus $\mathcal{T}(I_0)$ is completely resonant and supports the infinitely many T-periodic orbits of the family

$$\mathcal{F} := \big\{ I(t) = I_0, \quad \phi(t) = \phi_0 + \tilde{\omega} t, \quad \hat{z}(t) = 0 \big\}. \tag{48}$$

The family $\mathcal{F}$ will not persist in its entirety for the Hamiltonian $\widetilde{\mathcal{H}}$. However, we claim that if the period T is "sufficiently nonresonant" with the shifted elliptic frequencies, we can prove the persistence of at least N geometrically distinct T-periodic solutions of $\widetilde{\mathcal{H}}$ close to $\mathcal{F}$. More precisely, the required nonresonance condition is

$$\big| \ell - \tilde{\Omega}_i(I_0, \eta)\tau \big| \geq \frac{\text{const}}{i} \qquad \forall \, \ell \in \mathbb{Z}, \ \forall \, i \in \mathcal{I}^c. \tag{49}$$

We now consider the periodicity condition (47). As we have said in the introduction, we construct a set of actions and of completely resonant frequencies, parametrized by the (rescaled) periods τ. Such actions and frequencies are related by the action-to-frequency map $I \to (\partial_I \widetilde{\mathcal{H}}_{int})_{|\hat{z}=0} = \omega + \eta^2 AI$. Indeed, since A is invertible we can choose I_0 and k as functions of τ and η so that (47) is always satisfied:

$$I_0 := \frac{1}{\eta^2\tau} A^{-1}\left(\kappa - \{\omega\tau\}\right), \tag{50}$$

$$k := [\omega\tau] + \kappa, \tag{51}$$

where $[(x_1,\dots,x_N)] := ([x_1],\dots,[x_N])$, $\{(x_1,\dots,x_N)\} := (\{x_1\},\dots,\{x_N\})$, and

$$\kappa := (\kappa_i)_{i\in\mathcal{I}} \in \mathbb{Z}^N, \qquad \kappa_i := i^{-1}\tilde{\kappa}, \qquad \tilde{\kappa} := 10\prod_{j\in\mathcal{I}} j. \tag{52}$$

Here the functions $[\cdot] : \mathbb{R} \to \mathbb{Z}$ and $\{\cdot\} : \mathbb{R} \to [0,1)$ denote the integer part and the fractional part, respectively.

In order to have $I_0 \approx 1$ in (50), we choose the parameter η, which is related to the amplitude of the solution, as a function of the rescaled period τ such that

$$\eta^2\tau = 1, \qquad \text{namely,} \qquad \eta := 1/\sqrt{\tau}. \tag{53}$$

Consequently, we can express I_0, k, and $\tilde{\Omega}_i$ in (50), (51), (46) as functions of τ only:

$$I_0 := I_0(\tau) = A^{-1}\left(\kappa - \{\omega\tau\}\right), \tag{54}$$

$$k := k(\tau) = [\omega\tau] + \kappa, \tag{55}$$

$$\tilde{\Omega}_i := \tilde{\Omega}_i(\tau) = \omega_i + \eta^2\big(BI_0(\tau)\big)_i \qquad \text{for} \quad i \in \mathcal{I}^c. \tag{56}$$

We point out that the constant vector κ defined in (52) has been added to have $(I_0)_i > 0$ in view of (35). In particular the following lemma holds.

LEMMA 3.1. *If μ is small enough, then $(I_0)_i > \pi\omega_i$ for all $i \in \mathcal{I}$.*

Proof. By (38) and (54) we get

$$I_0 = \frac{8\pi}{3} D\tilde{A}^{-1} D\big(\kappa - \{\omega t\}\big).$$

Recalling (37) and (41) we have, for all $i \in \mathcal{I}$,

$$\begin{aligned}(I_0)_i &= \frac{8\pi\omega_i}{3(4N-1)}\left((1-4N)\big(\omega_i\kappa_i - \omega_i\{\omega_i t\}\big) + 4\sum_{j\in\mathcal{I}}\big(\omega_j\kappa_j - \omega_j\{\omega_j t\}\big)\right)\\ &= \frac{8\pi\omega_i}{3(4N-1)}\left((1-4N)\big(\tilde{\kappa} - i\{\omega_i t\}\big) + 4\sum_{j\in\mathcal{I}}\big(\tilde{\kappa} - j\{\omega_j t\}\big) + O(\mu)\right)\end{aligned}$$

since, recalling (52), $\omega_i\kappa_i = i^{-1}\omega_i\tilde{\kappa}$ and $1 < i^{-1}\omega_i = i^{-1}\sqrt{i^2+\mu} < 1+\mu$. Taking μ small enough we get

$$\begin{aligned}(I_0)_i &= \frac{8\pi\omega_i}{3(4N-1)}\left(\tilde{\kappa} + (4N-1)i\{\omega_i t\} - 4\sum_{j\in\mathcal{I}} j\{\omega_j t\} + O(\mu)\right)\\ &\geq \frac{8\pi\omega_i}{3(4N-1)}\left(\tilde{\kappa} - 4\sum_{j\in\mathcal{I}} j + O(\mu)\right) = \frac{8\pi\omega_i}{3(4N-1)}\left(10\prod_{j\in\mathcal{I}} j - 4\sum_{j\in\mathcal{I}} j + O(\mu)\right)\\ &\geq \frac{8\pi\omega_i}{3(4N-1)}\left(2\prod_{j\in\mathcal{I}} j + O(\mu)\right) \geq \frac{8\pi\omega_i}{3(4N-1)}\big(2N + O(\mu)\big) \geq \frac{8\pi\omega_i}{3(4N-1)}\frac{3N}{2}\\ &> \pi\omega_i,\end{aligned}$$

where we used that $2\prod_{j\in\mathcal{I}} j \geq \sum_{j\in\mathcal{I}} j$ and $\prod_{j\in\mathcal{I}} j \geq N$. □

We note that by the choice of η made in (53), it results that $|I_0| \leq$ const. Then I_0 belongs to the domain of definition of the Hamiltonian $\widetilde{\mathcal{H}}$, namely, it verifies (34), making the hypothesis

$$\mu\tau \geq c_3^{-1}. \tag{57}$$

We finally remark that, by (53) and (56), the quantities that we have to estimate in the crucial nonresonance condition (49) are

$$\ell - \tilde{\Omega}_i\tau = \ell - \tau\omega_i - \Big(BA^{-1}\big(\kappa - \{\omega\tau\}\big)\Big)_i, \qquad \ell \in \mathbb{Z}, i \in \mathcal{I}^c. \tag{58}$$

In this form, (49) clearly appears as a nonresonance condition between the frequency of the torus ω and the normal frequencies $\{\omega_i\}_{i\in\mathcal{I}^c}$.

3.1. Small divisors estimate. In order to estimate the quantities in (58) we will perform the expansion $\tau\omega_i = \tau\sqrt{i^2+\mu} = i\tau + \frac{\mu\tau}{2i} + O(\mu^2\tau)$, requiring that $\mu^2\tau$ is small, namely,

$$\mu^2\tau \leq c_4.$$

The other aspect of such a request is that, for any fixed μ, we have only a finite number of (rescaled) periods τ. Moreover, we note that the smallness of $\mu^2\tau$ implies that of μ since $\mu \leq c_3\mu^2\tau$ by (57). Hence, since for μ close to zero the Birkhoff normal form degenerates, it is not obvious that we can make $\mu^2\tau$ small for some fixed $\tau \geq 1$. Moreover, even if it is not necessary, for simplicity we limit our consideration to $\tau \sim \mu^{-2}$, namely, $c_4/2\mu^2 \leq \tau \leq c_4/\mu^2$. Again for technical reasons, we will need that τ is an *integer* and that $\mu\tau$ is far away from even integers. Let us define

$$\mathcal{T}_\mu := \left\{ \tau \in \mathbb{N}^+, \quad \frac{c_4}{2\mu^2} \leq \tau \leq \frac{c_4}{\mu^2} \quad \text{s.t.} \quad \mu\tau \in \mathcal{N} \right\}, \tag{59}$$

where

$$\mathcal{N} := \left\{ x > 0 \ \text{ s.t. } \ |x - 2m| \geq \frac{1}{2} \ \ \forall\, m \in \mathbb{Z} \right\} = \bigcup_{n>0 \text{ odd}} \left[n - \frac{1}{2}, n + \frac{1}{2} \right]. \tag{60}$$

The constant c_4 will be choose suitably small in the following.

PROPOSITION 3.2. *If μ is small enough and $\tau \in \mathcal{T}_\mu$, then the following estimate holds:*

$$\big|\ell - \tilde{\Omega}_i\tau\big| \geq \frac{c_5}{i} \qquad\qquad \forall\, \ell \in \mathbb{Z},\ i \in \mathcal{I}^c. \tag{61}$$

Proof. We will prove that

$$\big|\ell - \tau\sqrt{i^2+\mu} - \big(BI_0(\tau)\big)_i\big| \geq \frac{1}{6(4N-1)\, i} \qquad \forall\, \ell \in \mathbb{Z}, i \in \mathcal{I}^c. \tag{62}$$

Recalling (58), we see that the crucial estimate (61) follows from (62) taking $c_5 := \frac{1}{6(4N-1)}$. We first consider the term $BI_0(\tau)$. From (38) we have $BA^{-1} = 4(E^{-1}\tilde{B}\tilde{A}^{-1}D)$, while, by (41), $\tilde{B}\tilde{A}^{-1} = d^{-1}\tilde{B}$, where $d := 4N-1$. Recalling (54), we get

$$BI_0(\tau) = BA^{-1}\big(\kappa - \{\omega\tau\}\big) = 4d^{-1}\, E^{-1}\tilde{B}D\big(\kappa - \{\omega\tau\}\big)$$

and, in particular, for $i \in \mathcal{I}^c$,

$$(BI_0)_i = \frac{4}{d}\left(E^{-1}\tilde{B}D\big(\kappa - \{\omega\tau\}\big)\right)_i = \frac{4}{d\omega_i}\left(\tilde{B}D\big(\kappa - \{\omega\tau\}\big)\right)_i$$

$$= \frac{4}{d\omega_i}\left(\hat{\kappa} - \sum_{h\in\mathcal{I}} \omega_h\{\omega_h\tau\}\right), \tag{63}$$

where

$$\hat{\kappa} := \big(\tilde{B}D\kappa\big)_i = \left(\tilde{B}\begin{pmatrix} \vdots \\ \dfrac{\omega_h}{h}\,\tilde{\kappa} \\ \vdots \end{pmatrix}\right)_i = \sum_{h\in\mathcal{I}} \frac{\omega_h}{h}\,\tilde{\kappa} = \tilde{\kappa}N + O(\mu). \tag{64}$$

Now we need the following lemma.

LEMMA 3.3. *Let $0 < \mu < 1$. Then for all $h \in \mathbb{N}^+$ there exist $\delta_h^{(k)} > 0$, $k = 1, 2, 3$, such that*

$$\omega_h = h + \delta_h^{(1)}, \qquad \delta_h^{(1)} < \frac{\mu}{2h} \tag{65}$$

$$= h + \frac{\mu}{2h} - \delta_h^{(2)}, \quad \delta_h^{(2)} < \frac{\mu^2}{8h^3}. \tag{66}$$

Moreover, if $\tau \in \mathbb{N}^+$, there exists $n_h := n_h(\mu, \tau) \in \mathbb{Z}$ such that

$$\left|\omega_h\{\omega_h\tau\} - \frac{\mu\tau}{2} - n_h\right| \le \frac{\mu}{2h} + \frac{\mu^2\tau}{8h^2} \qquad \forall\, h \in \mathbb{N}^+. \tag{67}$$

By the elementary inequality $0 < 1 - (1+x)^{-1} < x$ for all $x > 0$, we get, using (65),

$$0 < \frac{1}{i} - \frac{1}{\omega_i} = \frac{1}{i}\left(1 - \frac{1}{1 + \delta_i^{(1)}/i}\right) < \frac{\delta_i^{(1)}}{i^2} < \frac{\mu}{2i^3},$$

namely,

$$\frac{1}{\omega_i} - \frac{1}{i} = O\left(\frac{\mu}{i^3}\right). \tag{68}$$

Since $|\{\omega_h\tau\}| \le 1$, by substituting (68) into (63) we get

$$(BI_0)_i = \frac{4}{di}\left(\hat{\kappa} - \sum_{h\in\mathcal{I}} \omega_h\{\omega_h\tau\}\right) + O\left(\frac{\mu}{i^3}\right);$$

hence, by (67) and (64),

$$(BI_0)_i = \frac{4}{di}\left(\tilde{\kappa}N - \sum_{h\in\mathcal{I}}\left(\frac{\mu\tau}{2} + n_h\right)\right) + O\left(\frac{\mu}{i} + \frac{\mu^2\tau}{i}\right). \tag{69}$$

Moreover, since by (66) we get

$$\tau\sqrt{i^2 + \mu} = \tau i + \frac{\mu\tau}{2i} + O\left(\frac{\mu^2\tau}{i^3}\right),$$

using (69) we have

$$\ell - \tau\sqrt{i^2+\mu} - \big(BI_0(\tau)\big)_i = \ell - \tau i - \frac{\mu\tau}{2i} + \frac{4}{di}\left(\sum_{h\in\mathcal{I}}\left(\frac{\mu\tau}{2}+n_h\right) - \tilde{\kappa}N\right) + O\left(\frac{\mu}{i}+\frac{\mu^2\tau}{i}\right). \tag{70}$$

From the hypothesis $\tau \in \mathcal{T}_\mu$ and for μ small enough, it follows that

$$\frac{\mu}{i} + \frac{\mu^2\tau}{i} < \frac{2c_4}{i}.$$

Hence, by (70) and choosing c_4 small enough, in order to prove (62) it is sufficient to show that

$$\left|\ell - \tau i - \frac{\mu\tau}{2i} + \frac{4}{di}\left(\sum_{h\in\mathcal{I}}\left(\frac{\mu\tau}{2}+n_h\right) - \tilde{\kappa}N\right)\right| \geq \frac{1}{4(4N-1)\,i}. \tag{71}$$

Now, since $d = 4N-1$ and $\sum_{h\in\mathcal{I}} 1 = N$, (71) is equivalent to

$$\left|\mu\tau + 2\left(id(\ell-\tau i) - 4\tilde{\kappa}N + 4\sum_{h\in\mathcal{I}} n_h\right)\right| \geq \frac{1}{2}. \tag{72}$$

Since

$$2\left(id(\ell-\tau i) - 4\tilde{\kappa}N + 4\sum_{h\in\mathcal{I}} n_h\right)$$

is an even integer, (72) follows by hypothesis $\tau \in \mathcal{T}_\mu$. □

Proof of Lemma 3.3. Since $\omega_h = h\sqrt{1+x}$ with $x := \mu/h^2$, $0 < x < 1$, (65) and (66) directly follow by the elementary inequalities

$$1 < \sqrt{1+x}, \qquad -\frac{x^2}{8} < \sqrt{1+x} - 1 - \frac{x}{2} < 0,$$

holding for any $0 < x < 1$. We now prove (67). Because $|\{\omega_h\tau\}| \leq 1$, we have

$$|\omega_h\{\omega_h\tau\} - h\{\omega_h\tau\}| = |(\omega_h - h)\{\omega_h\tau\}| \leq |\omega_h - h| \leq \frac{\mu}{2h}, \tag{73}$$

where in the last inequality we have used (65). Moreover, by (66)

$$\{\omega_h\tau\} = \omega_h\tau - [\omega_h\tau] = h\tau + \frac{\mu\tau}{2h} - \delta_h^{(2)}\tau - [\omega_h\tau],$$

from which we get

$$\left|h\{\omega_h\tau\} - \frac{\mu\tau}{2} - n_h\right| \leq \frac{\mu^2\tau}{8h^2}, \qquad \text{where} \quad n_h := h^2\tau - h[\omega_h\tau].$$

Then the proof follows by the previous inequality and (73). □

We now give a lower estimate on the cardinality of $\mathcal{T}_\mu$.

LEMMA 3.4. *For μ small enough,*

$$\sharp\, \mathcal{T}_\mu \geq \frac{c_4}{6\mu^2}.$$

Proof. We first claim that

$$\sharp\left\{\tau \in \mathbb{N} \quad \text{s.t.} \quad \mu\tau \in \left[n-\frac{1}{2}, n+\frac{1}{2}\right]\right\} \geq \left[\frac{1}{\mu}\right] \tag{74}$$

for any n odd. Indeed, if $\tau_0 := \min\{\tau \in \mathbb{N} \text{ s.t. } \mu\tau \geq n-1/2\}$, then $\mu\tau_0 < \mu + n - 1/2$, and therefore $\mu\tau_0 + \mu m \leq n + 1/2$ for any $0 \leq m \leq -1 + 1/\mu$, proving (74). Since

$$\sharp\left\{n \text{ odd} \quad \text{s.t.} \quad \frac{c_4}{2\mu} + \frac{1}{2} \leq n \leq \frac{c_4}{\mu} - \frac{1}{2}\right\} \geq \frac{c_4}{4\mu} - 2 \geq \frac{c_4}{5\mu},$$

then

$$\sharp\, \mathcal{T}_\mu \geq \frac{c_4}{5\mu}\left[\frac{1}{\mu}\right] \geq \frac{c_4}{6\mu^2},$$

for μ small enough. □

3.2. Functional setting. Since the problem is Hamiltonian, any T-periodic solution of the Hamiltonian equations for $\widetilde{\mathcal{H}}$ in (36), namely,

$$\begin{cases} \dot{I} = -\eta^4 \partial_\phi \widetilde{K}(I, \phi, \hat{z}, \bar{\hat{z}}), \\ \dot{\phi} = \omega + \eta^2 AI + \eta^2 B^t z\bar{z} + \eta^4 \partial_I \widetilde{K}(I, \phi, \hat{z}, \bar{\hat{z}}), \\ \dot{\hat{z}}_i = -\mathrm{i}(\Omega + \eta^2 BI)_i\, \hat{z}_i - \mathrm{i}\eta^2 \partial_{\bar{z}_i} \hat{G}(\hat{z}, \bar{\hat{z}}) - \mathrm{i}\eta^4 \partial_{\bar{z}_i} \widetilde{K}(I, \phi, \hat{z}, \bar{\hat{z}}), & i \in \mathcal{I}^c, \end{cases} \tag{75}$$

is a critical point of the Lagrangian action functional

$$S(I, \phi, \hat{z}) = \int_0^T \left(I \cdot \dot{\phi} - \mathrm{i} \sum_{i \in \mathcal{I}^c} z_i \dot{\bar{z}}_i - \widetilde{\mathcal{H}}(I, \phi, \hat{z}, \bar{\hat{z}}) \right) dt, \tag{76}$$

in the space of T-periodic, $\mathcal{P}_{a,s}$-valued curves $\big(I(t), \phi(t), \hat{z}(t)\big)$.

In particular we are looking for periodic orbits of the Hamiltonian $\widetilde{\mathcal{H}}$ near the family $\mathcal{F}$ defined in (48); namely, we seek solutions of the form

$$\begin{cases} I(t) = I_0 + J(t), \\ \phi(t) = \phi_0 + \tilde{\omega} t + \psi(t), \\ \hat{z}(t) = 0 + w(t), \end{cases} \tag{77}$$

where I_0 was defined in (54); $\phi_0 \in \mathbb{T}^N$ is a parameter to determine. Recalling (75), $\zeta(t) = (J(t), \psi(t), w(t))$ and $\phi_0 \in \mathbb{T}^N$ must satisfy

$$\begin{cases} \dot{\psi} - \eta^2 AJ = \eta^2 B^t w\bar{w} + \eta^4 \partial_I \widetilde{K}(I_0 + J, \phi_0 + \tilde{\omega} t + \psi, w, \bar{w}), \\ \dot{J} \qquad\qquad = -\eta^4 \partial_\phi \widetilde{K}(I_0 + J, \phi_0 + \tilde{\omega} t + \psi, w, \bar{w}), \\ \dot{w}_i + \mathrm{i}\tilde{\Omega}_i w_i = -\mathrm{i}\eta^2 \partial_{\bar{z}_i} \hat{G}(w, \bar{w}) - \mathrm{i}\eta^4 \partial_{\bar{z}_i} \widetilde{K}(I_0 + J, \phi_0 + \tilde{\omega} t + \psi, w, \bar{w}), \quad i \in \mathcal{I}^c. \end{cases} \tag{78}$$

We will look for $\zeta(t)$ as a T-periodic curve taking values in the covering space $\mathbb{R}^N \times \mathbb{R}^N \times \ell^{a,s}$ (that for simplicity we will still denote by $\mathcal{P}_{a,s}$) with $\int_0^T \psi(t) dt = 0$. For

$\zeta = (J, \psi, w) \in \mathbb{R}^N \times \mathbb{R}^N \times \ell^{a,s}$ we define the norm[5]

$$\|\zeta\|_{\mathcal{P}_{a,s}} = \|(J, \psi, w)\|_{\mathcal{P}_{a,s}} := |J| + |\psi| + \|w\|_{a,s}.$$

With this norm, $\mathcal{P}_{a,s}$ is a Banach algebra, recalling that $s > 1$ and Lemma 2.2.

In particular we will look for H^k-solutions $\zeta(t)$ in the Banach space

$$\overline{H}^k_{T,a,s} := \left\{ \zeta \in H^k_{T,a,s}, \quad \int_0^T \psi(t)dt = 0 \right\},$$

where $k \in \mathbb{N}$, $T > 0$,

$$H^k_{T,a,s} := \left\{ \zeta \in H^k(\mathbb{R}, \mathcal{P}_{a,s}), \quad \zeta(t+T) = \zeta(t) \right\},$$

and $H^k(\mathbb{R}, \mathcal{P}_{a,s})$ is the Sobolev space of the functions $\zeta : \mathbb{R} \to \mathcal{P}_{a,s}$ with k weak derivatives in $L^2(\mathbb{R}, \mathcal{P}_{a,s})$ (for $k = 0$, $H^0(\mathbb{R}, \mathcal{P}_{a,s}) = L^2(\mathbb{R}, \mathcal{P}_{a,s})$).

The space $H^k_{T,a,s}$ is endowed with the norm

$$\|\zeta\|_{H^k_{T,a,s}} := \sum_{h=0}^{k} T^h \|\partial_t^h \zeta\|_{T,a,s},$$

where

$$\begin{aligned}
\|\zeta\|_{T,a,s} &:= |J|_{L^2,T} + |\psi|_{L^2,T} + \|w\|_{L^2,T,a,s} + \|\bar{w}\|_{L^2,T,a,s}, \\
|J|^2_{L^2,T} &:= \frac{1}{T}\int_0^T |J(t)|^2 dt, \qquad |\psi|^2_{L^2,T} := \frac{1}{T}\int_0^T |\psi(t)|^2 dt, \\
\|w\|^2_{L^2,T,a,s} &:= \frac{1}{T}\int_0^T \|w(t)\|^2_{a,s} dt.
\end{aligned}$$

Note that $H^k_{T,a,s} = \overline{H}^k_{T,a,s} \oplus \mathbb{R}^N$.

Remark 3.5. With the above definitions the following result holds:

$$\|\zeta(t)\|_{\mathcal{P}_{a,s}} \leq \|\zeta\|_{H^1_{T,a,s}} \qquad \forall\, t \in \mathbb{R}. \tag{79}$$

Hence the spaces $H^k_{T,a,s}$, for $k \geq 1$, are Banach algebras, and the $H^k_{T,a,s}$-norm of the product of any component of a vector ζ with any component of a vector ζ' is bounded by $\|\zeta\|_{H^k_{T,a,s}} \|\zeta'\|_{H^k_{T,a,s}}$. We will consider system (78) as a functional equation in $H^k_{T,a,s}$.

To simplify notation we rewrite (78) in the form

$$L\zeta = N(\zeta; \phi_0), \tag{80}$$

where L is the linear operator

$$L\zeta = L(J, \psi, w) := (\dot{\psi} - \eta^2 AJ, \dot{J}, \dot{w}_i + \mathrm{i}\tilde{\Omega}_i w_i) \tag{81}$$

and N is the nonlinearity

$$N(\zeta; \phi_0) := \big(N_I(\zeta; \phi_0), N_\phi(\zeta; \phi_0), N_{\hat{z}}(\zeta; \phi_0)\big) \tag{82}$$

[5]Here $|\cdot|$ is the standard Euclidian norm on $\mathbb{R}^N$, $\|w\|^2_{a,s} = \sum_{i \in \mathcal{I}^c} |w_i|^2 i^{2s} e^{2ai}$.

defined by

$$
\begin{aligned}
(83)\quad N_I &:= \eta^2 B^t w\bar{w} + \eta^4 \partial_I \tilde{K}(I_0 + J, \phi_0 + \tilde{\omega}t + \psi, w),\\
N_\phi &:= -\eta^4 \partial_\phi \tilde{K}(I_0 + J, \phi_0 + \tilde{\omega}t + \psi, w),\\
(N_{z_i}) &:= -\mathrm{i}\eta^2 (BJ)_i w_i - \mathrm{i}\eta^2 \partial_{\bar{z}_i} \hat{G}(w) - \mathrm{i}\eta^4 \partial_{\bar{z}_i} \tilde{K}(I_0 + J, \phi_0 + \tilde{\omega}t + \psi, w), \quad i \in \mathcal{I}^c.
\end{aligned}
$$

We note that by (19) and Remark 3.5, we get that for all $\phi_0 \in \mathbb{T}^N$

$$
N(\cdot; \phi_0) \in C^\infty\big(H^k_{T,a,s}, H^k_{T,a,s+1}\big) \qquad \forall\, k \geq 1. \tag{84}
$$

Since A is invertible and the nonresonance condition (49) holds by Proposition 3.2, the kernel of the linear operator L is

$$
\mathcal{K} = \big\{\zeta(t) = (J(t), \psi(t), w(t)) \ \text{ s.t. } \ \psi(t) \equiv \text{const}, \ J(t) \equiv 0, w(t) \equiv 0\big\}.
$$

On the other hand, the range of L is composed by the curves $\tilde{\zeta}(t) := (\tilde{J}(t), \tilde{\psi}(t), \tilde{w}(t))$ with $\int_0^T \tilde{\psi} = 0$, as we will show in the next subsection concerning the inversion of L.

3.3. Inversion of the linear operator. Recall that $\tau = T/(2\pi) = \eta^{-2}$. By the theory of the symmetric operators, and since A is invertible, it possesses an orthonormal basis of eigenvectors $e^{(1)}, \dots, e^{(N)} \in \mathbb{R}^N$ with respective eigenvalues $\nu_1, \dots, \nu_N \in \mathbb{R} \setminus \{0\}$. In these coordinates we can write

$$
\begin{aligned}
\tilde{J}(t) &= \sum_{j=1}^N \tilde{J}^{(j)}(t) e^{(j)} = \sum_{j=1}^N e^{(j)} \sum_{\ell \in \mathbb{Z}} \tilde{J}^{(j)}_\ell \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right),\\
\tilde{\psi}(t) &= \sum_{j=1}^N \tilde{\psi}^{(j)}(t) e^{(j)} = \sum_{j=1}^N e^{(j)} \sum_{\ell \in \mathbb{Z}} \tilde{\psi}^{(j)}_\ell \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right),\\
\tilde{w}_i(t) &= \sum_{\ell \in \mathbb{Z}} \tilde{w}_{i\ell} \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right), \qquad\qquad i \in \mathcal{I}^c.
\end{aligned}
$$

We define the linear operator

$$
\mathcal{L}\tilde{\zeta} = \mathcal{L}(\tilde{J}, \tilde{\psi}, \tilde{w}) := (J, \psi, w) = \zeta \tag{85}
$$

in the following way:

$$
J(t) := \tau \sum_{j=1}^N e^{(j)} \left(-\frac{1}{\nu_j} \tilde{J}^{(j)}_0 + \sum_{\ell \neq 0} \frac{1}{\mathrm{i}\ell} \tilde{\psi}^{(j)}_\ell \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right)\right), \tag{86}
$$

$$
\psi(t) := \tau \sum_{j=1}^N e^{(j)} \sum_{\ell \neq 0} \frac{1}{\mathrm{i}\ell} \left(\frac{\nu_j}{\mathrm{i}\ell} \tilde{\psi}^{(j)}_\ell + \tilde{J}^{(j)}_\ell \right) \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right), \tag{87}
$$

$$
w_i(t) := \tau \sum_{\ell \in \mathbb{Z}} \frac{1}{\mathrm{i}(\ell - \tilde{\Omega}_i \tau)} \tilde{w}_{i\ell} \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right), \qquad i \in \mathcal{I}^c. \tag{88}
$$

The following proposition not only states that $\mathcal{L}$ is the inverse of L but also gives an upper bound of its norm. However, because of the small divisors $\ell - \tilde{\Omega}_i \tau$, appearing in (88), it turns out that $\mathcal{L}$ "loses one spatial derivative." On the other hand, $\mathcal{L}$ also

"gains one time derivative" if one gives up "two spatial derivatives." Estimate (61) will be crucial.

PROPOSITION 3.6. *Suppose $\tau \in \mathcal{T}_\mu$ and μ small enough. Take $k \in \mathbb{N}^+$. If $\tilde\zeta \in \overline{H}^k_{T,a,s+1}$, then $\zeta = \mathcal{L}\tilde\zeta \in \overline{H}^k_{T,a,s} \cap \overline{H}^{k+1}_{T,a,s-1}$ and*

$$\|\zeta\|_{H^k_{T,a,s}} + \|\zeta\|_{H^{k+1}_{T,a,s-1}} \le c_6^{-1}\,\tau\,\|\tilde\zeta\|_{H^k_{T,a,s+1}}. \tag{89}$$

Moreover,

$$L\mathcal{L}\tilde\zeta = L\zeta = \tilde\zeta. \tag{90}$$

Proof. First we note that the average of ψ is zero as it follows by (87). We now prove (89). Since

$$\frac{1}{T}\int_0^T \left|\sum_{\ell\in\mathbb{Z}} a_\ell \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right)\right|^2 dt = \sum_{\ell\in\mathbb{Z}} |a_\ell|^2 \tag{91}$$

and ($\{e^{(j)}\}_{1\le j\le N}$ being an orthonormal basis) $\left|\sum_{j=1}^N b_j e^{(j)}\right|^2 = \sum_{j=1}^N |b_j|^2$, we have that

$$f(t) = \sum_{j=1}^N e^{(j)} \sum_{\ell\in\mathbb{Z}} f_\ell^{(j)} \exp\left(\frac{\mathrm{i}\ell t}{\tau}\right) \Longrightarrow |f|^2_{L^2,T} = \frac{1}{T}\int_0^T |f(t)|^2 dt = \sum_{j=1}^N \sum_{\ell\in\mathbb{Z}} |f_\ell^{(j)}|^2.$$

Hence, if $\tilde C := 2\max_{1\le j\le N}\{|\nu_j|^2, 1/|\nu_j|^2\}$, from (87) we get

$$|\psi|^2_{L^2,T} \le \tilde C\tau^2 \sum_{j=1}^N \sum_{\ell\ne 0} \ell^{-2}\left(|\tilde\psi_\ell^{(j)}|^2 + |\tilde J_\ell^{(j)}|^2\right) \le \tilde C\tau^2\left(|\tilde\psi|^2_{L^2,T} + |\tilde J|^2_{L^2,T}\right), \tag{92}$$

$$\begin{aligned} |\partial_t^h\psi|^2_{L^2,T} &\le \tilde C\tau^{2(1-h)} \sum_{j=1}^N \sum_{\ell\ne 0} \ell^{2(h-1)}\left(|\tilde\psi_\ell^{(j)}|^2 + |\tilde J_\ell^{(j)}|^2\right) \\ &\le \tilde C\tau^2\left(|\partial_t^{h-1}\tilde\psi|^2_{L^2,T} + |\partial_t^{h-1}\tilde J|^2_{L^2,T}\right) \quad \text{for} \quad h \ge 1. \end{aligned} \tag{93}$$

Similar estimates hold for J defined in (86), namely,

$$|J|^2_{L^2,T} \le \tilde C\tau^2\left(\sum_{j=1}^N |\tilde J_0^{(j)}|^2 + \sum_{j=1}^N \sum_{\ell\ne 0} \ell^{-2}|\tilde\psi_\ell^{(j)}|^2\right) \le \tilde C\tau^2\left(|\tilde J|^2_{L^2,T} + |\tilde\psi|^2_{L^2,T}\right), \tag{94}$$

$$|\partial_t^h J|^2_{L^2,T} = |\partial_t^{h-1}\tilde\psi|^2_{L^2,T} \qquad \text{for} \quad h \ge 1. \tag{95}$$

We now go on to estimate w defined in (88) in which the small divisors $\ell - \tilde\Omega_j\tau$ appear. By (91) we have that if $w(t) = (w_i(t))_{i\in\mathcal{I}^c}$ with $w_i(t) = \sum_{\ell\in\mathbb{Z}} w_{i\ell}\exp(\mathrm{i}\ell t/\tau)$, then

$$\begin{aligned} \|w\|^2_{L^2,T,a,s} &= \frac{1}{T}\int_0^T \|w(t)\|^2_{a,s} = \frac{1}{T}\int_0^T \sum_{i\in\mathcal{I}^c} i^{2s}e^{2ai}|w_i(t)|^2 dt \\ &= \sum_{i\in\mathcal{I}^c} i^{2s}e^{2ai}\frac{1}{T}\int_0^T |w_i(t)|^2 dt = \sum_{i\in\mathcal{I}^c} i^{2s}e^{2ai}\sum_{\ell\in\mathbb{Z}} |w_{i\ell}|^2. \end{aligned}$$

Hence, recalling (88), we get

$$\begin{aligned}\|\partial_t^h w\|^2_{L^2,T,a,s} &= \tau^2 \sum_{i\in\mathcal{I}^c} i^{2s} e^{2ai} \sum_{\ell\in\mathbb{Z}} \frac{|\ell|^{2h}}{\tau^{2h}} \frac{|\tilde{w}_{i\ell}|^2}{|\ell-\tilde{\Omega}_i\tau|^2} \\ &\le \frac{\tau^2}{c_5^2} \sum_{i\in\mathcal{I}^c} i^{2(s+1)} e^{2ai} \sum_{\ell\in\mathbb{Z}} \frac{|\ell|^{2h}}{\tau^{2h}} |\tilde{w}_{i\ell}|^2 = \frac{\tau^2}{c_5^2} \|\partial_t^h \tilde{w}\|^2_{L^2,T,a,s+1} \end{aligned} \tag{96}$$

by the crucial estimate (61). Moreover, we claim that by (61)

$$\frac{|\ell|}{|\ell-\tilde{\Omega}_i\tau|} \le \frac{4\tau i^2}{c_5}. \tag{97}$$

To prove (97) we distinguish two cases, $\ell \le 2\tilde{\Omega}_i\tau$ and $\ell > 2\tilde{\Omega}_i\tau$. In the first case we have by (61)

$$\frac{|\ell|}{|\ell-\tilde{\Omega}_i\tau|} \le \frac{|\ell| i}{c_5} \le \frac{2\tilde{\Omega}_i\tau i}{c_5} \le \frac{4\tau i^2}{c_5}$$

since $\tilde{\Omega}_i \le 2i$. On the other hand, if $\ell > 2\tilde{\Omega}_i\tau$, we have $|\ell-\tilde{\Omega}_i\tau| \ge |\ell|/2$, which implies

$$\frac{|\ell|}{|\ell-\tilde{\Omega}_i\tau|} \le 2 \le \frac{4\tau i^2}{c_5},$$

and (97) follows. Using (97) we get

$$\begin{aligned}\|\partial_t^h w\|^2_{L^2,T,a,s-1} &= \tau^2 \sum_{i\in\mathcal{I}^c} i^{2(s-1)} e^{2ai} \sum_{\ell\in\mathbb{Z}} \frac{|\ell|^{2h}}{\tau^{2h}} \frac{|\tilde{w}_{i\ell}|^2}{|\ell-\tilde{\Omega}_i\tau|^2} \\ &\le \frac{16\tau^2}{c_5^2} \sum_{i\in\mathcal{I}^c} i^{2(s+1)} e^{2ai} \sum_{\ell\in\mathbb{Z}} \frac{|\ell|^{2(h-1)}}{\tau^{2(h-1)}} |\tilde{w}_{i\ell}|^2 \\ &= \frac{16\tau^2}{c_5^2} \|\partial_t^{h-1} \tilde{w}\|^2_{L^2,T,a,s+1}. \end{aligned} \tag{98}$$

Therefore (89) follows from (92)–(96) and (98).

Finally we note that (90) directly follows from the definition of $\mathcal{L}$ given in (85)–(88).

We also remark that the constant c_6 does not depend on k, a, s. □

3.4. Lyapunov–Schmidt reduction. From the previous section it results that the kernel $\mathcal{K}$ and the range $\mathcal{R}$ of the linear operator L are $\{\tilde{\psi} \equiv \text{const}\}$ and $\{\int_0^T \tilde{\psi} = 0\}$, respectively. For $\tilde{\zeta} = (\tilde{J}, \tilde{\psi}, \tilde{w})$ let us define the projections

$$\Pi_{\mathcal{K}}\tilde{\zeta} := \big(0, \langle\tilde{\psi}\rangle, 0\big), \qquad \Pi_{\mathcal{R}}\tilde{\zeta} := \big(\tilde{J}, \tilde{\psi} - \langle\tilde{\psi}\rangle, \tilde{w}\big),$$

where $\langle\tilde{\psi}\rangle := \int_0^T \tilde{\psi}$. In such a way the equation $L\zeta = N(\zeta;\phi_0)$ decomposes into the equation on the kernel,

$$0 = \Pi_{\mathcal{K}} N(\zeta;\phi_0), \quad \text{namely,} \quad \big\langle N_\phi(\zeta,\phi_0)\big\rangle = 0, \tag{99}$$

and the one on the range,

$$L\zeta = \Pi_{\mathcal{R}} N(\zeta;\phi_0), \tag{100}$$

respectively. The idea is to solve first the range equation for any fixed ϕ_0, finding a solution $\zeta(t) = \zeta_{\phi_0}(t)$ by the contraction mapping theorem, and thereafter the kernel equation (99) for $\zeta = \zeta_{\phi_0}$, namely the *finite dimensional* equation

$$\big\langle N_\phi(\zeta_{\phi_0}; \phi_0)\big\rangle = 0, \tag{101}$$

determining ϕ_0 by a variational argument.

3.5. Range equation. We rewrite the range equation (100) in a fixed-point form:

$$\zeta = \Phi(\zeta; \phi_0)$$

with

$$\Phi(\zeta; \phi_0) := \mathcal{L}\Pi_{\mathcal{R}} N(\zeta; \phi_0).$$

By Proposition 3.6, the operator $\mathcal{L}$ "loses one derivative," but, by the smoothing property (84), the nonlinearity N gains exactly *one* derivative. In particular, we have that, for any $\phi_0 \in \mathbb{T}^N$ fixed,

$$\Phi(\cdot; \phi_0) \in C^\infty\big(\overline{H}^k_{T,a,s}, \overline{H}^k_{T,a,s}\big) \qquad \forall\, k \geq 1. \tag{102}$$

In the following lemma we prove that Φ is a contraction on a suitable closed ball of $\overline{H}^1_{T,a,s}$.

LEMMA 3.7. *Suppose $\tau \in \mathcal{T}_\mu$ and μ small enough. For any $\phi_0 \in \mathbb{T}^N$ the map $\Phi(\cdot; \phi_0)$ is a contraction on the closed ball of radius $\rho := c_7^{-1}\mu$ of $\overline{H}^1_{T,a,s}$.*

Proof. Let $\zeta, \zeta', h \in \overline{H}^1_{T,a,s}$ and $\|\zeta\|_{H^1_{T,a,s}}, \|\zeta'\|_{H^1_{T,a,s}} \leq \rho$. From (82), (83), (84), and (32), and recalling Proposition 2.4 and Remark 3.5, we get the following estimates on the nonlinearity:

$$\|N(\zeta)\|_{H^1_{T,a,s+1}} \leq c_8^{-1}\left(\eta^2\rho^2 + \frac{\eta^4}{\mu}\right), \tag{103}$$

$$\|DN(\zeta)[h]\|_{H^1_{T,a,s+1}} \leq c_8^{-1}\left(\eta^2\rho + \frac{\eta^4}{\mu}\right)\|h\|_{H^1_{T,a,s}}. \tag{104}$$

Using Proposition 3.6 and (103), we obtain

$$\|\Phi(\zeta)\|_{H^1_{T,a,s}} \leq (c_6c_8)^{-1}\left(\rho^2 + \frac{\eta^2}{\mu}\right) = (c_6c_8)^{-1}\rho^2 + \frac{\rho}{2} \leq \rho,$$

taking $c_7 := c_4c_6c_8/4$ and μ small enough. Hence Φ maps the ball in itself. Nothing remains but to show that Φ is a contraction. By (104) we get

$$\|N(\zeta) - N(\zeta')\|_{H^1_{T,a,s+1}} \leq c_8^{-1}\left(\eta^2\rho + \frac{\eta^4}{\mu}\right)\|\zeta - \zeta'\|_{H^1_{T,a,s}},$$

and by Proposition 3.6 we have

$$\|\Phi(\zeta) - \Phi(\zeta')\|_{H^1_{T,a,s}} \leq (c_6c_8)^{-1}\left(\rho + \frac{\eta^2}{\mu}\right)\|\zeta - \zeta'\|_{H^1_{T,a,s}}.$$

Since, for μ small enough,

$$(c_6c_8)^{-1}\left(\rho+\frac{\eta^2}{\mu}\right)\leq\left(\frac{1}{c_6c_8}+\frac{1}{2}\right)\frac{\mu}{c_7}<1,$$

Φ is a contraction. □

By the contraction mapping theorem and noting that the dependence of the nonlinearity N, and therefore of Φ, on the parameter ϕ_0 is smooth, we conclude that there exists a smooth function $\mathbb{T}^N\ni\phi_0\mapsto\zeta_{\phi_0}\in\overline{H}^1_{T,a,s}$ solving $\zeta_{\phi_0}=\Phi(\zeta_{\phi_0};\phi_0)$. By (90), ζ_{ϕ_0} also solves the range equation (100), as the following corollary states.

COROLLARY 3.8. *Suppose $\tau\in\mathcal{T}_\mu$ and μ small enough. Then there exists a smooth function $\mathbb{T}^N\ni\phi_0\mapsto\zeta_{\phi_0}\in\overline{H}^1_{T,a,s}$ solving* (100) *and satisfying*

$$\|\zeta_{\phi_0}\|_{H^1_{T,a,s}}\leq\frac{\mu}{c_7}.$$

Remark 3.9. In [BaB], instead of (49) the authors imposed the weaker "diophantine-type" condition $|\ell-\tilde{\Omega}_i\tau|\geq\operatorname{const}i^{-\sigma}$, $\sigma>1$, on the small divisors. Then the operator $\mathcal{L}$ "loses σ derivatives." If the nonlinearity N is smoothing of order $d>1$, namely it "gains d derivatives," taking $1<\sigma<d$, the contraction mapping theorem can still be used in solving the range equation for almost every rescaled period τ. In particular, $d=2$ for the beam equation and $d>1$ for the NLS. Since we have exactly $d=1$ for the wave equation, in order to have a positive measure set of rescaled periods, a KAM analysis is necessary (see Remark 3.12).

3.6. Kernel equation. Once we have solved the range equation (100) by finding the smooth function

$$\mathbb{T}^N\ni\phi_0\mapsto\zeta_{\phi_0}=:(J_{\phi_0},\psi_{\phi_0},w_{\phi_0})\in\overline{H}^1_{T,a,s},$$

we still have to solve the "reduced" kernel equation (101). As the solutions of the Hamiltonian equations (75) are critical points of the action functional S defined in (76), so the solutions of the "reduced" kernel equation (101) are critical points of the *reduced action functional*

$$\mathcal{S}(\phi_0):=S(I_{\phi_0},\phi_{\phi_0},\hat{z}_{\phi_0})=\int_0^T\left(I_{\phi_0}\dot{\phi}_{\phi_0}-\mathrm{i}\hat{z}_{\phi_0}\dot{\bar{\hat{z}}}_{\phi_0}-\widetilde{\mathcal{H}}(\phi_{\phi_0},I_{\phi_0},\hat{z}_{\phi_0},\bar{\hat{z}}_{\phi_0})\right)dt, \tag{105}$$

where

$$I_{\phi_0}(t):=I_0+J_{\phi_0}(t),\qquad\phi_{\phi_0}(t):=\phi_0+\tilde{\omega}t+\psi_{\phi_0}(t),\qquad\hat{z}_{\phi_0}(t):=w_{\phi_0}(t). \tag{106}$$

Actually we are claiming that

$$\langle N_\phi(\zeta_{\phi_0};\phi_0)\rangle=0$$

or equivalently

$$\partial_{\phi_0}\mathcal{S}(\phi_0)=0. \tag{107}$$

Indeed, (107) is a corollary of (36), (83), and the following.

LEMMA 3.10. *The reduced action functional satisfies*

$$\partial_{\phi_0}\mathcal{S}(\phi_0)=-T\langle\partial_\phi\widetilde{\mathcal{H}}(I_{\phi_0},\phi_{\phi_0},\hat{z}_{\phi_0},\bar{\hat{z}}_{\phi_0})\rangle.$$

Proof. We have

$$\begin{aligned}\partial_{\phi_0}\mathcal{S}(\phi_0) &= \int_0^T \Big[\big(\dot\phi_{\phi_0} - \partial_I\widetilde{\mathcal{H}}\big)\partial_{\phi_0} I_{\phi_0} + I_{\phi_0}\partial_{\phi_0}\dot\phi_{\phi_0} - \partial_\phi\widetilde{\mathcal{H}}\ \partial_{\phi_0}\phi_{\phi_0} \\ &\qquad -\big(\mathrm{i}\dot{\bar w}_{\phi_0} + \partial_{\bar z}\widetilde{\mathcal{H}}\big)\partial_{\phi_0} w_{\phi_0} - \mathrm{i} w_{\phi_0}\partial_{\phi_0}\dot{\bar w}_{\phi_0} - \partial_{\bar z}\widetilde{\mathcal{H}}\ \partial_{\phi_0}\bar w_{\phi_0}\Big]\,dt \\ &= I_{\phi_0}\partial_{\phi_0}\phi_{\phi_0}\Big|_0^T - \mathrm{i} w_{\phi_0}\partial_{\phi_0}\bar w_{\phi_0}\Big|_0^T - \int_0^T \big\langle\partial_\phi\widetilde{\mathcal{H}}\big\rangle\partial_{\phi_0}\phi_{\phi_0}\ dt\end{aligned}$$

by an integration by parts and since ζ_{ϕ_0} satisfies the range equation (100). Moreover, since ζ_{ϕ_0} is periodic and $\int_0^T \psi_{\phi_0} = 0$, we get

$$\begin{aligned} I_{\phi_0}\partial_{\phi_0}\phi_{\phi_0}\Big|_0^T &= I_{\phi_0}(0)\partial_{\phi_0}\big(\phi_{\phi_0}(T) - \phi_{\phi_0}(0)\big) = I_{\phi_0}\partial_{\phi_0}\tilde\omega T = 0, \\ w_{\phi_0}\partial_{\phi_0}\bar w_{\phi_0}\Big|_0^T &= w_{\phi_0}(0)\partial_{\phi_0}\big(\bar w_{\phi_0}(T) - \bar w_{\phi_0}(0)\big) = 0 \\ \int_0^T \phi_{\phi_0}(t)\,dt &= \int_0^T \big(\phi_0 + \tilde\omega t\big)\,dt + \int_0^T \psi_{\phi_0}(t)\,dt = \phi_0 T + \tilde\omega\frac{T^2}{2}.\end{aligned}$$

Finally

$$\partial_{\phi_0}\mathcal{S}(\phi_0) = -\int_0^T \big\langle\partial_\phi\widetilde{\mathcal{H}}\big\rangle\partial_{\phi_0}\phi_{\phi_0}\ dt = -\big\langle\partial_\phi\widetilde{\mathcal{H}}\big\rangle\partial_{\phi_0}\int_0^T \phi_{\phi_0}(t)\,dt = -\big\langle\partial_\phi\widetilde{\mathcal{H}}\big\rangle T. \qquad \square$$

3.7. Existence. By (107) every critical point $\phi_0 \in \mathbb{T}^N$ of the reduced action functional $\mathcal{S}$ defined in (105) solves the "reduced" kernel equation (101) and, therefore, the curve $\big(I_{\phi_0}(t), \phi_{\phi_0}(t), \hat z_{\phi_0}(t)\big)$ defined in (106) is a solution of the Hamiltonian equations (75). In particular we expect the existence of at least N geometrically distinct T-periodic solutions, namely solutions not obtained from each other simply by time translations.

Indeed, let us consider the restriction of $\mathcal{S}$ to the plane $E := [\tilde\omega]^\perp$. The set $\mathbb{Z}^N \cap E$ is a lattice of E; hence $\mathcal{S}$ can be defined on the quotient space $\Gamma := E/(\mathbb{Z}^N \cap E) \sim \mathbb{T}^{N-1}$. Due to the invariance of $\mathcal{S}$ with respect to the time shift, any critical point of $\mathcal{S}_{|\Gamma}\colon \Gamma \to \mathbb{R}$ is also a critical point of $\mathcal{S}\colon \mathbb{T}^N \to \mathbb{R}$. By the Lusternik–Schnirelman category theory (see, for example, [A]), since $\operatorname{cat}\Gamma = \operatorname{cat}\mathbb{T}^{N-1} = N$, we can define the N min-max critical values $\mathrm{c}_1 \le \mathrm{c}_2 \le \cdots \le \mathrm{c}_N$ for the reduced action functional $\mathcal{S}_{|\Gamma}$. If the critical levels c_i are all distinct, the corresponding T-periodic solutions are surely geometrically distinct, since their actions c_i are different. On the other hand, if some min-max critical levels coincide, then $\mathcal{S}_{|\Gamma}$ possesses infinitely many critical points. Although not all the corresponding T-periodic solutions are necessarily geometrically distinct, since a periodic solution can cross Γ at most a finite number of times, the existence of infinitely many geometrically distinct orbits follows (see [BBiV] for further details).

PROPOSITION 3.11. *Suppose $\tau \in \mathcal{T}_\mu$ and μ small enough. Then system* (75) *possesses (at least) N geometrically distinct T-periodic solutions*

(108)
$$\big(I_{\phi_0^{(j)}}(t), \phi_{\phi_0^{(j)}}(t), \hat z_{\phi_0^{(j)}}(t)\big) = \big(I_0(\tau), \phi_0^{(j)} + \tilde\omega t, 0\big) + \zeta_{\phi_0^{(j)}}(t), \qquad \zeta_{\phi_0^{(j)}} \in \overline{H}^1_{T,a,s},$$

parametrized by suitable $\phi_0^{(j)} \in \mathbb{T}^N$, $1 \le j \le N$. Moreover,

$$\|\zeta_{\phi_0^{(j)}}\|_{H^1_{T,a,s}} \le \frac{\mu}{c_7} \qquad \forall\, 1 \le j \le N. \tag{109}$$

Remark 3.12. In (49) we have imposed a strong condition on the small divisors in order to use the standard contraction mapping theorem in solving the range equation. The other side of such a condition is that we are able to consider only a finite number of periods. The natural way to deal with the small divisors problem (49), in order to obtain a positive measure set of periods, should be via a KAM analysis. The range equation should be solved by a Nash–Moser implicit function theorem. Thereafter, one should prove that, for any fixed value of the perturbative parameter η, the bifurcation equation $0 = \Pi_{\mathcal{K}} N(\zeta_{\phi_0}; \phi_0)$ has a solution for ϕ_0 belonging to a suitable η-dependent Cantor set $\mathcal{C}_\eta$ (see [BBo06] and [GMPr05], where the extension of Weinstein's theorem [We73] for completely resonant wave equation is considered). A standard way to proceed is to develop the reduced action functional in powers of the perturbative parameter and to prove that the first nontrivial term in the development has a nondegenerate critical point. Such nondegeneracy is used to show that, for any η in a suitable positive measure set, there exists a critical point $\phi_0 = \phi_0(\eta)$ of the whole reduced action functional, belonging to $\mathcal{C}_\eta$. However, in the present case, while the perturbative parameter η goes to zero, the period T goes to infinity. Then, in the computation of $\mathcal{S}$ in (105), one has to average over an *infinite* time. As a consequence, all the low order terms in the development of $\mathcal{S}$ vanish; the first nontrivial terms appear only at a very high order in η, and it is difficult to deal with them.

3.8. Regularity. The solutions $(I_0(\tau), \phi_0^{(j)} + \tilde{\omega} t, 0) + \zeta_{\phi_0^{(j)}}(t)$ of system (75) described in (108) belong to $H^1_{T,a,s}$ and verify estimate (109). With the same procedure, for any fixed $k \geq 1$, we can also find solutions in $H^k_{T,a,s}$ verifying

$$\|\zeta_{\phi_0^{(j)}}\|_{H^k_{T,a,s}} \leq \frac{\mu}{c_7}. \tag{110}$$

Indeed, we can solve the range equation (100) in $\overline{H}^k_{T,a,s}$, adapting Lemma 3.7. However, in this case, the constant c_7 *depends on* k.

On the other hand, giving up the $H^k_{T,a,s}$-estimate in (110), it is possible to prove, by a bootstrap argument, that the solutions (108) of Proposition 3.11 actually belong to $H^k_{T,a,s}$ for any $k \geq 1$. In particular we show that ζ_{ϕ_0} of Corollary 3.8 belongs to $\overline{H}^k_{T,a,s}$.

Indeed, $\zeta_{\phi_0} \in \overline{H}^1_{T,a,s}$ solves the fixed point equation $\zeta_{\phi_0} = \Phi(\zeta_{\phi_0}; \phi_0) = \mathcal{L}\Pi_{\mathcal{R}} N(\zeta_{\phi_0}; \phi_0)$. Since $N(\zeta_{\phi_0}; \phi_0) \in H^1_{T,a,s+1}$ by (84), then $\zeta_{\phi_0} \in \overline{H}^2_{T,a,s-1}$ by Proposition 3.6. Noting that $\overline{H}^k_{T,a,s-1} \subset \overline{H}^k_{T,\tilde{a},s}$ for any $k \geq 1$ and $0 < \tilde{a} < a$, we have $\zeta_{\phi_0} \in \overline{H}^2_{T,\tilde{a},s}$. Defining $a_k := a\left(\frac{1}{2} + \frac{1}{2^k}\right)$, we prove that $\zeta_{\phi_0} \in \overline{H}^k_{T,a_k,s}$ for any $k \geq 1$ and finally, $\zeta_{\phi_0} \in \overline{H}^k_{T,a/2,s}$ for any $k \geq 1$. However, in this fashion, the $H^k_{T,a,s}$-estimates deteriorate while k increases.

Summarizing, by the Sobolev immersions we have that $\zeta_{\phi_0} \in C^k(\mathbb{R}, \mathcal{P}_{a/2,s})$ for any $k \geq 1$. We have shown the following.

COROLLARY 3.13. *The solutions* (108) *of* (75) *belong to* $C^\infty(\mathbb{R}, \mathcal{P}_{a/2,s})$.

3.9. Minimal period.

LEMMA 3.14. *Let* $h, k \in \mathbb{N}^+$, $h < k$. *We have*

$$[\omega_k \tau - k\tau] \leq [\omega_h \tau - h\tau] < \frac{\mu\tau}{2h}. \tag{111}$$

Proof. We first proof that

$$\omega_k - k < \omega_h - h, \tag{112}$$

which implies $\omega_k\tau - k\tau < \omega_h\tau - h\tau$ and the first inequality in (111). Dividing by k, (112) is equivalent to

$$f(x) := \sqrt{x^2 + \mu/k^2} - x - \sqrt{1 + \mu/k^2} + 1 > 0 \qquad \text{for} \quad 0 < x < 1,$$

where $x := h/k$. Since $f(1) = 0$ and $f'(x) = x(x^2 + \mu/k^2)^{-1/2} - 1 < 0$ for $0 < x < 1$, we get $f(x) > 0$, and (112) follows. Then the second inequality in (111) directly follows from (65). ∎

LEMMA 3.15. *Let $T^{\min}$ be the minimal period of a T-periodic orbit* (108) *of Proposition* 3.11. *If $N \geq 2$, then*

$$T^{\min} \geq \frac{c_9}{\mu}. \tag{113}$$

Proof. Let $\big(I(t), \phi(t), \hat{z}(t)\big)$ be a T-periodic solution of Proposition 3.11. We know that $\phi(T) - \phi(0) = 2\pi k$ with $k \in \mathbb{Z}^N$ defined in (55). Denoting by $T_\phi^{\min} \leq T^{\min}$ the minimal period of $\phi(t)$, we have that there exist $n \in \mathbb{N}^+$ such that $nT_\phi^{\min} = T$ and $\tilde{k} \in \mathbb{Z}^N$ such that $\phi(T_\phi^{\min}) - \phi(0) = 2\pi\tilde{k}$, verifying $n\tilde{k} = k$. Hence we deduce that n divides $g := \gcd(k_{i_1}, \dots, k_{i_N})$, and we get that

$$T^{\min} \geq T_\phi^{\min} = \frac{T}{n} \geq \frac{T}{g}. \tag{114}$$

We claim that

$$\tilde{g} := \gcd(k_{i_1}, k_{i_2}) < \frac{\mu\tau i_2}{i_1}. \tag{115}$$

Then the lemma follows by (114) and (115), noting that $\tilde{g} \geq g$ and recalling that $T = 2\pi\tau$ (with $c_9 = 2\pi i_1/i_2$).

We now prove (115). Since $\tau \in \mathbb{N}$ we have $k_i = [\omega_i\tau] + \kappa_i = i\tau + [\omega_i\tau - i\tau] + \kappa_i$ for all $i \in \mathcal{I}$ and

$$\begin{aligned} i_2k_{i_1} - i_1k_{i_2} &= i_2[\omega_{i_1}\tau - i_1\tau] - i_1[\omega_{i_2}\tau - i_2\tau] + i_2\kappa_{i_1} - i_1\kappa_{i_2} \\ &\geq (i_2 - i_1)[\omega_{i_1}\tau - i_1\tau] + \left(\frac{i_2}{i_1} - \frac{i_1}{i_2}\right)\tilde{\kappa} > 0 \end{aligned} \tag{116}$$

by (111) and recalling (52). Moreover, since $\tilde{g} = \gcd(k_{i_1}, k_{i_2})$ there exist $h_1, h_2 \in \mathbb{N}$ such that $k_{i_1} = h_1\tilde{g}$ and $k_{i_2} = h_2\tilde{g}$. From (116) we have that $i_2k_{i_1} - i_1k_{i_2} > 0$, and therefore

$$i_2k_{i_1} - i_1k_{i_2} = \big(i_2h_1 - i_1h_2\big)\tilde{g} \geq \tilde{g}. \tag{117}$$

Finally, by (111)

$$\begin{aligned} i_2k_{i_1} - i_1k_{i_2} &= i_2[\omega_{i_1}\tau - i_1\tau] - i_1[\omega_{i_2}\tau - i_2\tau] + i_2\kappa_{i_1} - i_1\kappa_{i_2} \\ &< i_2[\omega_{i_1}\tau - i_1\tau] + \frac{i_2}{i_1}\tilde{\kappa} \leq \frac{i_2}{i_1}\left(\frac{\mu\tau}{2} + \tilde{\kappa}\right) \leq \frac{i_2\mu\tau}{i_1} \end{aligned} \tag{118}$$

for μ small enough. Then (115) follows from (117) and (118). ∎

3.10. Distinct orbits. Take $\tau, \tau' \in \mathcal{T}_\mu$. For μ small enough, τ, τ' satisfy the hypotheses of Proposition 3.11. Therefore, let

$$\big(I(t), \phi(t), \hat{z}(t)\big) := \big(I_0(\tau), \phi_0 + \tilde{\omega} t, 0\big) + \zeta(t), \qquad \phi_0 \in \mathbb{T}^N,\ \zeta \in \overline{H}^1_{T,a,s},\ T = 2\pi\tau,$$
$$\big(I'(t), \phi'(t), \hat{z}'(t)\big) := \big(I_0(\tau'), \phi'_0 + \tilde{\omega} t, 0\big) + \zeta'(t), \quad \phi'_0 \in \mathbb{T}^N,\ \zeta' \in \overline{H}^1_{T',a,s},\ T' = 2\pi\tau',$$

be two solutions of (75) found in Proposition 3.11; we recall that by (109),

$$\|\zeta\|_{H^1_{T,a,s}} \le \frac{\mu}{c_7}, \qquad \|\zeta'\|_{H^1_{T',a,s}} \le \frac{\mu}{c_7}. \tag{119}$$

Suppose that they are geometrically the same solution, namely, up to a time translation,

$$\big(I(t), \phi(t), \hat{z}(t)\big) = \big(I'(t), \phi'(t), \hat{z}'(t)\big) \qquad \forall\, t \in \mathbb{R}. \tag{120}$$

We claim that

$$|I_0(\tau) - I_0(\tau')| \le \frac{2\mu}{c_7}. \tag{121}$$

Indeed, using $I(t) = I'(t)$,

$$\begin{aligned} |I_0(\tau) - I_0(\tau')| &= |J(t) - J'(t)| \le |J(t)| + |J'(t)| \\ &\le \|\zeta(t)\|_{\mathcal{P}_{a,s}} + \|\zeta'(t)\|_{\mathcal{P}_{a,s}} \le \|\zeta\|_{H^1_{T,a,s}} + \|\zeta'\|_{H^1_{T',a,s}} \le \frac{2\mu}{c_7}, \end{aligned}$$

recalling (79) and (119).

Moreover, we claim that (120) implies also

$$\{\omega_{i_1}(\tau' - \tau)\} < \frac{\mu}{c_{10}} \qquad \text{or} \qquad \{\omega_{i_1}(\tau - \tau')\} < \frac{\mu}{c_{10}}, \tag{122}$$

with $c_{10} := c_7/2\|A\|$. Indeed, since $|A^{-1}v| \ge \|A\|^{-1}|v|$ for any $v \in \mathbb{R}^N$, recalling (54) and choosing $v := \{\omega\tau'\} - \{\omega\tau\}$, we have $I_0(\tau) - I_0(\tau') = A^{-1}v$, and therefore

$$|I_0(\tau) - I_0(\tau')| \ge \|A\|^{-1}|\{\omega\tau'\} - \{\omega\tau\}| \ge \|A\|^{-1}|\{\omega_{i_1}\tau'\} - \{\omega_{i_1}\tau\}|. \tag{123}$$

Noting that[6]

$$|\{\omega_{i_1}\tau'\} - \{\omega_{i_1}\tau\}| = \{\omega_{i_1}(\tau' - \tau)\} \qquad \text{or} \qquad |\{\omega_{i_1}\tau'\} - \{\omega_{i_1}\tau\}| = \{\omega_{i_1}(\tau - \tau')\},$$

(122) follows by (121) and (123).

LEMMA 3.16. *Let*

$$\mathcal{M} := \left\{ n \in \mathbb{Z}, \quad |n| \le \frac{c_4}{2\mu^2} \quad s.t. \quad \{\omega_{i_1} n\} \le \frac{\mu}{c_{10}} \right\}. \tag{124}$$

Then

$$\sharp\, \mathcal{M} \le \frac{10 c_4}{c_{10}\mu}.$$

[6]Indeed, if $\{x\} \ge \{y\}$, then

$$|\{x\} - \{y\}| = \{x\} - \{y\} = \big\{\{x\} - \{y\}\big\} = \{x - y - [x] + [y]\} = \{x - y\};$$

on the other hand, if $\{y\} \ge \{x\}$, then $|\{x\} - \{y\}| = \{y - x\}$.

Proof. We first note that, for μ small enough, we get

$$\left[\frac{2c_4}{i_1\mu}\right]\left[\frac{i_1}{\mu}\right] \geq \left(\frac{2c_4}{i_1\mu}-1\right)\left(\frac{i_1}{\mu}-1\right) \geq \frac{c_4}{\mu^2}+1,$$

and therefore,

$$\left[-\frac{c_4}{2\mu^2},\frac{c_4}{2\mu^2}\right] \subseteq \bigcup_{1\leq m\leq\left[\frac{2c_4}{i_1\mu}\right]} \left[\left[-\frac{c_4}{2\mu^2}\right]+(m-1)\left[\frac{i_1}{\mu}\right], \left[-\frac{c_4}{2\mu^2}\right]+m\left[\frac{i_1}{\mu}\right]\right). \tag{125}$$

Now we claim that

$$\sharp\left(\mathcal{M}\cap\left[\bar n,\bar n+\left[\frac{i_1}{\mu}\right]\right]\right) \leq \frac{4i_1}{c_{10}}+1 \qquad \forall\,\bar n\in\mathbb{Z}. \tag{126}$$

Then from (125) and (126) we have

$$\sharp\,\mathcal{M} \leq \left(\frac{4i_1}{c_{10}}+1\right)\left[\frac{2c_4}{i_1\mu}\right] \leq \frac{5i_1}{c_{10}}\frac{2c_4}{i_1\mu} \leq \frac{10c_4}{c_{10}\mu},$$

and the lemma follows.

Nothing remains but to prove (126). If $\mathcal{M}\cap\big[\bar n,\bar n+[i_1/\mu]\big)=\emptyset$, then (126) is trivially true. Otherwise let

$$n_0 := \min\big(\mathcal{M}\cap\big[\bar n,\bar n+[i_1/\mu]\big)\big).$$

By definition, $n\notin\mathcal{M}$ for any $\bar n\leq n<n_0$. Moreover, by (65), (66) we have

$$\{\omega_{i_1}n\}=\{\delta^{(1)}_{i_1}n\} \qquad \text{with} \quad \frac{\mu}{4i_1}<\delta^{(1)}_{i_1}<\frac{\mu}{2i_1}. \tag{127}$$

We now prove that

$$n\in\mathcal{M}\cap\big[\bar n,\bar n+[i_1/\mu]\big) \Longrightarrow n=n_0+n', \qquad 0\leq n'<4i_1/c_{10}, \tag{128}$$

from which (126) follows. By definition of n_0 it is obvious that $n'\geq 0$. Let us consider $n'\in\mathbb{N}$ such that

$$4i_1/c_{10}\leq n'<\bar n-n_0+[i_1/\mu].$$

For such n' we will show that $\{\omega_{i_1}(n_0+n')\}>\mu/c_{10}$. We have that

$$\{\omega_{i_1}(n_0+n')\}=\{\delta^{(1)}_{i_1}n_0+\delta^{(1)}_{i_1}n'\}=\{\{\delta^{(1)}_{i_1}n_0\}+\delta^{(1)}_{i_1}n'\}. \tag{129}$$

By (127)

$$\frac{\mu}{c_{10}}=\frac{\mu}{4i_1}\frac{4i_1}{c_{10}}<\{\delta^{(1)}_{i_1}n_0\}+\delta^{(1)}_{i_1}n'<\{\omega_{i_1}n_0\}+\frac{\mu}{2i_1}n'<\frac{\mu}{c_{10}}+\frac12\leq 1. \tag{130}$$

Therefore

$$\{\delta^{(1)}_{i_1}n_0\}+\delta^{(1)}_{i_1}n'=\big\{\{\delta^{(1)}_{i_1}n_0\}+\delta^{(1)}_{i_1}n'\big\}=\{\omega_{i_1}(n_0+n')\}$$

from (129). Finally, by (130),

$$\{\omega_{i_1}(n_0+n')\} > \frac{\mu}{c_{10}}.$$

Hence $n_0+n' \notin \mathcal{M}$, and (128) follows. □

LEMMA 3.17. *Fix $\tau \in \mathcal{T}_\mu$. For μ small enough*

$$\sharp\Big\{\tau' \in \mathcal{T}_\mu \quad s.t. \quad (120) \quad holds\Big\} \leq \frac{10c_4}{c_{10}\mu}.$$

Proof. If (120) holds, then τ and τ' verify (122). Hence $\tau - \tau' \in \mathcal{M}$, defined in (124). We conclude by Lemma 3.16. □

By Lemmas 3.4 and 3.17 we conclude that the number of geometrically distinct solutions found in Proposition 3.11 is greater than

$$\frac{c_4}{6\mu^2}\left(\frac{10c_4}{c_{10}\mu}\right)^{-1} = \frac{c_{10}}{60\mu}.$$

Actually, since in Proposition 3.11 N geometrically distinct orbits correspond to any τ, then the total number of geometrically distinct solutions is greater than c_{11}/μ with $c_{11} := c_{10}N/60$.

COROLLARY 3.18. *The total number of geometrically distinct solutions found in Proposition* 3.11 *is greater than c_{11}/μ.*

3.11. Proof of Theorem 1.1. Suppose that $\tau \in \mathcal{T}_\mu$, μ small enough, and consider a solution found in Proposition 3.11. Such a solution is of the form

$$(131)\quad \big(I(t),\phi(t),\hat{z}(t)\big) = \big(I_0(\tau),\tilde{\omega}t+\phi_0,0\big)+\zeta(t) = \big(I_0(\tau)+J(t),\tilde{\omega}t+\phi_0+\psi(t),w(t)\big)$$

with

$$(132)\qquad |J(t)| + |\psi(t)| + \|w(t)\|_{a,s} \leq \frac{\mu}{c_7} \quad \forall\, t \in \mathbb{R},$$

by (109) and (79). We now want to rewrite such a solution in the z variables defined in (30); by (33) and recalling that $\hat{z}_i = z_i$ for $i \in \mathcal{I}^c$ we get

$$\begin{cases} z_i(t) = \sqrt{I_i(t)}\big(\cos\phi_i(t) - \mathrm{i}\sin\phi_i(t)\big) & \text{for } i \in \mathcal{I}, \\ z_i(t) = w_i(t) & \text{for } i \in \mathcal{I}^c. \end{cases}$$

In the z_* variables of Proposition 2.4 we have

$$\begin{cases} \mathrm{z}_{*i}(t) = \eta\sqrt{I_i(t)}\big(\cos\phi_i(t) - \mathrm{i}\sin\phi_i(t)\big) & \text{for } i \in \mathcal{I}, \\ \mathrm{z}_{*i}(t) = \eta w_i(t) & \text{for } i \in \mathcal{I}^c \end{cases}$$

by (30) and

$$(133)\qquad \sup_{t\in\mathbb{R}} \|\mathrm{z}_*(t)\|_{a,s} = O(\eta).$$

We define $\check{z} := (\check{z}_i)_{i\geq 1}$ by

$$\check{z}_i(t) := \begin{cases} \sqrt{(I_0)_i}\big(\cos\big(\tilde{\omega}t + (\phi_0)_i\big) - \mathrm{i}\sin\big(\tilde{\omega}t + (\phi_0)_i\big)\big) & \text{for } i \in \mathcal{I}, \\ 0 & \text{for } i \in \mathcal{I}^c. \end{cases}$$

By (132) we get

$$\sup_{t\in\mathbb{R}} \|\mathrm{z}_*(t) - \eta\check{z}(t)\|_{a,s} = O\left(\eta\mu\right). \tag{134}$$

Concerning the z variables defined in (15) we have, recalling (18), (133), and (134),

$$\sup_{t\in\mathbb{R}} \|\mathrm{z}(t) - \eta\check{z}(t)\|_{a,s} = O\left(\eta\mu + \eta^3\right) = O\left(\eta\mu\right) = O\big(\mu^2\big), \tag{135}$$

since

$$\eta^2 = 1/\tau \le 2\mu^2/c_4, \tag{136}$$

recalling (53) and (59). Regarding the q variables defined in (9) we have, recalling (15) and (135),

$$\sup_{t\in\mathbb{R}} \|q(t) - \eta\check{q}(t)\|_{a,s} = O\big(\mu^2\big), \tag{137}$$

where $\check{q} := (\check{q}_i)_{i\ge1}$ and

$$\check{q}_i(t) := \begin{cases} \sqrt{2(I_0)_i}\cos\left(\tilde{\omega}t + (\phi_0)_i\right) & \text{for } i \in \mathcal{I}, \\ 0 & \text{for } i \in \mathcal{I}^c. \end{cases}$$

We note that, by Corollary 3.13, the solution in (131) belongs to $C^\infty(\mathbb{R}, \mathcal{P}_{a/2,s})$, and therefore, $q \in C^\infty(\mathbb{R}, \ell^{a/2,s})$. Finally, by Lemma 2.1, we have that

$$u(t,x) := \sum_{i\ge1} q_i(t)\sqrt{\frac{2}{\pi\omega_i}}\,\sin ix$$

belongs to $C^\infty(\mathbb{R}\times[0,\pi],\mathbb{R})$ and is a solution of (1). Defining

$$\tilde{u}(t,x) := \eta\sum_{i\in\mathcal{I}} 2\sqrt{\frac{(I_0)_i}{\pi\omega_i}}\cos\left(\tilde{\omega}t + (\phi_0)_i\right)\sin ix, \tag{138}$$

we have, for any $t\in\mathbb{R}$ and $x\in[0,\pi]$,

$$\begin{aligned} |u(t,x) - \tilde{u}(t,x)| &= \left|\sum_{i\ge1}\left(q_i(t) - \eta\check{q}_i(t)\right)\sqrt{\frac{2}{\pi\omega_i}}\,\sin ix\right| \\ &\le \sum_{i\ge1}\left|q_i(t) - \eta\check{q}_i(t)\right|\sqrt{\frac{2}{\pi\omega_i}} \\ &\le c_{12}\|q(t) - \eta\check{q}(t)\|_{a,s}, \end{aligned}$$

where, in the last line, we have used the Cauchy–Schwarz inequality. Therefore, by (137), we get

$$\sup_{t\in\mathbb{R},\, x\in[0,\pi]} |u(t,x) - \tilde{u}(t,x)| = O\big(\mu^2\big). \tag{139}$$

Define, for $i\in\mathcal{I}$,

$$a_i := 2\frac{\eta}{\mu}\sqrt{\frac{(I_0)_i}{\pi\omega_i}}. \tag{140}$$

Since $\eta = 1/\sqrt{\tau} \geq \mu/\sqrt{c_4}$ (recall (53) and (59)), by Lemma 3.1 we get

$$a_i \geq \frac{2}{\sqrt{c_4}}.$$

Defining $\varphi_i := (\phi_0)_i$ for $i \in \mathcal{I}$, (6) follows by (138), (139), and (140).

Estimate (8) follows from Lemma 3.15, while (7) follows from (45) and (136). Finally, the statement about the total number of geometrically distinct solutions follows from Corollary 3.18.

Remark 3.19. We can improve estimate (6) or, equivalently, (139). Indeed, for any fixed $k \geq 1$ and ζ in (131), by (110), we have $\|\zeta\|_{H^k_{T,a,s}} \leq \mathrm{const}(k)\,\mu$, where $\mathrm{const}(k)$ is a suitable large constant depending on $\mathcal{I}$, a, s, and k. Arguing as above and using the Sobolev immersion $H^k \subset C^{k-1}$, we get $\|q - \eta\check{q}\|_{C^{k-1}(\mathbb{R},\ell^{a,s})} \leq \mathrm{const}(k)\,\mu^2$. Therefore $\sup_{\mathbb{R}\times[0,\pi]} \left|\partial_t^h(u-\tilde{u})\right| \leq \mathrm{const}(k)\,\mu^2$ for any $h \leq k-1$. Since the estimates on the x-derivatives directly follows by the analyticity, we conclude that $\|u-\tilde{u}\|_{C^k(\mathbb{R}\times[0,\pi])} \leq \mathrm{const}(k)\,\mu^2$. We remark that, if one needs the previous C^k-estimate, the constant c in Theorem 1.1 must depend on k.

Appendix.

LEMMA A.1. *Except a countable set of $\mu > 0$ for any $\mathcal{I} = \{i_1 < \cdots < i_N\} \subset \mathbb{N}^+$, $N \geq 1$, the vector $\omega = (\omega_{i_1}, \ldots, \omega_{i_N})$, $\omega_i = \sqrt{i^2 + \mu}$, is rationally independent.*

Proof. For any $n \in \mathbb{Z}^N \setminus \{0\}$ let us define $E_n := \big\{\mu > 0 \quad \text{s.t.} \quad \omega \cdot n = 0\big\}$. We claim that E_n is at the most countable. Indeed, for $\mu > -1$, let us consider the analytic function

$$f_n(\mu) := \sum_{j=1}^{N} \sqrt{i_j^2 + \mu} \cdot n_j = \omega \cdot n.$$

It is enough to show that f_n is not identically zero, so that the set of its zeros is at the most countable. Suppose, by contradiction, that $f_n(\mu) \equiv 0$ for any $\mu > -1$, then $d^k f_n/d\mu^k(0) = 0$ for any $k \geq 1$, and therefore $\sum_{j=1}^N n_j\, i_j^{1-2k} = 0$ for any $k \geq 1$. Hence, multiplying for i_1^{2k-1}, we have

$$n_1 + \sum_{j=2}^{N} n_j \left(\frac{i_1}{i_j}\right)^{2k-1} = 0 \qquad \forall\, k \geq 1.$$

Noting that $i_j > i_1$ for any $j \geq 2$, and taking the limit for $k \to \infty$, we get $n_1 = 0$. In the same way one can prove that $n_2 = 0$ and, by induction, that $n_1 = \cdots = n_N = 0$. □

Proof of Lemma 2.1. We first prove that $u(t,x)$ is C^k. Moreover, for any fixed $t \in \mathbb{R}$, the function $x \mapsto \partial_t^h u(t,x)$, $h < k+1$, is real analytic with an analytic extension in the complex strip $|\mathrm{Im}\, x| < a$. Since $q \in C^k(\mathbb{R}, \ell^{a,s})$, we have that for all $t \in \mathbb{R}$,

$$\sum_{i\geq 1} |q_i(t)|^2 i^{2s} e^{2ai} = \|q(t)\|_{a,s}^2 \leq \|q\|^2_{C^k(\mathbb{R},\ell^{a,s})} < \infty. \tag{141}$$

For any fixed $\tilde{a} < a$ by (141), the Cauchy–Schwarz inequality, and

$$\sup_{|\mathrm{Im}\, x|\leq \tilde{a}} |\chi_i(x)| \leq \sup_{|\mathrm{Im}\, x|\leq \tilde{a}} |\sin ix| \leq e^{\tilde{a}i},$$

we get, for $i_0 \geq 1$,

$$\sup_{|\operatorname{Im} x|\leq \tilde{a}} \left| \sum_{i\geq i_0} \frac{q_i(t)}{\sqrt{\omega_i}}\, \chi_i(x) \right|^2 \leq \sum_{i\geq i_0} |q_i(t)|^2 i^{2s} e^{2ai} \sum_{i\geq i_0} \omega_i^{-1} i^{-2s} e^{-2(a-\tilde{a})i}$$
$$\leq \|q\|^2_{C^k(\mathbb{R},\ell^{a,s})} \sum_{i\geq i_0} \omega_i^{-1} i^{-2s} e^{-2(a-\tilde{a})i} \quad \stackrel{i_0\to\infty}{\longrightarrow} \quad 0,$$

from which we have that, for any $t \in \mathbb{R}$, the series in (13) uniformly converges to a 2π-periodic real analytic odd function with analytic extension on the complex strip $|\operatorname{Im} x| < a$. Moreover, again by the Cauchy–Schwarz inequality, we get

$$|\partial_x^h u(t,x)| \leq \|q(t)\|^2_{a,s} \sum_{i\geq 1} \omega_i^{-1} i^{2(h-s)} e^{-2ai} \qquad \forall\, (t,x) \in \mathbb{R}\times[0,\pi], \quad h \in \mathbb{N},$$

and, from (141), we have that

$$\forall\, h \in \mathbb{N} \;\; \exists\, c_{h,a,s} > 0 \;\; \text{s.t.} \qquad \sup_{(t,x)\in\mathbb{R}\times[0,\pi]} |\partial_x^h u(t,x)| \leq c_{h,a,s}. \tag{142}$$

By similar arguments one proves that for any $x \in [0,\pi]$, the function

$$\mathbb{R} \ni t \mapsto u(t,x) = \sum_{i\geq 1} \frac{q_i(t)}{\sqrt{\omega_i}}\, \chi_i(x)$$

is continuous since the series uniformly converges on $\mathbb{R}$ (and the functions $t \mapsto q_i(t)$ are continuous since $q \in C^0(\mathbb{R}, \ell^{a,s})$). The continuity in both variables follows by (142) and the continuity in t for x_0 fixed:

$$|u(t,x) - u(t_0,x_0)| \leq |u(t,x) - u(t,x_0)| + |u(t,x_0) - u(t_0,x_0)|$$
$$\leq c_{1,a,s}|x - x_0| + |u(t,x_0) - u(t_0,x_0)|.$$

Arguing as above one proves that, for any fixed $x \in [0,\pi]$, the series

$$\sum_{i\geq 1} \frac{1}{\sqrt{\omega_i}}\, \partial_t q_i(t) \chi_i(x)$$

uniformly converges for $t \in \mathbb{R}$; hence we can differentiate inside the summation in (13), obtaining

$$\partial_t u(t,x) = \sum_{i\geq 1} \frac{1}{\sqrt{\omega_i}}\, \partial_t q_i(t) \chi_i(x).$$

Carrying on the above arguments we finally have that $u \in C^k$ and that

$$\partial_t^h u(t,x) = \sum_{i\geq 1} \frac{1}{\sqrt{\omega_i}}\, \partial_t^h q_i(t) \chi_i(x), \qquad h < k+1, \tag{143}$$

is, for any fixed $t \in \mathbb{R}$, a 2π-periodic real analytic odd function with analytic extension on the complex strip $|\operatorname{Im} x| < a$.

We now prove that u, defined in (13), is a classical solution of (1). From (9) and (10),

$$\frac{\partial G}{\partial q_i} = \frac{1}{\sqrt{\omega_i}} \int_0^\pi f(u)\chi_i \, dx; \tag{144}$$

hence, we have, by (143) and (11),

$$\begin{aligned} u_{tt} &= \sum_{i\geq 1} \frac{\ddot{q}_i(t)}{\sqrt{\omega_i}} \chi_i(x) \\ &= \sum_{i\geq 1} \frac{1}{\sqrt{\omega_i}} \left(-\omega_i^2 q_i - \omega_i \frac{\partial G}{\partial q_i} \right) \chi_i \\ &= -\sum_{i\geq 1} \frac{q_i(t)}{\sqrt{\omega_i}} \, (\mu - \partial_{xx}) \, \chi_i(x) - \sum_{i\geq 1} \chi_i \int_0^\pi f(u)\chi_i \, dx, \end{aligned}$$

because ω_i^2 are the eigenvalues of the operator $\mu - \partial_{xx}$. Moreover, as χ_i for $i \geq 1$ is a complete orthonormal basis for the L^2 functions on $[0, \pi]$, we obtain

$$u_{tt} = -(\mu - \partial_{xx})u - \sum_{i\geq 1} \chi_i \int_0^\pi f(u)\chi_i \, dx = -(\mu - \partial_{xx})u - f(u). \qquad \square$$

REFERENCES

[A] A. Ambrosetti, *Critical Points and Nonlinear Variational Problems*, Mém. Soc. Math. Fr. (N.S.) 49, 1992, Historical Jrl., Ann Arbor, MI.

[B00] D. Bambusi, *Lyapunov center theorem for some nonlinear PDEs: A simple proof*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 29 (2000), pp. 823–837.

[BP01] D. Bambusi and S. Paleari, *Families of periodic solutions of resonant PDEs*, J. Nonlinear Sci., 11 (2001), pp. 69–87.

[BaB] D. Bambusi and M. Berti, *A Birkhoff–Lewis-type theorem for some Hamiltonian PDEs*, SIAM J. Math. Anal., 37 (2005), pp. 83–102.

[BBiV] M. Berti, L. Biasco, and E. Valdinoci, *Periodic orbits close to elliptic tori and applications to the three-body problem*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5), 3 (2004), pp. 87–138.

[BBo03] M. Berti and P. Bolle, *Periodic solutions of nonlinear wave equations with general nonlinearities*, Comm. Math. Phys., 243 (2003), pp. 315–328.

[BBo04] M. Berti and P. Bolle, *Multiplicity of periodic solutions of nonlinear wave equations*, Nonlinear Anal., 56 (2004), pp. 1011–1046.

[BBo06] M. Berti and P. Bolle, *Cantor families of periodic solutions for completely resonant nonlinear wave equations*, Duke Math. J., 134 (2006), pp. 359–419.

[BDG] L. Biasco and L. Di Gregorio, *Periodic solutions of Birkhoff–Lewis type for the nonlinear wave equation*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei, 17 (2006), pp. 25–33.

[BL34] G. D. Birkhoff and D. C. Lewis, *On the periodic motions near a given periodic motion of a dynamical system*, Ann. Mat. Pura Appl. (4), 12 (1934), pp. 117–133.

[Bou99] J. Bourgain, *Periodic solutions of nonlinear wave equations*, in Harmonic Analysis and Partial Differential Equations, Chicago Lectures in Math., Univ. Chicago Press, Chicago, IL, 1999, pp. 69–97.

[C00] W. Craig, *Problèmes de petits diviseurs dans les équations aux dérivées partielles*, Panor. Synthèses 9, Société Mathématique de France, Paris, 2000.

[CW93] W. Craig and C. E. Wayne, *Newton's method and periodic solutions of nonlinear wave equations*, Comm. Pure Appl. Math., 46 (1993), pp. 1409–1498.

[GMPr05] G. Gentile, V. Mastropietro, and M. Procesi, *Periodic solutions for completely resonant nonlinear wave equations with Dirichlet boundary conditions*, Comm. Math. Phys., 256 (2005), pp. 437–490.

[K87] S. B. Kuksin, *Hamiltonian perturbations of infinite-dimensional linear systems with imaginary spectrum*, Funktsional. Anal. i Prilozhen., 21 (1987), pp. 22–37. (In Russian.)

[K93] S. B. Kuksin, *Nearly Integrable Infinite-Dimensional Hamiltonian Systems*, Lecture Notes in Math. 1556, Springer-Verlag, Berlin, 1993.

[L34] D. C. Lewis, *Sulle oscillazioni periodiche d'un sistema dinamico*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 19 (1934), pp. 234–237.

[Mo77] J. Moser, *Proof of a generalized form of a fixed point theorem due to G.D. Birkhoff*, in Geometry and Topology (Proc. III Latin Amer. School of Math., Inst. Mat. Pura Aplicada CNPq, Rio de Janeiro, 1976), Lecture Notes in Math. 597, Springer, Berlin, 1977, pp. 464–494.

[P96] J. Pöschel, *Quasi-periodic solutions for a nonlinear wave equation*, Comment. Math. Helv., 71 (1996), pp. 269–296.

[T] G. Tarantello, *Solutions with prescribed minimal period for nonlinear vibrating strings*, Comm. Partial Differential Equations, 12 (1987), pp. 1071–1094.

[W90] C. E. Wayne, *Periodic and quasi-periodic solutions of nonlinear wave equations via KAM theory*, Comm. Math. Phys., 127 (1990), pp. 479–528.

[We73] A. Weinstein, *Normal modes for nonlinear Hamiltonian systems*, Invent. Math., 20 (1973), pp. 47–57.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1126–1153

ON THE GLOBAL EXISTENCE OF WEAK SOLUTIONS FOR THE NAVIER–STOKES EQUATIONS OF COMPRESSIBLE FLUID FLOWS*

MIKHAIL PEREPELITSA†

Abstract. We study the Cauchy problem for the two dimensional Navier–Stokes equations of compressible fluid flows with periodic initial data. We assume that the bulk viscosity coefficient depends on the density of the flow. The global existence of a weak solution with uniform lower and upper bounds on the density, as well as the decay of the solution to an equilibrium state, is proved when the initial datum, $(\rho_0, \mathbf{u}_0)$, does not contain vacuum and belongs to the space $L^\infty\left(\mathbb{T}^2\right) \times \left[W^{1,2}\left(\mathbb{T}^2\right)\right]^2$, where $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$.

Key words. compressible Navier–Stokes equations, Orlicz spaces, commutator operators

AMS subject classifications. 76N10, 35Q30

DOI. 10.1137/040619119

1. Introduction.

In the mathematical model of the motion of a compressible, viscous, isentropic gas, the gas is described by its density (scalar), $\rho(t,x)$, and velocity (vector), $\mathbf{u}(t,x)$, both being functions of time t and position in physical space $x \in \mathbb{R}^3$. If there are no external forces acting on the gas, the flow of the gas is plane-parallel and does not depend on the displacement in the transversal direction. The model is described by the following system of the two dimensional compressible Navier–Stokes equations:

$$\rho_t + \operatorname{div} \rho\mathbf{u} = 0, \tag{1}$$

$$(\rho\mathbf{u})_t + \operatorname{div} \rho\mathbf{u} \otimes \mathbf{u} = \operatorname{div} \mathbb{S}, \tag{2}$$

$$(t,x) \in \mathbb{R}^+ \times \mathbb{R}^2,$$

where $\mathbf{u} = (u_1(t,x),\, u_2(t,x))$, $\mathbf{u} \otimes \mathbf{u}$ is a tensor with components $\{\mathbf{u} \otimes \mathbf{u}\}_{ij} = u_i u_j$, $i,j = 1,2$, and the stress tensor, $\mathbb{S}$, is given by the Stokes formula:

$$\mathbb{S} = (-P + \zeta \operatorname{div} \mathbf{u})\mathbb{I} + 2\mu\left(\mathbb{D} - \frac{1}{3}\operatorname{div} \mathbf{u}\, \mathbb{I}\right). \tag{3}$$

In the above formula, $\mu > 0$ and $\zeta \geq 0$ are the shear and the bulk viscosity coefficients, $P = a\rho^\gamma$, $\gamma > 1$, $a > 0$, is the pressure, $\mathbb{I}$ is the identity tensor, and $\mathbb{D}$ is the rate of strain tensor that has components $\mathbb{D}_{ij} = \frac{1}{2}\left(\partial_i u_j + \partial_j u_i\right)$, $i,j = 1,2$. We stress here that the condition on the flow to be isentropic holds only approximately, since the entropy is generated in the irreversible process set up in the flow by the friction; see [1, section 3.4]. Nevertheless, the full understanding of this model is essential in studying compressible flows and is not yet complete; see Feireisl [4], Hoff [7], Lions [11], [10], and Novotný and Straskraba [12] for detailed treatment of various models of motion of compressible flows.

The viscosity coefficients, in general, depend on the thermodynamic parameters. For example, for polyatomic gases it is expected that ζ, μ are proportional to θ^β,

*Received by the editors November 18, 2004; accepted for publication (in revised form) November 9, 2005; published electronically December 11, 2006.
http://www.siam.org/journals/sima/38-4/61911.html
†Department of Mathematics, 1326 Stevenson Center, Vanderbilt University, Nashville, TN 37240 (mikhail.perepelitsa@vanderbilt.edu).

$\beta > 0$, where θ is the temperature; see [1, section 1.7]. Since for isentropic flows $\theta \sim \rho^{\gamma-1}$, we are led to the following hypothesis:

$$\zeta = b\rho^{\beta}, \quad \mu = \text{const.} > 0, \quad b > 0,\ \beta > 0. \tag{4}$$

The assumption on the shear viscosity being a constant is the technical restriction that reflects the fact that at present we do not have tools to treat a nontrivial dependence $\mu = \mu(\rho)$.

We consider the Cauchy problem with the given initial data ρ_0 and u_0, which are periodic with period 1 in each space direction x_i, $i = 1, 2$, i.e., functions defined on $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$. We require that

$$\mathbf{u}(0, x) = \mathbf{u}_0(x), \quad \rho(0, x) = \rho_0(x), \quad x \in \mathbb{T}^2. \tag{5}$$

There is a well-developed theory of weak solutions for the system of the multidimensional compressible Navier-Stokes equations with both viscosity coefficients, $\mu > 0$, $\zeta \geq 0$, being constants. In Lions [11], the global in time existence of weak solutions of the multidimensional isentropic Navier–Stokes equations is proved for the large values of γ. Later, Feireisl, Novotný, and Petzeltová [5] proved the existence of weak solutions for the physical range of γ, i.e., $\gamma > 1$ if $N = 2$ and $\gamma > \frac{5}{3}$ if $N = 3$. Weak solutions considered in those works have *a minimal amount of regularity*, and there are some important questions, such as formation of vacuum or blow up of uniform bounds on density, that are still unanswered.

A development in this direction was made in Desjardins [3], where the author proves that in dimensions $N = 2, 3$, with the assumption on the initial data, $\rho_0 \in L^{\infty}(\mathbb{T}^N)$, $\mathbf{u}_0 \in \left[W^{1,2}(\mathbb{T}^N)\right]^N$, there exists a weak solution such that the L^{∞} norm of the density does not blow up and no vacuum forms at *small times* $t \in [0, T_0[$. A related result of Hoff [7] establishes the global in time existence of weak solutions that are *small perturbations of an equilibrium state*, measured in weak norms. The weak solution constructed in the later work has a fair amount of regularity, and in particular, the density is a L^{∞} function and no vacuum occurs at positive times.

In contrast, under the assumptions (4), which have a certain physical background behind them, the problem (1)–(5) admits a *globally* defined solution with *no restrictions* on the size of initial data, and its regularity is controlled by the regularity of initial data. This result was proved in Kazhikhov and Waigant [8] and is stated below. Some extensions of these theorems that take into account the presence of external forces are possible; see [8].

The existence of the classical solution is established in the following theorem.

THEOREM A (see [8]). *Let $\beta > 3$ and $\gamma \geq 0$. Let $\rho_0(x)$, $\mathbf{u}_0(x)$ be such that*

$$0 < m < \rho_0(x) < M < +\infty, \quad x \in \mathbb{T}^2,$$
$$(\rho_0, \mathbf{u}_0) \in C^{1+\omega}\left(\mathbb{T}^2\right) \times \left[C^{2+\omega}\left(\mathbb{T}^2\right)\right]^2, \quad 0 < \omega < 1,$$

for some m, M. Then, there exists the unique, classical solution of the problem (1)–(5),

$$\rho \in C^{1+\omega/2, 1+\omega}\left([0, +\infty) \times \mathbb{T}^2\right),$$
$$\mathbf{u} \in \left[C^{1+\omega/2, 2+\omega}\left([0, +\infty) \times \mathbb{T}^2\right)\right]^2.$$

The solution exists for all times $t > 0$, and the density is bounded away from vacuum.

The existence of a weak solution is established in the next theorem.

THEOREM B (see [8]). *Let $\beta > 3$ and $\gamma \geq 0$. Let $\rho_0(x)$, $\mathbf{u}_0(x)$ be such that*

$$0 < m < \rho_0(x) < M < +\infty, \quad x \in \mathbb{T}^2,$$
$$(\rho_0, \mathbf{u}_0) \in W^{1,q}\left(\mathbb{T}^2\right) \times \left[W^{2,2}\left(\mathbb{T}^2\right)\right]^2, \quad q > 2,$$

for some m, M. Then, there exists a strong solution of the problem (1)–(5). The solution exists for all times $t > 0$, and the density is an L^∞ function and is bounded away from vacuum.

The existence of weak solutions for which the density is an element of L^p, $p > 1$, spaces is proved in the next theorem.

THEOREM C (see [8]). *Let $\beta > 3$ and $\gamma \geq 0$. Let $\rho_0(x)$, $\mathbf{u}_0(x)$ such that*

$$(\rho_0, \mathbf{u}_0) \in L^\infty\left(\mathbb{T}^2\right) \times \left[W^{1,2}\left(\mathbb{T}^2\right)\right]^2.$$

Then, there exists a weak solution of the problem (1)–(5). The solution exists for all times $t > 0$.

In Theorems A and B it is shown that the solution does not admit vacuum at positive times and the density of the flow is uniformly bounded. At the same time, the density at time $t = 0$ must be differentiable, in the classical or weak sense, ruling out the solutions with jump discontinuities.

On the other hand, in Theorem C, there is only one assumption on the density, $\rho_0 \in L^\infty$, but it is not known if the solution is bounded at positive times and/or there are vacuum zones appearing in the flow. This is the question we address in the present paper.

We derive *new a priori estimates* that provide the uniform bounds on the density form above and below.

We investigate, as well, the long time behavior and obtain a weak solution that converges in time to a constant state $(\int_{\mathbb{T}^2} \rho_0(\cdot), \mathbf{0})$ in the norm $\|\cdot\|_{L^\infty} \times \|\cdot\|_{L^p}$, $p > 1$, in the case that the bulk viscosity is comparable with the pressure ($\beta = \gamma$). We prove the following theorem.

THEOREM. *Let $\beta > 3$ and $\gamma > 1$. Let*

$$(\rho_0, \mathbf{u}_0) \in L^\infty\left(\mathbb{T}^2\right) \times \left[W^{1,2}\left(\mathbb{T}^2\right)\right]^2$$

and

$$0 < m < \operatorname{infess}_{\mathbb{T}^2} \rho_0(\cdot) \leq \operatorname{supess}_{\mathbb{T}^2} \rho(\cdot) < M \tag{6}$$

for some m, M. Then, there exists a global weak solution of the problem (1)–(5) with the following properties.

$$(\rho, \mathbf{u}) \in C\left([0, +\infty); L^2\left(\mathbb{T}^2\right)\right) \times \left[C\left([0, +\infty); L^2_w\left(\mathbb{T}^2\right)\right)\right]^2,$$

and for any $T > 0$ there are $\bar{M} = \bar{M}(T)$ and $\bar{m} = \bar{m}(T) > 0$ such that

$$\bar{m} < \rho(t, x) < \bar{M}, \quad t \in (0, T), \ x \text{ a.e. } \mathbb{T}^2.$$

If $\beta = \gamma$, then there exists a weak solution such that, for any $p > 1$,

$$\|\rho(t, \cdot) - \int_{\mathbb{T}^2} \rho_0(s)\, ds\|_{L^\infty(\mathbb{T}^2)} + \|\mathbf{u}(t, \cdot)\|_{L^p(\mathbb{T}^2)} \to 0, \quad t \to +\infty. \tag{7}$$

REMARK 1. *L^2_w is the space of L^2 functions with L^2 weak topology. The additional regularity for velocity* $\mathbf{u}$ *is available. The complete list of properties of* $\mathbf{u}$ *can be found in section* 6.

The central point of the proof is contained in the analysis of the rate of production of the density given by the quantity

$$-\rho \operatorname{div} \mathbf{u} = -\rho(4/3\mu+\zeta)^{-1}B - \rho(4/3\mu+\zeta)^{-1}P,$$

where $B = (4\mu/3+\zeta)\operatorname{div}\mathbf{u} - P$, the effective viscous flux, solves the following Poisson equation:

$$\Delta B = \partial_t(\operatorname{div}\rho\mathbf{u}) + \operatorname{div}\operatorname{div}\rho\mathbf{u}\otimes\mathbf{u}.$$

Accordingly, we can write

$$B + \frac{d}{dt}(-\Delta)^{-1}[\operatorname{div}\rho\mathbf{u}] = \mathbf{u}\cdot(-\Delta)^{-\frac{1}{2}}\nabla(-\Delta)^{-\frac{1}{2}}\operatorname{div}[\rho\mathbf{u}] - (-\Delta)^{-\frac{1}{2}}\operatorname{div}(-\Delta)^{-\frac{1}{2}}\operatorname{div}[\rho\mathbf{u}\otimes\mathbf{u}], \tag{8}$$

and the terms on the right-hand side can be written as a sum of commutators of Riesz transforms and the operators of multiplication by u_i. The framework of this analysis originates in the works of Hoff [7] and Lions [11]. Using the smoothing properties of the commutators we obtain an estimate on the L^∞ norm of the commutators in terms of L^∞ norm of ρ and $\|\nabla\mathbf{u}\|_{L^2}$. This is a delicate point of the analysis, as we have to use Poincaré–Sobolev type estimates stated in terms of Orlicz norms determined by the convex functions $r^2\log^p(2+r)$ to show that the right-hand side of (8) can be bounded by $\|\rho\|_{L^\infty}\log^\alpha(2+\|\nabla\mathbf{u}\|_{L^2})$, $\alpha > 2$.

On the other hand, using energy type estimates (at this point it is essential to work with function defined on $\mathbb{R}^2$) and dvi-curl lemma we show that $\log(2+\|\nabla\mathbf{u}\|_{L^2})$ does not exceed a linear function of $\|\rho\|_{L^\infty}$. Both estimates lead to the a priori estimate on $\|\rho\|_{L^\infty}$ when $\beta > 3$ and consequently on $\|\nabla\mathbf{u}\|_{L^2}$.

In the above argument $(\rho, \mathbf{u})$ is the classical solution which exists globally in time if initial data are smooth; see Theorem A. Naturally, the next step is the study of the compactness properties of the sequence of such solutions, $(\rho^n, \mathbf{u}^n)$, which correspond to the initial data approximating the given pair $(\rho_0, \mathbf{u}_0)$ from $L^\infty \times \left(W^{1,2}\right)^2$. Note that in the case when the viscosity coefficient depends on density, the argument of Lions [11] *cannot* be applied here since the effective viscous fluxes, B^n, incorporate the term $\zeta(\rho^n)\operatorname{div}\mathbf{u}^n$ that is not in general weakly continuous. This difficulty was resolved in Kazhikhov and Waigant [8] by utilizing a uniform boundedness of the norms $\|\nabla B^n\|_{L^2((0,T)\times\mathbb{T}^2)}$, $T > 0$, which provides the strong compactness of $\{B^n\}$. For the purpose of completeness of the work we present the proof of strong compactness of the sequence of smooth solutions in the last section.

The work is organized as follows. Section 2 contains notations, the necessary mathematical apparatus, and the derivation of the complimentary system of equations. In section 3 we obtain estimates based on the first energy inequality. Section 4 deals with additional energy type estimates. Section 5 is devoted to the derivation of uniform upper and lower bounds on the density. In section 6 we establish the existence of a weak solution as limit of smooth solutions for which bounds from the previous sections apply.

2. Notations and preliminary lemmas. We use the following notation:

$$\begin{cases} \partial_i = \frac{\partial}{\partial x_i},\ i=1,2, & \nabla = (\partial_1, \partial_2), \\ \nabla^\perp = (\partial_2, -\partial_1), & \frac{\mathrm{d}}{\mathrm{dt}} = \frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla, \\ \operatorname{div}\mathbf{u} = \nabla\cdot\mathbf{u}, & \operatorname{rot}\mathbf{u} = \nabla^\perp\cdot\mathbf{u}, \\ \Pi = \left[-\frac{1}{2},\frac{1}{2}\right]\times\left[-\frac{1}{2},\frac{1}{2}\right]\subset\mathbb{R}^2, & f_\Pi = \frac{1}{|\Pi|}\int_{\mathbb{T}^2} f(x)\,dx. \end{cases}$$

Let $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ be a two dimensional torus. We identify functions defined on $\mathbb{T}^2$ with their periodic counterparts defined on $\mathbb{R}^2$. By $\|\cdot\|_p$, $1 \le p \le +\infty$, we denote Lebesgue $L^p(\mathbb{T}^2)$ norms of a function. We use standard notation $W^{k,q}\left(\mathbb{T}^2\right)$, $q \in [1,+\infty]$, k an integer, for Sobolev spaces of functions that possess weak derivative up to the kth order, which belong to $L^q(\mathbb{T}^2)$. $W^{-k,\frac{q}{q-1}}\left(\mathbb{T}^2\right)$, $q \in [1,+\infty)$ is the dual space of $W^{k,q}\left(\mathbb{T}^2\right)$. We use the same notation for norms of scalar and vector functions.

The material on Fourier series and integral operators presented below can be found, for example, in [14] and [15]. Let $f \in L^1\left(\mathbb{T}^2\right)$. The set of Fourier coefficients of f, $\{c_n\}_{n\in Z^2}$, where

$$c_n = \int_\Pi f(y) e^{-2\pi\mathbf{i}(n\cdot y)}\,dy,$$

is well defined, and we formally write $f \sim \sum_{n\in Z^2} c_n e^{2\pi\mathbf{i}(n\cdot x)}$. For $i = 1,2$, the ith Riesz transform of f is defined as a multiplier

$$R_i(f) \sim \sum_{n\in Z^2, n\neq 0} -\mathbf{i}\frac{n_i}{|n|} c_n e^{2\pi\mathbf{i}(n\cdot x)}.$$

It is well known that R_i is a linear continuous operator from any $L^p(\mathbb{T}^2)$ to itself, for $1 < p < +\infty$. We use this property in the following form.

LEMMA 1. *There is $C = C(p)$, $p \in]1,+\infty[$, such that for any $\mathbf{f} = (f_1, f_2)$, a vector function with components in $L^1(\mathbb{T}^2)$ such that $\mathbf{F} = (F_1, F_2) = (\operatorname{div}\mathbf{f}, \operatorname{rot}\mathbf{f})$ belongs to $\left(L^p(\mathbb{T}^2)\right)^2$, it holds that*

$$\|\nabla\mathbf{f}\|_p \le C\|\mathbf{F}\|_p. \tag{9}$$

Proof. Indeed, for $i = 1,2$ we can formally write

$$\begin{aligned} -\partial_i f_1 &= R_i \circ (R_1(F_1) + R_2(F_2)), \\ -\partial_i f_2 &= R_i \circ (R_2(F_1) - R_1(F_2)). \end{aligned}$$

Therefore, L^p estimates on singular integral operators apply, and using a suitable approximation argument we conclude. ∎

Let L^1_0 be the space of integrable functions with zero mean. Define normed spaces

$$\mathcal{H}^1 = \left\{ f \in L^1_0\left(\mathbb{T}^2\right) : ||f||_{\mathcal{H}^1} \triangleq ||f||_1 + ||R_1(f)||_1 + ||R_2(f)||_1 < +\infty \right\},$$

where R_i, $i = 1,2$, are Riesz transforms introduced above, and

$$\mathcal{BMO} = \left\{ f \in L^1_{loc}\left(\mathbb{T}^2\right) \,:\, ||f||_{\mathcal{BMO}} < +\infty \right\},$$

where $||f||_{\mathcal{BMO}} = \sup_{\{0<\epsilon<1,\, x\in\Pi\}} \frac{1}{|B_\epsilon(0)|}\int_{B_\epsilon(x)} |f(y) - f_{B_\epsilon(x)}|\,dy$. In the case of $\mathcal{BMO}$ space, functions are defined modulo additive constants.

We will use standard notation for operators $(\Delta)^{-1} : L_0^p\left(\mathbb{T}^2\right) \to W^{2,p}\left(\mathbb{T}^2\right)$, the inverse Laplace operator, and $(-\Delta)^{-1}\mathrm{div} : L^p\left(\mathbb{T}^2\right) \to W^{1,p}\left(\mathbb{T}^2\right)$, where $1 < p < +\infty$ and L_0^p is the space of L^p integrable functions with zero mean. In terms of Fourier series these operators are defined as follows:

$$\begin{cases} (-\Delta)^{-1}(f) & \sim & \sum_{n\in Z^2, n\neq 0} \frac{1}{4\pi^2|n|^2} c_n e^{2\pi \mathbf{i}(n\cdot x)}, \\ (-\Delta)^{-1}\mathrm{div}\,(f) & \sim & \sum_{n\in Z^2, n\neq 0} -\frac{\mathbf{i}}{2\pi} \frac{n_1+n_2}{|n|^2} c_n e^{2\pi \mathbf{i}(n\cdot x)}. \end{cases}$$

Consider the composition of two Riesz transforms, $R_i \circ R_j$, $i, j = 1, 2$. There is a representation of this operator as a singular integral

$$R_i \circ R_j(f)(x) = p.v. \int_{\mathbb{T}^2} K_{ij}(x-y) f(y)\, dy,$$

where the kernel $K_{ij}(x)$, $i, j = 1, 2$, has a singularity of the second order at 0 and

$$|K_{ij}(x)| \leq C|x|^{-2}, \quad x \in \mathbb{T}^2,$$

for some constant $C > 0$. Given function g, define the linear operator

$$[g, R_iR_j](f) \equiv gR_i \circ R_j(f) - R_i \circ R_j(gf), \quad i, j = 1, 2.$$

This operator can be written as a convolution with the singular kernel K_{ij},

$$[g, R_iR_j](f)(x) = p.v. \int_{\mathbb{T}^2} K_{ij}(x-y)(g(x) - g(y)) f(y)\, dy, \quad i, j = 1, 2.$$

The commutator possesses a number of remarkable properties. We will use the following result of Coifman et al. [2].

LEMMA 2. *Let g, $f \in C^\infty(\mathbb{T}^2)$, and $p > 1$. Then, there is $C = C(p)$ such that*

$$\|[g, R_iR_j](f)\|_p \leq C\|g\|_{\mathcal{BMO}}\|f\|_p, \quad i, j = 1, 2. \tag{10}$$

We will use Orlicz norms of functions as they arise naturally from the type of methods we use, i.e., energy estimates for gradients of functions defined on $\mathbb{T}^2$. The extensive study of these spaces is contained in [9].

DEFINITION 1. *Convex, even function $\Phi : \mathbb{R} \to \mathbb{R}^+$ is called a Young function if $\Phi(0) = 0$ and $\lim_{r\to\infty} \Phi(r) = +\infty$.*

Let Φ be a Young function.

DEFINITION 2. *The functional $N_\Phi : L^1(\mathbb{T}^2) \to \mathbb{R}^+$ defined by*

$$N_\Phi(f) = \inf\left\{\lambda > 0 : \int_{\mathbb{T}^2} \Phi\left(\frac{f}{\lambda}\right) dx \leq 1\right\} \tag{11}$$

is called an Orlicz norm of function f.

$N_\Phi(\cdot)$ is a norm. There is an analogue of Hölder inequalities for Orlicz norms. It is contained in the next lemma. The proof is an easy consequence of the definition of $N_\Phi(\cdot)$ and the convexity of the defining function Φ.

LEMMA 3. *Let Φ_i, $i = 1, 2, 3$, be left-continuous Young functions that satisfy the following inequality for all x, $y > 0$ and some positive constants C_1, C_2:*

$$\Phi_3(xy) \leq C_1\Phi_1(x) + C_2\Phi_2(y). \tag{12}$$

Then, for all f_1, f_2, $f_3 \in L^1(\mathbb{T}^2)$, and some constant $C > 0$,

$$N_{\Phi_3}(f_1 f_2) \le C N_{\Phi_1}(f_1) N_{\Phi_2}(f_2). \tag{13}$$

The use of Lemma 3 will be limited to the Young functions appearing in the next lemma, the proof of which is omitted.

LEMMA 4. *There are constants C_i, $i = 1, \dots, 3$, such that for all $x, y \ge 0$ the following inequalities hold:*

$$xy \le C_1(p) x^2 \left[\log(e^2 + x)\right]^p + C_1(p) y^2 \left[\log(e^2 + y)\right]^{-p}, \quad p > 0, \tag{14}$$

$$xy \le C_2(p) \left[\exp(x^p) - 1\right] + C_2(p) y \left[\log(e^2 + y)\right]^{\frac{1}{p}}, \quad p > 0, \tag{15}$$

$$(xy)^2 \left[\log(e^2 + xy)\right]^p \le C_3 \Psi_s(x) + C_3 y^2 \left[\log(e^2 + y)\right]^r \tag{16}$$

with $r > p > 1$, $s > 0$, and a Young function Ψ_s such that

$$\lim_{x \to +\infty} \frac{\exp\exp(2x^{\frac{1}{s}})}{\Psi_s(x)} = 1. \tag{17}$$

In what follows we use notation $\Phi_p(x) = x^2 \left[\log(e^2 + x)\right]^p$ and $\Upsilon_p(x) = e^{x^p} - 1$, $p > 1$, $x \ge 0$. We now derive Poincaré–Sobolev type inequalities.

LEMMA 5. *For any $p \in]1, +\infty[$ there are C and ϵ_0 such that for any $g \in C^\infty(\mathbb{T}^2)$ and $x \in \mathbb{T}^2$ it holds that*

$$|g(x + z) - g(x)| \le C \left[-\log \epsilon\right]^{-\frac{p-1}{2}} N_{\Phi_p}(\nabla g \,:\, B_{3\epsilon}(x)), \quad |z| < \epsilon < \epsilon_0.$$

Proof. Indeed, for all $y \in B_\epsilon(x)$, the integration by parts can be used to establish the inequality

$$|g(y) - g_{B_\epsilon(x)}| \le \frac{1}{|B_\epsilon(0)|} \int_{B_{3\epsilon}(0)} \frac{|\nabla g|(y + z)}{|z|}\, dz,$$

where $g_{B_\epsilon(x)}$ is the average of $g(x)$ over the ball $B_\epsilon(x) = \{|y - x| < \epsilon\}$. We now apply Lemma 3 together with Lemma 4, where we set $f_1(z) = |z|^{-1}$, $f_2(z) = |\nabla g|(y+z)$ with $\Phi_1(r) = \Phi_{-p}(r)$, $\Phi_2(r) = \Phi_p(r)$ and $\Phi_3(r) = r$. Let us compute $N_{\Phi_1}(|z|^{-1} \,:\, B_{3\epsilon}(0))$:

$$\begin{aligned} N_{\Phi_1}(|z|^{-1} \,:\, B_{3\epsilon}(0)) &= \inf\left\{\lambda > 0 \,:\, \int_{B_{3\epsilon}(0)} \frac{1}{|z|^2 \lambda^2 \log^p(e^2 + (|z|\lambda)^{-1})} < 1\right\} \\ &\le \inf\left\{0 < \lambda < (3\epsilon)^{-1} \,:\, 2\pi \int_0^{3\epsilon} \frac{1}{r\lambda^2 \log^p(r\lambda)^{-1}}\, dr < 1\right\}. \end{aligned}$$

Then, there is $C = C(p)$ such that

$$N_{\Phi_1}(|z|^{-1} : B_{3\epsilon}(0)) \le C \left[-\log \epsilon\right]^{-\frac{p-1}{2}},$$

when ϵ is sufficiently small (depending on p), and this concludes the proof. ∎

With the help of the last lemma we establish the L^∞ estimates on the commutators that have been introduced above.

LEMMA 6. *Let g, $f \in C^\infty(\mathbb{T}^2)$. Let $\Upsilon_\omega(r) = \exp(r^\omega) - 1$, $\omega > 0$. Then, if*

$$\frac{p-1}{2} - \frac{1}{\omega} > 1,$$

there is $C > 0$, independent of f and g, with the property

$$\sup_{\mathbb{T}^2} |[g, R_i R_j](f)(\cdot)| \le C N_{\Upsilon_\omega}(f) N_{\Phi_p}(\nabla g).$$

Proof. The proof is a simple application of the Hölder inequality given by Lemma 3. Let us use the letter C to denote the generic constant that is independent of functions f and g. In formula (2) we split the integral in two: one taken over the set $|y - x| > \epsilon$ and the singular part $|y - x| < \epsilon$, for some $\epsilon > 0$:

$$\begin{aligned}[g, R_i R_j](f)(x) = &\int_{\mathbb{T}^2 \cap \{|x-y| > \epsilon\}} K_{ij}(x-y)(g(x) - g(y)) f(y)\, dy \\ &+ p.v. \int_{\mathbb{T}^2 \cap \{|x-y| < \epsilon\}} K_{ij}(x-y)(g(x) - g(y)) f(y)\, dy.\end{aligned}$$

By the fact that $K_{ij}(x)$ has a singularity of the second order at 0, and using the estimate on $g(x) - g(y)$ from Lemma 5, we obtain

$$\begin{aligned}|[g, R_i R_j](f)(x)| \le{}& C(\epsilon) N_{\Phi_p}(|\nabla g|) \int_{\mathbb{T}^2} |f(y)| \\ &+ C N_{\Phi_p}(|\nabla g| : B_{3\epsilon}(x)) \int_{B_\epsilon(0)} |z|^{-2} [-\log |z|]^{-\frac{p-1}{2}} f(x+z).\end{aligned}$$

Young functions $\Upsilon_\omega(r)$ and $r \log^{\frac{1}{\omega}}(e^2 + r)$ satisfy conditions of Lemma 3 which we apply for the last integral. It is easy to see that there is $C > 0$ for which

$$N_{r \log^{\frac{1}{\omega}}(e^2+r)} \left(|z|^{-2} [-\log |z|]^{-\frac{p-1}{2}} \,:\, B_\epsilon(0) \right) < C.$$

This proves the lemma, since $\int_{\mathbb{T}^2} |f(x)|\, dx \le C N_{\Upsilon_\omega}(f)$, for suitable constant C. ∎

The next two lemmas are the well-known Poincaré–Sobolev inequalities. Proofs can be found, for example, in Ziemer [17].

LEMMA 7. *Let $2 < p < +\infty$, and $\mathcal{A} \subset \mathbb{T}^2$ of a nonzero Lebesgue measure. There is a constant $C = C(p, |\mathcal{A}|)$ for which*

$$\|f - f_{\mathcal{A}}\|_p \le C \|\nabla f\|_{\frac{2p}{2+p}}$$

holds for any $f \in C^\infty(\mathbb{T}^2)$. We use notation $f_{\mathcal{A}} = |\mathcal{A}|^{-1} \int_{\mathcal{A}} f(\cdot)$.

LEMMA 8. *For any $p > 2$ there is a $C = C(p) > 0$ such that*

$$\sup_{\mathbb{T}^2} |f| \le |f^2_{\mathbb{T}}| + C \|\nabla f\|_p$$

holds for any $f \in C^\infty(\mathbb{T}^2)$. Also, there are absolute constants c_1, $c_2 > 0$, such that for any cube C it holds that

$$\int_C \exp\left(\frac{|f(x) - f_C|^2}{c_1 \|\nabla f\|_2^2} \right) dx \le c_2 |C|$$

for any $f \in C^\infty(\mathbb{T}^2)$, *provided* $f \not\equiv \text{const}$.

REMARK 2. *The last inequality in the previous lemma can be used to show that*

$$N_{\Upsilon_2}(f) \le C\|\nabla f\|_2 + |\bar{f}|, \tag{18}$$

with $\bar{f} = \int_{\mathbb{T}^2} f(\cdot)$ *and some* $C > 0$ *independent of* f.

Lemma 1 can also be stated for Orlicz norms if the underlying Young functions are "comparable" with power functions. In particular, the next lemma holds.

LEMMA 9. *Let* $\mathbf{f} = (f_1, f_2) \in \left[C^\infty\left(\mathbb{T}^2\right)\right]^2$ *and* $\mathbf{F} = (F_1, F_2) = (\operatorname{div}\mathbf{f}, \operatorname{rot}\mathbf{f})$ *Then, for* $1 \le p < +\infty$, *there is a* $C = (p) > 0$ *such that*

$$N_{\Phi_p}(\nabla\mathbf{f}) \le C N_{\Phi_p}(\mathbf{F}).$$

Proof. We just have to notice that $\Phi_p(r)$ can be approximated by the linear combinations $\alpha r^2 + \beta r^3$, $r > 0$, and use Lemma 1 along with the definition of norm $N_{\Phi_p}(\cdot)$. □

Let us now assume that ρ_0 and $\mathbf{u}_0$ are $C^\infty\left(\mathbb{T}^2\right)$ functions. Let $\bar{\rho} = \int_{\mathbb{T}^2} \rho_0(\cdot)$ and $\bar{P} = P(\bar{\rho})$. By Theorem A, there exists the unique, global in time, *classical* solution $(\rho, \mathbf{u})$ of the problem (1)–(5). From (1) and (2) we derive equations needed for studying higher regularity of the solution. Note that in $\mathbb{R}^2$, $\operatorname{div}\mathbb{D} = \nabla\operatorname{div}\mathbf{u} + \frac{1}{2}\nabla^\perp \operatorname{rot}\mathbf{u}$ and system (2) can be written in the following form:

$$\frac{d\mathbf{u}}{dt} + \frac{1}{\rho}\nabla\left(P - \bar{P} - (4\mu/3 + \zeta)\operatorname{div}\mathbf{u}\right) - \frac{1}{\rho}\mu\nabla^\perp \operatorname{rot}\mathbf{u} = 0. \tag{19}$$

Introducing the shorthand notation

$$\left\{\begin{array}{lcl} A & = & \operatorname{rot}\mathbf{u}, \\ B & = & (4\mu/3 + \zeta)\operatorname{div}\mathbf{u} - P + \bar{P}, \end{array}\right. \tag{20}$$

and after applying differential operators rot and div to the above system, we obtain

$$\frac{dA}{dt} - (\partial_1\mathbf{u}\cdot\nabla)u_2 + (\partial_2\mathbf{u}\cdot\nabla)u_1 = \operatorname{rot}\frac{1}{\rho}\left(\nabla B + \mu\nabla^\perp A\right), \tag{21}$$

$$\frac{d}{dt}\operatorname{div}\mathbf{u} + (\partial_1\mathbf{u}\cdot\nabla)u_1 + (\partial_2\mathbf{u}\cdot\nabla)u_2 = \operatorname{div}\frac{1}{\rho}\left(\nabla B + \mu\nabla^\perp A\right). \tag{22}$$

Consequently, (22) can be written as

$$\begin{aligned} \frac{d}{dt}\left(\frac{B}{4\mu/3 + \zeta}\right) + (\partial_1\mathbf{u}\cdot\nabla)u_1 + (\partial_2\mathbf{u}\cdot\nabla)u_2 - \frac{d}{dt}\frac{P - \bar{P}}{4\mu/3 + \zeta} \\ = \operatorname{div}\frac{1}{\rho}\left(\nabla B + \mu\nabla^\perp A\right). \end{aligned} \tag{23}$$

The next equation is essential for the analysis of compressible, viscous flows and was studied in [7], [11], [13]. Let us apply operator $(-\Delta)^{-1}\operatorname{div}$ to system (2). With the notation

$$\psi = (-\Delta)^{-1}\operatorname{div}\rho\mathbf{u},\ \bar{B} = \int_{\mathbb{T}^2} B(t,\cdot), \tag{24}$$

we have

$$(4\mu/3 + \zeta)\operatorname{div}\mathbf{u} = \bar{B} + P - \bar{P} - \frac{d}{dt}\psi + \sum_{i,j=1,2}[u_i, R_iR_j](\rho u_j), \tag{25}$$

where we used the fact that $\mathbf{u}\cdot\nabla\psi-(-\Delta)^{-1}\operatorname{div}\operatorname{div}\rho\mathbf{u}\otimes\mathbf{u}$ can be written as a sum of commutators: $\sum_{i,j}[u_i,R_iR_j](\rho u_j)$. Using this in (1) we obtain

$$\frac{4\mu/3+\zeta}{\rho}\frac{d}{dt}\rho-\frac{d}{dt}\psi+P-\bar{P}=-\bar{B}+\sum_{i,j=1,2}[u_i,R_iR_j](\rho u_j). \tag{26}$$

Finally, by Calderón–Zygmund estimates of singular integral operators, for any $1<p<+\infty$, there is $C=C(p)>0$ such that

$$\|\nabla\psi(t,\cdot)\|_p\le C\|\rho(t,\cdot)\mathbf{u}(t,\cdot)\|_p, \tag{27}$$

and consequently, there is $C=C(p)$ such that

$$\|\nabla\psi(t,\cdot)\|_{\Phi_p}\le C\|\rho(t,\cdot)\mathbf{u}(t,\cdot)\|_{\Phi_p}. \tag{28}$$

In analyzing the system of equations, we encounter the problem of obtaining a "good" estimate for the term $\int_{\mathbb{T}^2}B\nabla u_1\cdot\nabla^\perp u_2\,dx$. One sees directly that it is bounded by $\|B\|_\infty\|\nabla\mathbf{u}\|_2^2$. Unfortunately, $\|B\|_\infty$ is not controlled by the $\|\nabla B\|_2$, the quantity naturally arising in energy estimates. This leads us to a more subtle estimate that takes into account the structure of the product term $\nabla u_1\cdot\nabla^\perp u_2$. The estimate is obtained by using the duality between Hardy $\mathcal{H}^1$ and $\mathcal{BMO}$ spaces. In [6], Fefferman proved that $\mathcal{BMO}$ is the dual space of $\mathcal{H}^1$. In particular, there is an absolute constant C such that

$$\left|\int_{\mathbb{T}^2}B\nabla u_1\cdot\nabla^\perp u_2\,dx\right|\le C||B||_{\mathcal{BMO}}||\nabla u_1\cdot\nabla^\perp u_2||_{\mathcal{H}^1}. \tag{29}$$

One readily notices that $\operatorname{rot}\nabla u_1=\operatorname{div}\nabla^\perp u_2=0$ and one can use the so-called "div-curl" lemma of compensated compactness; see [2, Theorem II.1]. It yields the following estimate with the suitable choice of an absolute constant C:

$$||\nabla u_1\cdot\nabla^\perp u_2||_{\mathcal{H}^1}\le C||\nabla u_1||_2||\nabla^\perp u_2||_2.$$

Finally, from Lemma 8 and definition of $||\cdot||_{\mathcal{BMO}}$ it follows that there is a constant $C>0$ such that

$$||B||_{\mathcal{BMO}}\le C||\nabla B||_2.$$

Combining the last two estimates in (29) we obtain

$$\left|\int_{\mathbb{T}^2}B\nabla u_1\cdot\nabla^\perp u_2\,dx\right|\le C||\nabla B||_2||\nabla\mathbf{u}||_2^2. \tag{30}$$

3. The first energy inequality. It follows from the equation of conservation of mass that $\rho(t,x)\ge0$ and $\int_{\mathbb{T}^2}\rho(t,\cdot)\,dx=\int_{\mathbb{T}^2}\rho_0(\cdot)\,dx=\bar{\rho}$. By multiplying (2) by $\mathbf{u}$, integrating over the domain Π, and using periodicity of the flow we obtain

$$\frac{d}{dt}\int_{\mathbb{T}^2}\rho\frac{|\mathbf{u}|^2}{2}\,dx+\int_{\mathbb{T}^2}(-P+(4\mu/3+\zeta)\operatorname{div}\mathbf{u})\operatorname{div}\mathbf{u}+\mu(\operatorname{rot}\mathbf{u})^2\,dx=0.$$

Since $P = a\rho^\gamma$, $\gamma > 1$ we deduce from (1) that $-(\gamma-1)P\operatorname{div}\mathbf{u} = (P)_t + \operatorname{div}(P\mathbf{u})$. Substituting this in the last equation we get

$$\sup_{t\in[0,+\infty)} \int_{\mathbb{T}^2} \rho(t,\cdot)\frac{|\mathbf{u}(t,\cdot)|^2}{2} + \frac{P(t,\cdot)}{\gamma-1}\,dx + \iint_{(0,+\infty)\times\mathbb{T}^2} (4\mu/3+\zeta)(\operatorname{div}\mathbf{u})^2 + \mu(\operatorname{rot}\mathbf{u})^2\,dxdt \le C\int_{\mathbb{T}^2} \rho_0(\cdot)\frac{|\mathbf{u}_0(\cdot)|^2}{2} + \frac{P(\rho_0(\cdot))}{\gamma-1}\,dx.$$

Let $E(t) = \int_{\mathbb{T}^2} \rho(t,\cdot)\frac{|\mathbf{u}(t,\cdot)|^2}{2} + \frac{P(t,\cdot)}{\gamma-1}\,dx$ denote the total energy of the gas and $E_0 = E(0)$. We will refer to the next inequality as the first energy inequality,

$$\sup_{t\in[0,+\infty)} E(t) + \iint_{(0,+\infty)\times\mathbb{T}^2} (4\mu/3+\zeta)(\operatorname{div}\mathbf{u})^2 + \mu(\operatorname{rot}\mathbf{u})^2\;\,dx\,dt \;\le\; E_0. \tag{31}$$

For the purpose of studying long time behavior of the solutions we will need a variant of the first energy inequality. Following [7] we introduce $G(\rho) = P(\rho) - P(\bar\rho) - P'(\bar\rho)(\rho-\bar\rho)$. Then, using the same arguments that lead to (31) we obtain

$$\frac{d}{dt}\int_{\mathbb{T}^2} \rho(t,x)\frac{|\mathbf{u}|(t,x)^2}{2} + G(\rho(t,x))\,dx + \int_{\mathbb{T}^2} (4\mu/3+\zeta(\rho(t,x)))|\operatorname{div}\mathbf{u}|^2(t,x) + \mu(\operatorname{rot}\mathbf{u})^2(t,x)\,dx = 0. \tag{32}$$

For the purpose of using imbedding theorems we will need bounds on averages of various functions. They are obtained in the following way. First,

$$\begin{aligned}\left|\int_{\mathbb{T}^2} (\mathbf{u}-\bar{\mathbf{u}})\,\rho\,dx\right| &\le \|\rho\|_\gamma \|\mathbf{u}-\bar{\mathbf{u}}\|_{\gamma/(\gamma-1)}\\ &\le C\|\nabla\mathbf{u}\|_2,\\ \left|\int_{\mathbb{T}^2} (\mathbf{u}-\bar{\mathbf{u}})\,\rho\,dx\right| &\ge |\bar{\mathbf{u}}|\int_{\mathbb{T}^2}\rho_0\,dx - \left|\int_{\mathbb{T}^2}\rho\mathbf{u}\,dx\right|\\ &= |\bar{\mathbf{u}}|\int_{\mathbb{T}^2}\rho_0\,dx - \left|\int_{\mathbb{T}^2}\rho_0\mathbf{u}_0\,dx\right|,\end{aligned}$$

where we used the fact that the momentum is conserved; $\frac{d}{dt}\int_{\mathbb{T}^2}\rho\mathbf{u}\,dx = 0$. Assume that $\int_{\mathbb{T}^2}\rho_0\mathbf{u}_0\,dx = 0$. This is a harmless assumption since system (1)–(2) is invariant under Galilean transformations, and it is satisfied in the reference frame (coordinate system) that moves with the constant velocity $\frac{\int_{\mathbb{T}^2}\rho_0\mathbf{u}_0\,dx}{\int_{\mathbb{T}^2}\rho_0\,dx}$ relative to the initially chosen frame. Then, since $\int_{\mathbb{T}^2}\rho(t)\,dx$ is positive and independent of time, we conclude that

$$|\bar{\mathbf{u}}| \le C\|\nabla\mathbf{u}\|_2. \tag{33}$$

By the definition of function A and periodicity of the flow,

$$\int_{\mathbb{T}^2} A(t,\cdot) = 0. \tag{34}$$

4. The second energy inequality.

REMARK 3. *We will prove the theorem only for the case $\beta \geq \gamma$. The case $\beta > 3$, $\gamma > 1$ is a straightforward adaptation of the proof if one takes into account the following estimate proved in* [8]*:*

$$\sup_{s\in(0,T)} \|\rho(s,\cdot)\|_p \leq C(T,p), \quad 1 \leq p < +\infty.$$

In this case all estimates become time dependent.

Let

$$\left.\begin{array}{rcl} Y^2(t) & = & \int_{\mathbb{T}^2} A^2(t,\cdot) + \frac{B^2(t,\cdot)}{\frac{4\mu}{3}+\zeta(t,\cdot)}, \\ X^2(t) & = & \int_{\mathbb{T}^2} \frac{(\nabla B(t,\cdot) + \mu\nabla^\perp A(t,\cdot))^2}{\rho(t,\cdot)}, \\ D^2(t) & = & \int_{\mathbb{T}^2} \left(\frac{4\mu}{3}+\zeta(t,\cdot)\right) |\operatorname{div}\mathbf{u}(t,\cdot)|^2 + \mu|\operatorname{rot}\mathbf{u}(t,\cdot)|^2 \end{array}\right\} \tag{35}$$

and $|\rho|(t) = 1 + \|\rho(t,\cdot)\|_\infty$.

LEMMA 10. *For any $\omega \in]0,1[$ there is $C = C(\omega,\mu,\beta,\gamma,b,E_0)$ such that*

$$\begin{aligned} \frac{d}{dt}Y^2(t) + X^2(t) \leq C_\omega \Big(& Y(t)D(t) + Y^2(t)D(t) \\ & + D(t)(Y(t)+D(t))X(t)|\rho|(t)^{\frac{1+\beta\omega}{2}} \Big). \end{aligned} \tag{36}$$

Proof. Multiplying (21) by $2A$ and using the fact that $-(\partial_1\mathbf{u}\cdot\nabla)u_2 + (\partial_2\mathbf{u}\cdot\nabla)u_1 = A\operatorname{div}\mathbf{u}$, we get

$$\frac{dA^2}{dt} + 2A^2\operatorname{div}\mathbf{u} = 2A\operatorname{rot}\frac{1}{\rho}(\nabla B + \mu\nabla^\perp A). \tag{37}$$

Multiplying (23) by $2B$, using the fact that

$$(\partial_1\mathbf{u}\cdot\nabla)u_1 + (\partial_2\mathbf{u}\cdot\nabla)u_2 = (\operatorname{div}\mathbf{u})^2 - 2\nabla u_1\cdot\nabla^\perp u_2,$$

and using the equation of mass conservation (1), we get

$$\begin{aligned} & \frac{d}{dt}\left(\frac{B^2}{4\mu/3+\zeta}\right) + \frac{B^2}{4\mu/3+\zeta}\operatorname{div}\mathbf{u} \\ & \quad - \frac{B^2}{4\mu/3+\zeta}\operatorname{div}\mathbf{u} + B^2\frac{d}{dt}\frac{1}{4\mu/3+\zeta} - 2B\frac{d}{dt}\frac{P-\bar{P}}{4\mu/3+\zeta} \\ & \quad + 2B(\operatorname{div}\mathbf{u})^2 - 4B\nabla u_1\cdot\nabla^\perp u_2 = 2B\operatorname{div}\frac{1}{\rho}(\nabla B + \mu\nabla^\perp A). \end{aligned} \tag{38}$$

Using identity (55) in (39) and adding (37) and (39), after the integration of the result

over $\mathbb{T}^2$, we get

$$\begin{aligned}(39)\quad &\frac{d}{dt}\left[\int_{\mathbb{T}^2} A^2+\frac{B^2}{4\mu/3+\zeta}\,dx\right]+2\int_{\mathbb{T}^2}\frac{(\nabla B+\mu\nabla^\perp A)^2}{\rho}\,dx\\ &\qquad=-\int_{\mathbb{T}^2} A^2\mathrm{div}\,\mathbf{u}\,dx+\int_{\mathbb{T}^2}-2B(\mathrm{div}\,\mathbf{u})^2+4B\nabla u_1\cdot\nabla^\perp u_2\\ &\quad-\int_{\mathbb{T}^2}\frac{(\beta-1)\zeta-4\mu/3}{(4\mu/3+\zeta)^2}B^2\mathrm{div}\,\mathbf{u}\,dx+\int_{\mathbb{T}^2}(P-\bar P)\frac{2\beta\zeta}{(4\mu/3+\zeta)^2}B\mathrm{div}\,\mathbf{u}\\ &\qquad-\frac{2\gamma P}{4\mu/3+\zeta}B\mathrm{div}\,\mathbf{u}\,dx+\int_{\mathbb{T}^2}(P-\bar P)\mathrm{div}\,\mathbf{u}\int_{\mathbb{T}^2}\frac{2(\gamma-1)B}{4\mu/3+\zeta}\,dx.\end{aligned}$$

By the Young inequality there is $C>0$ such that for all $t>0$

$$(40)\qquad Y^2(t)\le CD^2(t)+C\int_{\mathbb{T}^2}\frac{(P-\bar P)^2}{4\mu/3+\zeta}.$$

Also, from the energy inequality (31) and $\beta\ge\gamma$ it follows that there is $C>0$, independent of $t>0$, that verifies the inequality

$$(41)\qquad Y^2-C\le 2D^2(t)\le 4Y^2(t)+C.$$

Our goal will be to find the optimal estimate of the right-hand side of (39) in terms of Y, X, and $\|\rho\|_\infty$.

We will use the following result, based on Lemma 1, applied to the function $\mathbf{f}=(B,\mu A)$, and $\mathbf{F}=(\mathrm{div}\,\mathbf{f},\,\mathrm{rot}\,\mathbf{f})=\nabla B+\mu\nabla^\perp A$. The next estimate holds:

$$(42)\qquad \|\nabla A\|_2+\|\nabla B\|_2\le C\|\nabla B+\mu\nabla^\perp A\|_2.$$

Consider terms in (39). Using the estimate (31) and notation (20) we get

$$\begin{aligned}(43)\quad &\left|\int_{\mathbb{T}^2}\frac{B}{4\mu/3+\zeta}\,dx\right|\left|\int_{\mathbb{T}^2}P\mathrm{div}\,\mathbf{u}\,dx\right|\\ &\qquad\le C\left\|\frac{B}{\sqrt{4\mu/3+\zeta}}\right\|_2\|\sqrt{4\mu/3+\zeta}\mathrm{div}\,\mathbf{u}\|_2\le CY(t)D(t).\end{aligned}$$

The set of the following estimates is obtained by using the appropriate Young inequalities and the definition of B, (20):

$$\begin{aligned}(44)\qquad \int_{\mathbb{T}^2}|B|(\mathrm{div}\,\mathbf{u})^2\,dx&=\int_{\mathbb{T}^2}|B|\left|\frac{B+P-\bar P}{4\mu/3+\zeta}\right||\mathrm{div}\,\mathbf{u}|\,dx\\ &\le C\left\|\frac{B^2}{4\mu/3+\zeta}\right\|_2 D(t)+CY(t)D(t),\end{aligned}$$

$$(45)\qquad \int_{\mathbb{T}^2}\frac{B^2\zeta}{(4\mu/3+\zeta)^2}|\mathrm{div}\,\mathbf{u}|\,dx\le C\left\|\frac{B^2}{4\mu/3+\zeta}\right\|_2 D(t),$$

$$(46)\qquad \int_{\mathbb{T}^2}\frac{|B(P-\bar P)\mathrm{div}\,\mathbf{u}|\zeta}{(4\mu/3+\zeta)^2}\,dx\le CY(t)D(t),$$

$$(47)\qquad \int_{\mathbb{T}^2}\frac{|BP\mathrm{div}\,\mathbf{u}|}{4\mu/3+\zeta}\,dx\le CY(t)D(t).$$

Let us take $0 < \omega < 1$ and use the Hölder inequality to estimate the term

$$\left\| \frac{B^2}{4\mu/3+\zeta} \right\|_2 \le C \left\| \frac{B}{4\mu/3+\zeta} \right\|_2^{1-\omega} \| |B|^{1+\omega} \|_{\frac{2}{\omega}}. \tag{48}$$

Then, by Lemma 7, the Hölder inequality, and the estimate (42), we get with a number $K > 0$,

$$\begin{aligned} \| |B|^{1+\omega} \|_{\frac{2}{\omega}} &\le \left| \overline{|B|^{1+\omega} \chi_{\{\rho(t,\cdot)<K\}}} \right| + \| |B|^{1+\omega} - \overline{|B|^{1+\omega} \chi_{\{\rho(t,\cdot)<K\}}} \|_{\frac{2}{\omega}} \\ &\le \left| \overline{|B|^{1+\omega} \chi_{\{\rho(t,\cdot)<K\}}} \right| + C \| |B|^{\omega} \nabla B \|_{\frac{2}{1+\omega}} \\ &\le CY^{1+\omega} + C\|B\|_2^{\omega} \|\nabla B\|_2 \le CY^{1+\omega} + CY^{\omega} X |\rho|^{\frac{1+\beta\omega}{2}}, \end{aligned} \tag{49}$$

where we used the fact that $|\{\rho(t,\cdot) < K\}| > \frac{1}{2}$ when $K > 2 \int_{\mathbb{T}^2} \rho(t,\cdot) = 2 \int_{\mathbb{T}^2} \rho_0(\cdot)$. Substituting the last inequality in (48) we get

$$\left\| \frac{B^2}{4\mu/3+\zeta} \right\|_2 \le CY^2 + CYX |\rho|^{\frac{1+\beta\omega}{2}}, \tag{50}$$

where $C > 0$ depends upon $\omega \in (0,1)$ among other parameters.

Consider now the term $\int_{\mathbb{T}^2} |A|^2 \operatorname{div} \mathbf{u}$. By the same type of arguments that were used in derivation of the last estimates we obtain that

$$\begin{aligned} \left| \int_{\mathbb{T}^2} A^2 \operatorname{div} \mathbf{u} \right| &\le \|A^2\|_2 D(t) \\ &\le CY^2(t) D(t) + Y(t) D(t) X(t) |\rho|(t)^{\frac{1+\beta\omega}{2}}, \end{aligned} \tag{51}$$

with $\omega \in]0,1[$ and some $C > 0$. The term $\int_{\mathbb{T}^2} B \nabla u_1 \cdot \nabla^{\perp} u_2 \, dx$ is estimated by using the duality between $\mathcal{BMO}$ and Hardy $\mathcal{H}^1$ space, followed by "div-curl" lemma as it is explained in section 2,

$$\int_{\mathbb{T}^2} |B \nabla u_1 \cdot \nabla^{\perp} u_2| \, dx \le C \|\nabla B\|_2 \|\nabla u\|^2 \le C D^2(t) X(t) |\rho|^{\frac{1}{2}}(t). \tag{52}$$

Using estimate (43) in (45) and (45) and combining (43)–(47), (51), and (52) in (39) we get the inequality (36). □

We will use the estimate obtained in the following lemma.

LEMMA 11. *Let $\beta = \gamma$. There are $C = C(\mu, \beta, \gamma, b, E_0) > 0$ such that for all $t > 0$, it holds that (modulo a constant)*

$$\int_{\mathbb{T}^2} \frac{|P(t,\cdot) - \bar{P}|^2}{4\mu/3 + \zeta(t,\cdot)} \lesssim D^2(t) + \frac{d}{dt} \Psi(t), \tag{53}$$

where $|\Psi(t)| \le C$.

Proof. Equation (25) multiplied by $(P - \bar{P})(4\mu/3 + \zeta)^{-1}$ can be written as

$$\begin{aligned} \frac{(P-\bar{P})^2}{4\mu/3+\zeta} &= -\bar{B} \frac{P-\bar{P}}{4\mu/3+\zeta} - \psi \frac{d}{dt} \frac{P-\bar{P}}{4\mu/3+\zeta} \\ &\quad + \partial_t \frac{\psi(P-\bar{P})}{4\mu/3+\zeta} + \operatorname{div} \left(\frac{\psi(P-\bar{P})}{4\mu/3+\zeta} \mathbf{u} \right) \\ &\quad - \frac{\psi(P-\bar{P})}{4\mu/3+\zeta} \operatorname{div} \mathbf{u} + [\mathbf{u}, RR](\rho \mathbf{u}) \frac{P-\bar{P}}{4\mu/3+\zeta}, \end{aligned} \tag{54}$$

where we used the shorthand notation $[\mathbf{u}, RR](\rho\mathbf{u})$ to mean $\sum_{i,j}[u_i, R_iR_j](\rho u_j)$. Using the equation of mass conservation (1), we can write

$$\psi\frac{d}{dt}\frac{P-\bar{P}}{4\mu/3+\zeta} = \psi(P-\bar{P})\frac{\beta\zeta}{(4\mu/3+\zeta)^2}\operatorname{div}\mathbf{u} - \psi\frac{\gamma P}{4\mu/3+\zeta}\operatorname{div}\mathbf{u}.$$
$$+\psi\frac{(\gamma-1)}{4\mu/3+\zeta}\int_{\mathbb{T}^2}(P-\bar{P})\operatorname{div}\mathbf{u}\,dx. \tag{55}$$

Then, by the repeated use of Hölder inequalities, Lemma 7, the estimate (27), and (33) we derive the following inequalities:

$$\begin{aligned}\left|\int_{\mathbb{T}^2}\psi\frac{(P-\bar{P})\beta\zeta}{(4\mu/3+\zeta)^2}\operatorname{div}\mathbf{u}\,dx\right| &\le C\int_{\mathbb{T}^2}|\psi\operatorname{div}\mathbf{u}|\,dx \le C\|\psi\|_2\|\operatorname{div}\mathbf{u}\|_2\\ &\le C\|\nabla\psi\|_{1+\frac{\gamma-1}{2}}\|\operatorname{div}\mathbf{u}\|_2 \le C\|\rho\mathbf{u}\|_{1+\frac{\gamma-1}{2}}\|\operatorname{div}\mathbf{u}\|_2\\ &\le C\|\rho\|_\gamma\|\mathbf{u}\|_{\frac{\gamma(\gamma+1)}{\gamma-1}}\|\operatorname{div}\mathbf{u}\|_2 \le C\|\nabla\mathbf{u}\|_2^2.\end{aligned} \tag{56}$$

Similarly,

$$\left|\int_{\mathbb{T}^2}\psi\frac{\gamma P}{4\mu/3+\zeta}\operatorname{div}\mathbf{u}\,dx\right| \le C\|\nabla\mathbf{u}\|_2^2.$$

Collecting the above estimates in (55) we obtain

$$\left|\int_{\mathbb{T}^2}\psi\frac{d}{dt}\frac{P-\bar{P}}{4\mu/3+\zeta}\right| \le \frac{1}{8}\int_{\mathbb{T}^2}\frac{(P-\bar{P})^2}{4\mu/3+\zeta} + CD^2(t).$$

Since, by Hölder inequalities, Lemma 7, and the estimate (31),

$$\|\psi\|_1 \le C\|\nabla\psi\|_{\frac{2\gamma}{\gamma+1}} \le C\|\rho\mathbf{u}\|_{\frac{2\gamma}{\gamma+1}} \le C\|\sqrt{\rho}\mathbf{u}\|_2\|\rho\|_\gamma^{\frac{1}{2}} \le C(E_0),$$

we get that for $t \ge 0$,

$$\left|\int_{\mathbb{T}^2}\psi\frac{P(t,\cdot)-\bar{P}}{4\mu/3+\zeta(t,\cdot)}\,dx\right| \le C\|\psi(t,\cdot)\|_1 \le C \tag{57}$$

and

$$\left|\int_{\mathbb{T}^2}\psi\frac{P(t,\cdot)-\bar{P}}{4\mu/3+\zeta(t,\cdot)}\operatorname{div}\mathbf{u}\,dx\right| \le CD^2(t). \tag{58}$$

We denote $\Psi(t) = \int_{\mathbb{T}^2}\psi\frac{P(t,\cdot)-\bar{P}}{4\mu/3+\zeta(t,\cdot)}\,dx$.

By Lemmas 2 and 7 we obtain that

$$\begin{aligned}\|[\mathbf{u}, RR](\rho\mathbf{u})\|_1 &\le C\|[\mathbf{u}, RR](\rho\mathbf{u})\|_{\frac{2\gamma}{\gamma+1}} \le C\|\nabla\mathbf{u}\|_2\|\rho\mathbf{u}\|_{\frac{2\gamma}{\gamma+1}}\\ &\le C\|\nabla\mathbf{u}\|_2\|\mathbf{u}\|_{\frac{2\gamma}{\gamma-1}}\|\rho\|_\gamma \le C\|\nabla\mathbf{u}\|_2^2\|\rho\|_\gamma.\end{aligned}$$

In this way we obtain

$$\int_{\mathbb{T}^2}|[\mathbf{u}, RR](\rho\mathbf{u})|\frac{|P-\bar{P}|}{4\mu/3+\zeta} \le CD^2. \tag{59}$$

Finally, with Young's inequality we deduce that

$$\left|\int_{\mathbb{T}^2} -\bar{B}(t)\frac{P(t,\cdot)-\bar{P}}{4\mu/3+\zeta(t,\cdot)}\right| \le CD^2(t)+\frac{1}{8}\int_{\mathbb{T}^2}\frac{|P(t,\cdot)-\bar{P}|^2}{4/3\mu+\zeta(t,\cdot)}. \tag{60}$$

Collecting estimates (**??**)–(60) we conclude the proof. □

COROLLARY 1. *If $\rho(t,x)$ is uniformly bounded in space and time, using the fact that $G(\rho)\lesssim|P-\bar{P}|^2$, we obtain the following estimate:*

$$\int_0^{+\infty}\int_{\mathbb{T}^2} G(\rho(\cdot,\cdot)) \lesssim 1+\int_0^{+\infty} D^2(\cdot). \tag{61}$$

COROLLARY 2. *After dividing the inequality* (36) *by $e^2+Y^2(t)$ and integrating over time interval $]0,T[$, we deduce, using* (31), (40), (41), *and* (53), *that for $\omega\in]0,1[$ there is $C_\omega>0$ independent of T and such that*

$$\begin{aligned}\log(e^2+Y^2(T))+\int_0^T\frac{X^2(s)}{e^2+Y^2(s)} &\le \log(e^2+Y^2(0))\\ &+C_\omega\left(1+\sup_{[0,T]}|\rho|\int_0^T D^2(\cdot)+\int_0^T D^2(\cdot)|\rho|^{1+\omega\beta}(\cdot)\right).\end{aligned} \tag{62}$$

Note that from the first energy inequality (31) and (62) it follows that if $\rho(t,x)$ is bounded uniformly in space and time, then $Y(t)$ and consequently $D(t)$, the rate of energy dissipation, are bounded uniformly in time.

5. Uniform bounds on density. We use (26) to get pointwise bounds on $\rho(t,x)$. Let us fix time $t\in]0,+\infty[$ and estimate the $L^\infty(\mathbb{T}^2)$ norm of the terms on the right-hand side of this equation.

Consider the first term on the right-hand side of (26). We have

$$\begin{aligned}|\bar{B}(t)| &= \left|\overline{(4\mu/3+\zeta)\operatorname{div}\mathbf{u}}\right|\\ &\le \delta+C_\delta D(t)^2\end{aligned} \tag{63}$$

for any $\delta>0$ and some $C_\delta>0$. Lemma 6 and Remark 2 are used to get estimates on the commutator $[\mathbf{u},RR](\rho\mathbf{u})$. Specifically, in Lemma 6 we set $g(\cdot)=\mathbf{u}(t,\cdot)$, $f(\cdot)=\rho\mathbf{u}(t,\cdot)$, $p>4$, and $\omega=2$. Then, there is $C=C(p)$ such that

$$\begin{aligned}\|[\mathbf{u},RR](\rho\mathbf{u})\|_\infty &\le CN_{\Phi_p}(|\nabla\mathbf{u}|)N_{\Upsilon_2}(\rho\mathbf{u})\\ &\le C|\rho|_\infty N_{\Phi_p}(|\nabla\mathbf{u}|)N_{\Upsilon_2}(\mathbf{u}) \le C|\rho|_\infty N_{\Phi_p}(|\nabla\mathbf{u}|)D.\end{aligned} \tag{64}$$

Next, consider the term $N_{\Phi_p}(|\nabla\mathbf{u}|)$. By the estimate (28) we have

$$\begin{aligned}N_{\Phi_p}(|\nabla\mathbf{u}|) &\le CN_{\Phi_p}(|\operatorname{div}\mathbf{u}|)+CN_{\Phi_p}(|\operatorname{rot}\mathbf{u}|)\\ &= CN_{\Phi_p}(A)+CN_{\Phi_p}\left(\frac{B+P-\bar{P}}{4\mu/3+\zeta}\right) \le CN_{\Phi_p}(A)\\ &+CN_{\Phi_p}\left(\frac{B}{4\mu/3+\zeta}\right)+CN_{\Phi_p}\left(\frac{P-\bar{P}}{4\mu/3+\zeta}\right).\end{aligned} \tag{65}$$

From the first energy estimate (31) and condition $\beta\ge\gamma$ it follows that there is $C>0$ independent of time with the property

$$N_{\Phi_p}\left(\frac{P(t,\cdot)-\bar{P}}{4\mu/3+\zeta(t,\cdot)}\right)\le C. \tag{66}$$

The N_{Φ_p} norms of functions appearing in the above inequalities can be interpolated between the norms of L^2 and N_{Ψ_s}, where $\Psi_s(x)$ is defined in Lemma 4. Indeed, using Lemmas 3 and 4 with $r > p > 4$ and $2s > r$ we have

$$\begin{aligned} N_{\Phi_p}(A) &= N_{\Phi_p}\left(\frac{A}{\log^s(e^2+|A|)}\log^s(e^2+|A|)\right) \\ &\le N_{\Phi_r}\left(\frac{A}{\log^s(e^2+|A|)}\right) N_{\Psi_s}\left(\log^s(e^2+|A|)\right). \end{aligned} \tag{67}$$

Both norms appearing in the last inequality are estimated in the next lemma.

LEMMA 12. *Let p, r, s be chosen such that $4 < r < 2s < r+4$. There is $C = C(r,s,\mu)$ such that*

$$N_{\Phi_r}\left(\frac{A}{\log^s(e^2+|A|)}\right) \le C\left(1 + \frac{||A||_2}{[\log(e^2+||A||_2)]^{s-r/2}}\right),$$

$$N_{\Psi_s}\left(\log^s(e^2+|A|)\right) \le C\log^s(e^2+||\nabla A||_2).$$

Proof. By definition of N_{Φ_r},

$$N_{\Phi_r}\left(\tfrac{A}{\log^s(e^2+|A|)}\right) = \inf\left\{\lambda > 0 \,:\, \int_{\mathbb{T}^2} \tfrac{A^2}{\lambda^2\log^{2s}(e^2+|A|)}\log^r\left(e^2 + \tfrac{|A|}{\lambda\log^s(e^2+|A|)}\right) \le 1\right\}.$$

For any $\lambda > 1$,

$$\begin{aligned} &\int_{\mathbb{T}^2} \tfrac{A^2}{\lambda^2\log^{2s}(e^2+|A|)}\log^r\left(e^2 + \tfrac{|A|}{\lambda\log^s(e^2+|A|)}\right)dx \\ &\le \int_{\mathbb{T}^2} \tfrac{A^2}{\lambda^2\log^{2s-r}(e^2+|A|)}\,dx \le 2^{2s-r}\int_{\mathbb{T}^2} \tfrac{A^2}{\lambda^2\log^{2s-r}(e^2+|A|^2)}\,dx \\ &\le 2^{2s-r}\tfrac{1}{\lambda^2}\tfrac{y}{\log^{2s-r}(e^2+y)} \circ \int_{\mathbb{T}^2} A^2\,dx, \end{aligned}$$

where the last inequality is Jensen's inequality. Note that function

$$y\log^{-2s+r}(e^2+y)$$

for $y > 0$ and $2s < r+4$ is concave and $\mathrm{meas}(\mathbb{T}^2) = 1$. Thus, there is $C > 0$ such that

$$N_{\Phi_r}\left(\frac{A}{\log^s(e^2+|A|)}\right) \le 1 + C\frac{\left(\int_{\mathbb{T}^2} A^2\right)^{\frac{1}{2}}}{\log^{s-r/2}(e^2+\int_{\mathbb{T}^2} A^2)}.$$

Next, we have

$$N_{\Psi_s}\left(\log^s(e^2+|A|)\right) = \inf\left\{\lambda > 0 \,:\, \int_{\mathbb{T}^2}\left(\exp\exp\left(2\tfrac{\log(e^2+|A|)}{\lambda^{\frac{1}{s}}}\right) - e\right)dx \le 1\right\}.$$

Let c_1 be the constant from Lemma 8. Let

$$\lambda = 2^s N^s \log^s(e^2 + \sqrt{c_1}N\|\nabla A(t,\cdot)\|_2),$$

where $N > 1$ will be chosen later and

$$\mathcal{A} = \left\{x \in \mathbb{T}^2 \,:\, |A(t,x)| < \sqrt{c_1}N||\nabla A(t,\cdot)||_2\right\}.$$

Chebyshev's inequality and Lemma 8 are used to obtain the next bound,

$$|\mathbb{T}^2 \setminus \mathcal{A}| \leq \frac{||A||_1}{\sqrt{c_1}N||\nabla A||_2} \leq \frac{C}{N} \tag{68}$$

for some $C > 0$ independent of A. Then, for $x \in \mathbb{T}^2 \setminus \mathcal{A}$, it holds that

$$\frac{\log(e^2 + |A(t,x)|)}{\lambda^{\frac{1}{s}}} \leq \log\left(\frac{|A(t,x)|}{\sqrt{c_1}N\|\nabla A(t,\cdot)\|_2}\right). \tag{69}$$

Thus, we have

$$\int_{\mathbb{T}^2} \exp\exp\left(2\frac{\log(e^2+|A|)}{\lambda^{\frac{1}{s}}}\right) \;\leq\; \int_{\mathcal{A}} e^{e^{N^{-1}}} \;+\; \int_{\mathbb{T}^2\setminus\mathcal{A}} \exp\left(\frac{A^2}{c_1N^2\|\nabla A\|_2^2}\right). \tag{70}$$

Jensen's inequality and Lemma 7 can be used to show that (note that $\bar{A} = 0$, $|\mathbb{T}^2| = 1$)

$$\int_{\mathbb{T}^2\setminus\mathcal{A}} \exp\left(\frac{A^2}{c_1N^2\|\nabla A\|_2^2}\right) \leq |\mathbb{T}^2 \setminus \mathcal{A}|^{1-N^{-2}} \left(\int_{\mathbb{T}^2} \exp\left(\frac{A^2}{c_1\|\nabla A\|_2^2}\right)\right)^{N^{-2}}$$
$$\leq |\mathbb{T}^2 \setminus \mathcal{A}|^{1-N^{-2}} (c_2)^{N^{-2}}.$$

From this inequality, (70), and (68), it follows that

$$\int_{\mathbb{T}^2} \exp\exp\left(2\frac{\log(e^2+|A|)}{\lambda^{\frac{1}{s}}}\right) - e\,dx \leq e^{e^{N^{-1}}} + \left(\frac{C}{N}\right)^{1-N^{-2}} (c_2)^{N^{-2}} - e.$$

The right-hand side of the above inequality is smaller than 1 for N big enough (independently of A). With this choice of N we conclude that

$$N_{\Psi_s}\left(\log^s(e^2+|A|)\right) \leq \lambda(N).$$

This finishes the proof of the lemma. □

We apply this lemma to (67). For (p, r, s), such that $r + 4 > 2s > r > p > 4$, and a suitable $C = C(p, r, s, \mu) > 0$, it holds that

$$N_{\Phi_p}(A) \leq C\left(1 + \frac{\|A\|_2}{\log^{s-\frac{r}{2}}(e^2 + \|\nabla A\|_2)}\right)\log^s(e^2 + \|\nabla A\|_2^2). \tag{71}$$

The estimates for B are just slightly more involved because of the fact that B may not have a zero average over $\mathbb{T}^2$. With the same choice of (p, r, s) the following inequality holds:

$$N_{\Phi_p}\left(\frac{B}{4\mu/3+\zeta}\right) = N_{\Phi_p}\left(\frac{B}{(4\mu/3+\zeta)\log^s(e^2+|B|)}\log^s(e^2+|B|)\right)$$
$$\leq N_{\Phi_r}\left(\frac{B}{(4\mu/3+\zeta)\log^s(e^2+|B|)}\right)$$
$$\times \left(N_{\Psi_s}\left(\log^s(e^2+|B-\bar{B}|)\right) + C\log^s(e^2+|\bar{B}|)\right). \tag{72}$$

By arguments almost identical to those from the last lemma we get

$$N_{\Phi_p}\left(\frac{B}{4\mu/3+\zeta}\right) \leq C\left(1 + \frac{\|B/\sqrt{(4\mu/3+\zeta)}\|_2}{\log^{s-\frac{r}{2}}(e^2 + \|B/\sqrt{(4\mu/3+\zeta)}\|_2)}\right)$$
$$\times\, C\left(\log^s(e^2+\|\nabla B\|_2^2) + C\log^s(e^2+|\bar{B}|)\right). \tag{73}$$

By using notation (35) and estimates (42) we derive

$$\left.\begin{array}{lll} \|A\|_2 \le Y, & & \|B/\sqrt{(4\mu/3+\zeta)}\| \le Y, \\ \log^s(e^2+\|\nabla A\|_2^2) & \le & C\log^s\left(e^2+\frac{X^2}{(e^2+Y^4)(e^2+\|\rho\|_\infty)}\right) \\ & + & 12^s\log^s(e^2+Y^2) \\ & + & 6^s\log^s(e^2+|\rho|_\infty^{\frac{1}{2}}), \\ \log^s(e^2+\|\nabla B\|_2^2) & \le & C\log^s\left(e^2+\frac{X^2}{(e^2+Y^4)(e^2+\|\rho\|_\infty)}\right) \\ & + & 12^s\log^s(e^2+Y^2) \\ & + & 6^s\log^s(e^2+|\rho|_\infty^{\frac{1}{2}}). \end{array}\right\}$$

Systematic use of the set of the above inequalities and a Young inequality $xy \le Ce^{x^{s^{-1}}} + Cy\log^s(e^2+y)$, $x,y>0$, which is contained in Lemma 4, in (71) and (73) results in the following:

$$\begin{aligned} N_{\Phi_p}(A) + N_{\Phi_p}\left(\frac{B}{4\mu/3+\zeta}\right) \lesssim 1 + |\bar{B}| + |\rho|_\infty^{\frac{1}{2}} + Y\log^{\frac{r}{2}}(e^2+Y^2) \\ + \frac{X^2}{(e^2+Y^4)(e^2+|\rho|_\infty)}, \end{aligned} \tag{74}$$

where by $\lesssim$ we mean that the inequality holds modulo some positive, multiplicative constant C that depends on $(p,r,\mu,\gamma,b,\beta,E_0)$, but not time t (from now on we fix s in the interval $]r,r+4[$). Then, using the last inequality and (66) in (66), we get

$$\begin{aligned} N_{\Phi_p}(|\nabla\mathbf{u}|) \lesssim 1 + |\bar{B}| + |\rho|_\infty^{\frac{1}{2}} + Y\log^{\frac{r}{2}}(e^2+Y^2) \\ + \frac{X^2}{(e^2+Y^4)(e^2+|\rho|_\infty)}, \end{aligned}$$

with $r>p>4$. Now we can continue estimate (65) by making use of the last inequality, the inequality $xy \le \delta x^2 + y^2/(4\delta)$, $\delta>0$, and (**??**). We derive

$$\begin{aligned} \|[\mathbf{u},\,RR](\rho\mathbf{u})\|_\infty - \delta \lesssim D^2|\rho|_\infty + D^2|\rho|_\infty^2 + D^2|\rho|^{\frac{9}{4}} \\ + D^2|\rho|^{\frac{\beta-\gamma+2}{2}} + DY\log^{\frac{r}{2}}(e^2+Y^2)|\rho|_\infty + \frac{X^2}{e^2+Y^2}. \end{aligned} \tag{75}$$

The inequality holds modulo some positive multiplicative constant

$$C = C(\delta,p,r,\mu,\gamma,b,\beta,E_0).$$

(δ,p,r) will be set later. Upon using inequality (41) and the fact that $\beta>3$, we can write

$$\begin{aligned} \|[\mathbf{u},\,RR](\rho\mathbf{u})\|_\infty - \delta \lesssim D^2|\rho|^\beta + D^2\log^{\frac{r}{2}}(e^2+Y^2)|\rho| \\ + \frac{X^2}{e^2+Y^2}. \end{aligned} \tag{76}$$

Summarizing the result, we can write (26) and the estimates (63), (66), and (76) as

$$\left.\begin{aligned} &\tfrac{(4\mu/3+\zeta)}{\rho}\tfrac{d\rho}{dt} - \tfrac{d\psi}{dt} + P - \bar{P} = F, \\ &\sup_{\mathbb{T}^2} |F(t,\cdot)| \le \delta + C\left(D^2(t)|\rho|^\beta(t) + \tfrac{X^2(t)}{e^2+Y^2(t)}\right. \\ &\qquad \left. + D^2(t)\log^{\frac{r}{2}}(e^2+Y^2(t))|\rho|(t)\right), \\ &t>0,\ x\in\mathbb{T}^2, \quad C = C(\delta,p,r,\mu,\gamma,b,\beta,E_0). \end{aligned}\right\} \tag{77}$$

Consider the function $\psi(t,\cdot) = (-\Delta)^{-1}\mathrm{div}\,\rho(t,\cdot)\mathbf{u}(t,\cdot)$ that appears in (26). The elliptic estimate (28) that involves the Orlicz norm defined by the function $\Phi_k(x) = x^2\log^k(e^2+x)$, $k>1$, asserts that

$$N_{\Phi_k}\left(\nabla\psi(t,\cdot)\right) \le C N_{\Phi_k}\left(\rho(t,\cdot)\mathbf{u}(t,\cdot)\right). \tag{78}$$

Lemma 8 can be used to provide the following estimate on the $\sup|\psi|$ (note that $\overline{\psi(t,\cdot)} = 0$):

$$\sup_{\mathbb{T}^2}|\psi(t,\cdot)| \le C N_{\Phi_k}\left(\nabla\psi(t,\cdot)\right) \le C N_{\Phi_k}\left(\rho(t,\cdot)\mathbf{u}(t,\cdot)\right).$$

For some $2l > m > k$ we write

$$\begin{aligned} N_{\Phi_k}\left(\rho(t,\cdot)\mathbf{u}(t,\cdot)\right) &\le |\rho|^{\frac12} N_{\Phi_k}\left(\frac{\rho^{\frac12}|\mathbf{u}|}{\log^l(e^2+|\mathbf{u}|)}\log^l(e^2+|\mathbf{u}|)\right) \\ &\le |\rho|^{\frac12} N_{\Phi_m}\left(\frac{\rho^{\frac12}|\mathbf{u}|}{\log^l(e^2+|\mathbf{u}|)}\right) N_{\Psi_l}\left(\log^l(e^2+|\mathbf{u}|)\right), \end{aligned} \tag{79}$$

where Ψ_l was defined in Lemma 4. Proceeding with the same arguments as in the proof of Lemma 12 and using the first energy estimate (31) we obtain the following bounds for $2l > m > k$:

$$\begin{aligned} &N_{\Phi_m}\left(\tfrac{\rho^{\frac12}|\mathbf{u}|}{\log^l(e^2+|\mathbf{u}|)}\right) \\ &= \inf\left\{\lambda>0 \,:\, \textstyle\int_{\mathbb{T}^2} \tfrac{\rho|\mathbf{u}|^2}{\lambda^2\log^{2l}(e^2+|\mathbf{u}|)}\log^m\left(e^2+\tfrac{\rho^{\frac12}|\mathbf{u}|}{\lambda\log^l(e^2+|\mathbf{u}|)}\right)dx \le 1\right\}. \end{aligned}$$

For any $\lambda > 1$,

$$\begin{aligned} &\textstyle\int_{\mathbb{T}^2} \tfrac{\rho|\mathbf{u}|^2}{\lambda^2\log^{2l}(e^2+|\mathbf{u}|)}\log^m\left(e^2+\tfrac{\rho^{\frac12}|\mathbf{u}|}{\lambda\log^l(e^2+|\mathbf{u}|)}\right)dx \\ &\le \textstyle\int_{\mathbb{T}^2} \tfrac{\rho|\mathbf{u}|^2}{\lambda^2\log^{2l-m}(e^2+|\mathbf{u}|)}\,dx + \tfrac{\log^m(e^2+\|\rho\|_\infty^{\frac12})}{\lambda^2}\textstyle\int_{\mathbb{T}^2}\rho|\mathbf{u}|^2\,dx \\ &\le C\lambda^{-2}E_0|\rho|^{\frac12}. \end{aligned}$$

Thus, there is a constant $C(m,l,E_0)$ such that

$$N_{\Phi_m}\left(\frac{\rho^{\frac12}|\mathbf{u}|}{\log^l(e^2+|\mathbf{u}|)}\right) \le C|\rho|^{\frac14}. \tag{80}$$

Also, using the same arguments as in the proof of the second estimate of Lemma 12 and estimate (33) we obtain that

$$\begin{aligned} N_{\Psi_l}\left(\log^l(e^2+|\mathbf{u}-\mathbf{u}_{\mathbb{T}^2}|)\right) &\le N_{\Psi_l}\left(\log^l(e^2+|\mathbf{u}|)\right) + C\log^l(e^2+|\mathbf{u}_{\mathbb{T}^2}|) \\ &\le C\left(\log^l(e^2+|\mathbf{u}_{\mathbb{T}^2}|)+\log^l(e^2+\|\nabla\mathbf{u}\|_2)\right) \\ &\le C\log^l(e^2+\|\nabla\mathbf{u}\|_2). \end{aligned}$$

We now apply Lemma 1 with $\mathbf{f}=\mathbf{u}$ and use definition of A and B to get

$$\begin{aligned} \|\nabla\mathbf{u}\|_2 \;&\le\; C\left(\|\operatorname{div}\mathbf{u}\|_2+\|\operatorname{rot}\mathbf{u}\|_2\right) \le C\|\tfrac{B}{4\mu/3+\zeta}\|_2 + C\|A\|_2 \\ &+\; C\|\tfrac{P-\bar P}{4\mu/3+\zeta}\|_2 \le C\left(1+Y\right). \end{aligned}$$

And thus,

$$N_{\Psi_l}\left(\log^l(e^2+|\mathbf{u}|)\right) \le C\log^l(e^2+Y^2). \tag{81}$$

In the view of (80) and (81) the estimate (79) can be written as

$$\sup_{\mathbb{T}^2}|\psi(t,\cdot)| \le C|\rho|(t)\log^l(e^2+Y^2(t)),$$

and consequently, using the fact that $\beta>3$ we obtain

$$\sup_{\mathbb{T}^2}|\psi(t,\cdot)| \le \epsilon|\rho|^\beta(t)+C_\epsilon\log^{2l}(e^2+Y^2), \quad 2l>1, \tag{82}$$

where C_ϵ depends on l, ϵ, and other parameters, but not time.

Let us go back to (26). Let $\theta=\theta(\rho)$ be the solution of the Cauchy problem

$$\frac{d\theta}{d\rho} = \left(\frac{4\mu}{3}+\zeta(\rho)\right)\rho^{-1}, \quad \theta(1)=1.$$

$\theta(\rho)$ is uniquely defined and smooth for $\rho\in(0,+\infty)$. In fact,

$$\theta = 4\mu/3\log\rho + b\beta^{-1}\rho^\beta + 1 - b\beta^{-1}. \tag{83}$$

Let $\bar\theta=\theta\circ\left(a^{-2}\bar P\right)^{\frac{1}{\gamma}}$. Let $X_t(x_0)$ be a trajectory of the flow generated by $\mathbf{u}$, i.e.,

$$\frac{d}{dt}X_t(x_0) = \mathbf{u}\left(t,\, X_t(x_0)\right), \quad X_t(x_0)=x_0, \quad t=0,\, x_0\in\mathbb{T}^2,\, t>0,$$

and consider (26) at points $(t,\,x)=(t,\,X_t(x_0))$. Let $w(t)=\theta\circ\rho(t,X_t(x_0))-\bar\theta$. Then $P(\rho(t,x))-\bar P=\frac{a^2\gamma\rho_1^\gamma}{4\mu/3+\zeta(\rho_1)}w(t)$, where ρ_1 is a value in the interval with endpoints $(a^{-2}\bar P)^{\frac{1}{\gamma}}$ and $\rho(t,\,X_t(x_0))$. Let $S=\{t\,|\,w(t)>0\}$. Note that $\bar P=a^2\bar\rho^\gamma$, where $\bar\rho$ is the total mass of the gas, and so the inequality $w(t)>0$ implies $\rho(t,X_t(x_0))\ge\bar\rho$. For $t\in S$, factor $e^{-1}(t)\equiv\frac{4\mu/3+\zeta(\rho_1)}{a\gamma\rho_1^\gamma}$ is bounded by $c|\rho|^{\beta-\gamma}(t)$ with some constant c. Set S is open and consists of a countable disjoint union of intervals. Let $(\alpha,\,\alpha_1)$ be one

of those intervals. After multiplying the equation in (77) by $\exp\int_\alpha^t e(t)\,d\tau$ we get

$$\begin{aligned}\left(w(t)\exp\int_\alpha^t e(\tau)\,d\tau\right)' &\le \frac{d}{dt}\left[(\psi-\bar\theta)\exp\int_\alpha^t e(\tau)\,d\tau\right]\\ &\quad-(\psi-\bar\theta)\frac{d}{dt}\exp\int_\alpha^t e(\tau)\,d\tau\\ &\quad+\delta\exp\int_\alpha^t e(\tau)\,d\tau+\left(\sup_{\mathbb{T}^2}|F(t,\cdot)|-\delta\right)\exp\int_\alpha^t e(\tau)\,d\tau,\end{aligned}$$

where $\delta>0$ will be chosen later. Integrate above inequality over $(\alpha,\,t)$, $t<\alpha_1$ to get

$$\begin{aligned}w(t)\le C(\delta)|\rho|^{\beta-\gamma}(t)+4\sup_{(0,\,t)\times\mathbb{T}^2}|\psi(\cdot,\cdot)|+4|\bar\theta|\\ +\int_\alpha^t\max\{|F(\tau)|-\delta,0\}\,d\tau.\end{aligned}\tag{84}$$

Note that the right-hand side does not depend on x_0. Using the estimate (82) in (84) we can write for $t>0$

$$\begin{aligned}|\rho|^\beta(t)\le C\Bigg(1+\sup_{(0,\,t)}\log^{2l}(e^2+Y^2(\cdot))\\ +\int_0^t\max\{|F(\tau)|-\delta,0\}\,d\tau\Bigg),\ 2l>1.\end{aligned}\tag{85}$$

We use the estimates for $|F|$ obtained in (77) with $\delta=1$ and the bounds for $\log(e^2+Y^2)$ and $\int_0^t\frac{X^2}{e^2+Y^2}$ contained in the second energy inequality (62). Choose r and ω such that $4<r<2(\beta-1)$ and $0<r\beta\omega<2(\beta-1)-r$. This is possible only when $\beta>3$. Then, we can derive the following estimate for $t>0$, using the fact that $\int_0^{+\infty}D^2(\cdot)\le E_0$:

$$\begin{aligned}\int_0^t\max\{|F(t)|-1,0\}\,dt\le C\sup_{[0,t]}|\rho|^{1+\frac{r}{2}}\\ +C\int_0^t D^2(s)\sup_{[0,s]}|\rho|^\beta(s)\,ds.\end{aligned}\tag{86}$$

Returning to (85) we have

$$\begin{aligned}|\rho|^\beta(t)\le C\sup_{[0,t]}|\rho|^{1+\frac{r}{2}}+C\int_0^t D^2\sup_{[0,s]}|\rho|^\beta\\ +C\left[\int_0^t D^2\left(\sup_{[0,s]}|\rho|^{\frac{\beta-1}{2}}+\sup_{[0,s]}|\rho|^{\beta-\gamma}\right)\right]^{2l},\end{aligned}$$

where $\int_0^{+\infty}D^2\le E_0$. Finally, let us choose l that satisfies conditions $1<2l<\frac{\beta}{\beta-\gamma}$ and $l(\beta-\gamma)<\beta$. Then, using Gronwall type estimates we obtain the global in time upper bound for density,

$$\|\rho(t,\cdot)\|_\infty\le\bar M,\quad t>0.\tag{87}$$

Now we revisit the second energy estimate (62) to derive global in time bounds on $\int_{\mathbb{T}^2} |\nabla \mathbf{u}|^2(t,\cdot)$. It follows from (62) and (87) that

$$Y^2(t) + \int_0^t X^2(s)\, ds \le C, \quad t > 0, \tag{88}$$

with some $C > 0$ independent of time.

Let us now derive the lower bound on density. The equation of evolution of density (77) can be used to obtain the inequality

$$\frac{d}{dt}\theta \ge \frac{d}{dt}\psi + a^2(\bar{M})^\gamma - |F|, \tag{89}$$

where θ is defined by (83). Inequality (82) together with estimates (87) and (88) imply that

$$\sup_{]0,+\infty[\times\mathbb{T}^2} |\psi(\cdot,\cdot)| \le C. \tag{90}$$

The estimate for $|F|$ is contained in (77), where we set $\delta = a^2(\bar{M})^\gamma$. We integrate inequality (89) along a trajectory, and after using estimates (87) and (88) we obtain that

$$\theta(\rho(t, X_t(x_0)) \ge \theta(\rho_0(x_0)) - C, \quad t \in]0,+\infty[, \quad x_0 \in \mathbb{T}^2, \tag{91}$$

with $C > 0$ independent of (t, x_0). Since $\theta(\rho)$ behaves like a $\log \rho$ for small $\rho > 0$ and $\rho_0(x_0) > m > 0$ we conclude the existence of $\bar{m} > 0$ such that

$$\rho(t,x) \ge \bar{m}, \quad (t,x) \in]0,+\infty[\times\mathbb{T}^2. \tag{92}$$

Let us prove the decay estimate (7). First, we prove that the kinetic energy $\int_{\mathbb{T}^2} \rho(t,\cdot)|\mathbf{u}|^2(t,\cdot)$ decays to 0 with time. Indeed, by Lemma 7 and estimates (33) and (87) we can write

$$C\int_{\mathbb{T}^2} \rho(t,\cdot)|u|^2(t,\cdot) \le D^2(t), \quad t > 0,$$

for some $C > 0$ independent of time. We use the last inequality in (32) to get

$$\frac{d}{dt}\int_{\mathbb{T}^2} \rho(t,\cdot)\frac{|\mathbf{u}|^2(t,\cdot)}{2} + C\int_{\mathbb{T}^2} \rho\frac{|\mathbf{u}|^2(t,\cdot)}{2} \le -\frac{d}{dt}\int_{\mathbb{T}^2} G(\rho(t,\cdot)), \quad t > 0.$$

Integrating this inequality we get

$$\begin{aligned}\int_{\mathbb{T}^2} \rho(t_2,\cdot)\frac{|\mathbf{u}|^2(t_2,\cdot)}{2} &\le e^{-C(t_2-t_1)}E_0 \\ &\quad + C\int_{t_1}^{t_2}\int_{\mathbb{T}^2} G(\cdot,\cdot), t_2 > t_1 > 0.\end{aligned} \tag{93}$$

Consider the equation in (77). With notation $w = \theta(\rho) - \theta(\bar{\rho})$, we get some $\rho_1 \in (\bar{m}, \bar{M})$, such that

$$P(\rho) - \bar{P} = \frac{a^2\gamma(\rho_1)^\gamma}{4\mu/3 + \zeta(\rho_1)} w \triangleq m(t,x_0)w.$$

In view of uniform bounds (87) and (92) there exist $C, c > 0$ independent of time t and position x_0 such that

$$c \le m(t, x_0) \le C.$$

The equation in (77) can be integrated to yield

$$\begin{aligned} w(t_2, x_0) = {} & e^{-\int_{t_1}^{t_2} m(\cdot), x_0}(w(t_1, x_0) - \psi(t_1, x_0)) + \psi(t_2, X_{t_2}(x_0)) \\ & + \int_{t_1}^{t_2} \psi(s, X_s(x_0)) m(s, x_0) e^{-\int_s^{t_2} m(\cdot, x_0)}\, ds \\ & + \int_{t_1}^{t_2} F(s, X_s(x_0)) e^{-\int_s^{t_2} m(\cdot, x_0)}\, ds. \end{aligned} \tag{94}$$

In the view of uniform bounds on $m(t, x_0)$ stated above and bounds on ρ we derive

$$\begin{aligned} |w(t_2, x_0)| \le {} & e^{-c(t_2 - t_1)} w|(t_1, x_0)| + C\delta + C \sup_{s \in (t_1, t_2)} \|\psi(s, \cdot)\|_\infty \\ & + \int_{t_1}^{t_2} \max\{|F(s, X_s(x_0))| - \delta,\, 0\}|. \end{aligned} \tag{95}$$

The estimate for $\max\{|F| - \delta,\, 0\}|$ is given by (77). Employing bounds (87), (88) we derive

$$\int_{t_1}^{t_2} \max\{|F| - \delta,\, 0\}| \le C_\delta \int_{t_1}^{t_2} D^2(\cdot) + X^2(\cdot).$$

Using Lemma 7, estimate (27), the Hölder inequality, and uniform in time bounds (87), (88) we can write

$$\begin{aligned} \|\psi(t, \cdot)\|_\infty \le C \|\nabla \psi(t, \cdot)\|_4 \le C \|\rho(t, \cdot) \mathbf{u}(t, \cdot)\|_4 \\ \le C \|\sqrt{\rho(t, \cdot)} \mathbf{u}(t, \cdot)\|_2^{\frac{1}{3}} \|\mathbf{u}(t, \cdot)\|_4^{\frac{2}{3}} \le C \|\sqrt{\rho(t, \cdot)} \mathbf{u}(t, \cdot)\|_2^{\frac{1}{3}}, \end{aligned}$$

with C independent of t. Collecting the last two estimates in (95) we get

$$\begin{aligned} \|w(t_2, \cdot)\|_\infty \le {} & e^{-c(t_2 - t_1)} \|w(t_1, \cdot)\|_\infty + C\delta \\ & + C \sup_{s \in (t_1, t_2)} \|\sqrt{\rho}(s, \cdot) \mathbf{u}(s, \cdot)\|_2^{\frac{1}{3}} + C_\delta \int_{t_1}^{t_2} D^2(\cdot) + X^2(\cdot). \end{aligned} \tag{96}$$

The last inequality and (93) were derived for the approximate smooth solution $(\rho, \mathbf{u})$, but in the view of the convergence results that will be established in the next chapter these inequalities are also true for the weak solution itself. Moreover, for a weak solution, as it follows from (31), (61), and (62), it holds that

$$D^2,\, X^2,\, \int_{\mathbb{T}^2} G(\rho(\cdot, x))\, dx \in L^1\left(]0, +\infty[\right),$$

and consequently (93) and (96) imply that

$$\|\theta(\rho(t, \cdot)) - \theta(\bar{\rho})\|_\infty + \|\sqrt{\rho}(t, \cdot) \mathbf{u}(t, \cdot)\|_2 \to 0, \quad t \to +\infty,$$

which in turn implies (7) since ρ is bounded from below and above as well as $\|\nabla \mathbf{u}(t, \cdot)\|_2$.

6. Convergence of approximate solutions. In this section we consider the limit of classical solutions, for which bounds derived in the last section hold. These solutions solve the problem with smooth initial datum $(\rho_0^n,\, \mathbf{u}_0^n)$, $n > 0$, that converges strongly to $(\rho_0,\, \mathbf{u}_0)$.

Let $(\rho_0^n,\, \mathbf{u}_0^n) \in C^\infty\left(\mathbb{T}^2\right) \times \left[C^\infty\left(\mathbb{T}^2\right)\right]^2$ be such that

$$\begin{aligned} m < \rho_0^n(x) < M, \quad & x \in \mathbb{T}^2, \\ \rho_0^n \to \rho_0 \quad & \text{in } L^2(\mathbb{T}^2), \\ \mathbf{u}_0^n \to \mathbf{u} \quad & \text{in } \left[W^{1,2}\left(\mathbb{T}^2\right)\right]^2, \\ & n \to +\infty. \end{aligned}$$

Let $(\rho^n,\, \mathbf{u}^n)$ be the classical solution of the problem with $(\rho_0^n,\, \mathbf{u}_0^n)$ as the initial datum that exists by Theorem A. The bounds obtained in the previous section apply for these solutions so that

$$\begin{aligned} \mathbf{u}^n \text{ is bounded in } & L^\infty \cap L^2\left((0,T);\, \left[W^{1,2}\left(\mathbb{T}^2\right)\right]^2\right), \\ A^n,\, B^n \text{ is bounded in } & L^2\left((0,T);\, W^{1,2}\left(\mathbb{T}^2\right)\right), \\ \text{bounded in } & L^\infty\left((0,T);\, L^2(\mathbb{T}^2)\right), \\ \bar{m} < \rho^n(t,x) < \bar{M}, \quad & (t,x) \in (0,T) \times \mathbb{T}^2, \\ & T > 0, \end{aligned}$$

for some $\bar{m}$, $\bar{M}$ that are independent of T in the case $\beta \geq \gamma$. Here $A^n = \operatorname{rot} \mathbf{u}^n$, $B^n = (4\mu/3 + \zeta(\rho^n))\operatorname{div} \mathbf{u}^n$. Moreover, using the system of equations (19), (21), (23) and Lemma 8 we deduce

$$\begin{aligned} \partial_t \rho^n \text{ is bounded in } & L^2\left((0,T);\, W^{-1,2}\left(\mathbb{T}^2\right)\right), \\ \partial_t \mathbf{u}^n \text{ is bounded in } & L^2\left((0,T);\, \left[W^{-1,2}\left(\mathbb{T}^2\right)\right]^2\right), \\ \partial_t A^n,\, \partial_t B^n \text{ is bounded in } & L^2\left((0,T);\, W^{-1,\frac{4}{3}}\left(\mathbb{T}^2\right)\right), \\ & T > 0. \end{aligned}$$

The embedding $W^{1,2}\left(\mathbb{T}^2\right) \subset L^2\left(\mathbb{T}^2\right)$ is compact and $L^2\left(\mathbb{T}^2\right) \subset W^{-1,\frac{4}{3}}\left(\mathbb{T}^2\right)$ (or $W^{-1,2}\left(\mathbb{T}^2\right)$) is continuous. Aubin's lemma can be used; see [16, Theorem 2.1, Chapter III] to conclude that there are

$$\begin{aligned} \mathbf{u} \in L^\infty \cap L^2\left((0,T);\, \left[W^{1,2}\left(\mathbb{T}^2\right)\right]^2\right), \\ A,\, B \in L^2\left((0,T);\, W^{1,2}\left(\mathbb{T}^2\right)\right) \cap L^\infty\left((0,T);\, L^2(\mathbb{T}^2)\right), \end{aligned}$$

such that

$$\begin{cases} \mathbf{u}^n \to \mathbf{u} & \text{in } L^2\left((0,T);\, \left[L^2(\mathbb{T}^2)\right]^2\right), \\ A^n \to A & \text{in } L^2\left((0,T);\, L^2(\mathbb{T}^2)\right), \\ B^n \to B & \text{in } L^2\left((0,T);\, L^2(\mathbb{T}^2)\right). \end{cases} \tag{97}$$

From the uniform boundedness of ρ^n we conclude that there is $\rho(t,x)$ such that

$$\begin{aligned} \bar{m} \leq \rho(t,x) \leq \bar{M}, & (t,x) \in (0,T) \times \mathbb{T}^2, \\ \rho^n \to \rho, \text{ * - weakly in } & L^\infty\left((0,T) \times \mathbb{T}^2\right). \end{aligned} \tag{98}$$

This information is sufficient for passing to the limit in (1). We obtain

$$\begin{cases} \partial_t \rho + \operatorname{div} \rho \mathbf{u} = 0 & \text{in } \mathcal{D}'\left((0,+\infty)\times \mathbb{T}^2\right), \\ \rho(0,x) = \rho_0(x) & \text{a.e. } x \in \mathbb{T}^2. \end{cases} \tag{99}$$

Also, since $\rho \in L^\infty$, ρ is a renormalized solution of (2), i.e., for any $b = b(z) \in C^\infty(\mathbb{R})$ it holds that

$$\begin{cases} \partial_t b(\rho) + \operatorname{div} b(\rho)\mathbf{u} + (\rho b'(\rho) - b(\rho))\operatorname{div} \mathbf{u} = 0, & \mathcal{D}'\left((0,+\infty)\times \mathbb{T}^2\right), \\ b(\rho(0,x)) = b(\rho_0(x)) & \text{a.e. } x \in \mathbb{T}^2, \end{cases} \tag{100}$$

where $b'(z)$ is the derivative of b. Also,

$$\rho \in C\left([0,+\infty); L^p\left(\mathbb{T}^2\right)\right), \quad 1 < p < +\infty$$

(see the proof of Theorem 2.4 in [10]).

Due to the presence of nonlinear terms $\zeta(\rho)\operatorname{div} \mathbf{u}$ and $P(\rho)$, convergence results established in (97) and (98) are not enough to conclude that $(\rho, \mathbf{u})$ solves (2) in the weak sense. The missing ingredient is a.e. convergence of ρ_n. If we assume for a moment that, additionally to (97) and (98),

$$\rho^n \to \rho \text{ a.e. in } (0,+\infty)\times \mathbb{T}^2, \quad n \to +\infty, \tag{101}$$

holds, then it is easy to see that

$$\begin{aligned} \zeta(\rho^n)\operatorname{div} \mathbf{u}^n \to \zeta(\rho)\operatorname{div} \mathbf{u} \quad & \text{weakly in } L^2\left((0,T)\times \mathbb{T}^2\right), \\ P(\rho^n) \to P(\rho) \quad & \text{in } L^2\left((0,T)\times \mathbb{T}^2\right), \quad T > 0, \end{aligned}$$

and (ρ, u) is a weak solution of (1)–(5).

Let us prove (101). Let $b(z)$ be as above. The sequence $b(\rho^n)$ is uniformly bounded and so there is $\overline{b} \in L^\infty\left((0,T)\times \mathbb{T}^2\right)$, $T > 0$ such that $b(\rho^n) \to \overline{b}$ * - weakly in this space. Also we can write

$$\operatorname{div} \mathbf{u}^n = \frac{B^n + P(\rho^n)_{\mathbb{T}^2}}{4\mu/3 + \zeta(\rho^n)} - \frac{P(\rho^n)}{4\mu/3 + \zeta(\rho^n)},$$

and thus

$$\operatorname{div} \mathbf{u} = (B + \bar{P}_{\mathbb{T}^2})\overline{\left(\frac{1}{4\mu/3 + \zeta(\rho)}\right)} - \overline{\left(\frac{P(\rho)}{4\mu/3 + \zeta(\rho)}\right)}, \tag{102}$$

where $^-$ stays for the * - weak limit of the functions and the equality holds a.e. in space and time. Let $b(z) = z^2$ and $\nu = (4\mu/3 + \zeta(\rho))^{-1}$. The pair $(\rho^n, \mathbf{u}^n)$ satisfies (1) in the classical sense and so it holds that

$$\begin{cases} \partial_t(\rho^n)^2 + \operatorname{div}(\rho^n)^2\mathbf{u}^n \\ \qquad = -(\rho^n)^2 \operatorname{div} \mathbf{u}^n \\ \qquad = -\frac{(B^n + P(\rho^n)_{\mathbb{T}^2})(\rho^n)^2}{4\mu/3 + \zeta(\rho^n)} + \frac{(\rho^n)^2 P(\rho^n)}{4\mu/3 + \zeta(\rho^n)}, \\ \\ (\rho^n)^2(0,x) = (\rho_0^n)^2(x), \quad x \in \mathbb{T}^2. \end{cases} \tag{103}$$

Taking the limit in the above equation, in the view of (97) and (98), we obtain (note that initial datum converges strongly)

$$\begin{cases} \partial_t \overline{\rho^2} + \operatorname{div} \overline{\rho^2}\mathbf{u} &= -(B + \overline{P}_{\mathbb{T}^2})\overline{\rho^2\nu} + \overline{\rho^2 P(\rho)\nu} \\ & \quad \text{in } \mathcal{D}'\left((0,+\infty)\times\mathbb{T}^2\right), \\ \overline{\rho^2}(0,x) = \rho_0^2(x) & \quad \text{a.e. } x \in \mathbb{T}^2. \end{cases} \tag{104}$$

Also, from (100) and (102) it follows that

$$\begin{cases} \partial_t \rho^2 + \operatorname{div} \rho^2\mathbf{u} &= -(B + \overline{P}_{\mathbb{T}^2})\rho^2\overline{\nu} + \rho^2\overline{P(\rho)\nu} \\ & \quad \text{in } \mathcal{D}'\left((0,+\infty)\times\mathbb{T}^2\right), \\ \rho^2(0,x) = \rho_0^2(x) & \quad \text{a.e. } x \in \mathbb{T}^2. \end{cases} \tag{105}$$

Setting $\psi = \overline{\rho^2} - \rho^2 \geq 0$, and subtracting the last two equations we obtain

$$\begin{cases} \partial_t \psi + \operatorname{div} \psi\mathbf{u} &= (B + \overline{P}_{\mathbb{T}^2})\left(\rho^2\overline{\nu} - \overline{\rho^2\nu}\right) - \left(\rho^2\overline{P\nu} - \overline{\rho^2 P\nu}\right) \\ & \quad \text{in } \mathcal{D}'\left((0,+\infty)\times\mathbb{T}^2\right), \\ \psi(0,x) = 0 & \quad \text{a.e. } x \in \mathbb{T}^2. \end{cases} \tag{106}$$

By writing $(\rho^n)^2 - \rho^2 = 2\rho(\rho^n - \rho) + (\rho^n - \rho)^2$, we see that

$$\lim_{n\to\infty} \int_{\mathbb{T}^2} \phi(\cdot)(\rho^n(t,\cdot) - \rho(t,\cdot))^2\, dx = \int_{\mathbb{T}^2} \phi(\cdot)\psi(t,\cdot),$$

$t > 0$, for any test function ϕ. Also, for b—a smooth function—we have

$$\begin{aligned} (\rho^n)^2 b(\rho^n) - \rho^2 b(\rho^n) &= (\rho^n)^2 b(\rho^n) - \rho^2 b(\rho) - \rho^2(b(\rho^n) - b(\rho)) \\ &= (\rho^2 b(\rho))'\Big|_{\rho(t,x)}(\rho^n - \rho) + (\rho^2 b(\rho)/2)''\Big|_{\tilde{\rho}^n}(\rho^n - \rho)^2 \\ &\quad - \rho^2\left(b(\rho)'\Big|_{\rho(t,x)}(\rho^n - \rho) + (b(\rho)/2)''\Big|_{\hat{\rho}^n}(\rho^n - \rho)^2\right) \end{aligned}$$

for some $\tilde{\rho}^n$ and $\hat{\rho}^n$ in between ρ^n and ρ. Multiplying the above equality by $B + \overline{P}_{\mathbb{T}^2}$, integrating over domain $\mathbb{T}^2$, taking the limit $n \to +\infty$, and using the uniform boundedness of ρ^n we obtain

$$\begin{aligned} &\left|\int_{\mathbb{T}^2} (B(t,\cdot) + \overline{P}_{\mathbb{T}^2}(t))\left(\rho^2\overline{\nu}(t,\cdot) - \overline{\rho^2\nu}(t,\cdot)\right)\right| \\ &\qquad \leq C|\overline{P}_{\mathbb{T}^2}(t) + \bar{B}(t)| \int_{\mathbb{T}^2} \psi(t,\cdot) + C\int_{\mathbb{T}^2} |B - \bar{B}|(t,\cdot)\psi(t,\cdot) \end{aligned}$$

for some constant $C > 0$. By estimating the second term on the right-hand side in (106) in similar fashion and integrating the equation in (106) over $\mathbb{T}^2$ we derive the inequality

$$\begin{cases} \frac{d}{dt}\int_{\mathbb{T}^2} \psi(t,\cdot) \leq C\left(1 + |\bar{B}|(t)\int_{\mathbb{T}^2}\psi(t,\cdot) + \int_{\mathbb{T}^2}|B - \bar{B}|(t,\cdot)\psi(t,\cdot)\right) \\ \qquad\qquad \text{in } \mathcal{D}'\left((0,+\infty)\right), \\ \int_{\mathbb{T}^2}\psi(0,x) = 0. \end{cases}$$

We use the Hölder inequality for Orlicz spaces defined by functions $\exp(x^2)-1$ and $x\log^{\frac{1}{2}}(e^2+x)$ followed by Remark 2 to get the estimate

$$\begin{aligned}\int_{\mathbb{T}^2}|B-\bar{B}|\psi &\le N_{e^{x^2}-1}\left(B-\bar{B}\right)N_{x\log^{\frac{1}{2}}(e^2+x)}(\psi)\\ &\le C||\nabla B||_2\Gamma\left(\int_{\mathbb{T}^2}\psi\right),\end{aligned}$$

where $\Gamma(r)$, $r\ge 0$ is the inverse function of $\frac{r}{\log^{1/2}(e^2+c/r)}$, with some $c>0$ that bounds $\int_{\mathbb{T}^2}\psi(t,\cdot)$ for all $t>0$. Note that $||\nabla B(t,\cdot)||_2$, $|\bar{B}|(t)$ belong to $L^1((0,T))$ for any $T>0$. Also,

$$\int_{0+}\frac{1}{\Gamma(r)}\,dr=+\infty.$$

Thus, it follows that $\psi(t,x)=\overline{\rho^2}-\rho^2=0$ a.e. $(0,+\infty)\times\mathbb{T}^2$, and consequently (101). The proof of Theorem is now complete.

Acknowledgments. A part of this work was done as a Ph.D. thesis at Northwestern University (USA). I express much gratitude to Professor Gui-Qiang Chen for support and advice.

REFERENCES

[1] G.K. Batchelor, *Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 1967.
[2] R. Coifman, P.-L. Lions, Y. Meyer, and S. Semmes, *Compensated compactness and Hardy spaces*, J. Math. Pures Appl., 72 (1993), pp. 247–286.
[3] B. Desjardins, *Regularity of weak solutions of the compressible isentropic Navier-Stokes equations*, Comm. Partial Differential Equations, 22 (1997), pp. 977–1008.
[4] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, 2004.
[5] E. Feireisl, A. Novotný, and H. Petzeltová, *On the existence of globally defined weak solutions to the Navier-Stokes equations*, J. Math. Fluid Mech., 3 (2001), pp. 358–392.
[6] C. Fefferman, *Characterizations of bounded mean oscillation*, Bull. Amer. Math. Soc., 77 (1971), pp. 587–588.
[7] D. Hoff, *Discontinuous solutions of the Navier-Stokes equations for compressible flow*, Arch. Rational Mech. Anal., 114 (1991), pp. 15–46.
[8] A.V. Kazhikhov and V.A. Waigant, *On the existence of global solutions of two-dimensional Navier-Stokes equations of a compressible viscous flow*, Sibirsk. Mat. Zh., 36 (1995), pp. 1283–1316 (in Russian).
[9] M.A. Krasnosel'skii and Ja. B. Rutickii, *Convex Functions and Orlicz Spaces*, P. Noordhoff Ltd., Groningen, The Netherlands, 1961.
[10] P.-L. Lions, *Mathematical Topics in Fluid Dynamics*, Vol. 1, *Incompressible Models*, Oxford Science Publications, New York, 1998.
[11] P.-L. Lions, *Mathematical Topics in Fluid Dynamics*, Vol. 2, *Compressible Models*, Oxford Science Publications, New York, 1998.
[12] A. Novotný and I. Straskraba, *Introduction to the Mathematical Theory of Compressible Flow*, Oxford University Press, Oxford, 2004.
[13] D. Serre, *Variations de grande amplitude pour la densite d'un fluide visqueux compressible*, Phys. D, 48 (1991), pp. 113–128.
[14] E.M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[15] E.M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.
[16] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1997.
[17] W.P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1154–1185

ON A HELE–SHAW TYPE DOMAIN EVOLUTION WITH CONVECTED SURFACE ENERGY DENSITY: THE THIRD-ORDER PROBLEM*

MATTHIAS GÜNTHER† AND GEORG PROKERT‡

Abstract. We investigate a moving boundary problem with a gradient flow structure which generalizes Hele–Shaw flow driven solely by surface tension to the case of nonconstant surface tension coefficient taken along with the liquid particles at boundary. The resulting evolution problem is first order in time, contains a third-order nonlinear pseudodifferential operator and is degenerate parabolic. Well-posedness of this problem in Sobolev scales is proved. The main tool is the construction of a variable symmetric bilinear form so that the third-order operator is semibounded with respect to it. Moreover, we show global existence and convergence to an equilibrium for solutions near trivial equilibria (balls with constant surface tension coefficient). Finally, numerical examples in 2D and 3D are given.

Key words. free boundary motion, degenerate nonlocal parabolic evolution

AMS subject classifications. 35R35, 76B07

DOI. 10.1137/050626995

1. Introduction. It is the aim of the present paper to consider the generalization of the well-investigated Hele–Shaw flow problem to the case of nonconstant surface tension coefficient (or surface energy density). While experiments on such situations have been reported in the literature (e.g., [3, 12]), theoretical investigations of this seem to be lacking. A first step in this direction has been made in [10] where short-time solvability was proved for a Hele–Shaw problem with nonconstant surface tension coefficient and so-called kinetic undercooling. Here we discuss the problem without this regularization, using again the simple assumption that the surface energy density is convectively transported along the moving boundary.

This leads to the following moving boundary problem: For a given bounded domain $\Omega(0) \subset \mathbb{R}^m$ and a given nonnegative function γ_0 defined on $\partial\Omega(0)$ one looks for a family of C^2-domains $\Omega(t) \subseteq \mathbb{R}^m$, $t > 0$ and functions $\varphi(\cdot,t) \in C^2\big(\overline{\Omega(t)}\big)$, $\psi(\cdot,t) \in C^2\big(\overline{\Omega(t)}\big)$, $\gamma_t \in C^2\big(\Gamma(t)\big)$ such that

$$\left.\begin{aligned} \Delta\varphi(\cdot,t) &= 0 && \text{in } \Omega(t),\\ \Delta\psi(\cdot,t) &= 0 && \text{in } \Omega(t),\\ \partial_n\psi(\cdot,t) &= \Delta_{\Gamma(t)}\gamma_t && \text{on } \Gamma(t),\\ \varphi(\cdot,t) &= \gamma_t\kappa(t) - \psi(\cdot,t) && \text{on } \Gamma(t),\\ V_n &= \partial_n\varphi(\cdot,t) && \text{on } \Gamma(t). \end{aligned}\right\} \tag{1.1}$$

Here γ_t represents the variable surface tension coefficient (local surface energy density) on $\Gamma(t)$, $\varphi(\cdot,t)$ is the velocity potential inside the flow domain $\Omega(t)$, and ψ is an auxiliary function describing the nonlocal influence of the spatial variation of γ_t.

*Received by the editors March 17, 2005; accepted for publication (in revised form) May 10, 2006; published electronically December 15, 2006.
http://www.siam.org/journals/sima/38-4/62699.html
†Mathematisches Institut, Universität Leipzig, Augustusplatz 10/11, 04109 Leipzig, Germany (Matthias.Guenther@math.uni-leipzig.de).
‡Faculty of Mathematics and Computer Science, Technische Universiteit Eindhoven, P. O. Box 513, 5600 MB Eindhovn, The Netherlands (g.prokert@tue.nl).

Furthermore, $\kappa(t)$ is the $(m-1)$-fold mean curvature of $\Gamma(t)$, with the sign taken such that κ is negative for convex domains, ∂_n is the outer normal derivative, and $V_n(t)$ is the (outer) normal velocity of $\Gamma(t)$, determining its time evolution. Note that by setting $\Phi = \varphi + \psi$, (1.1) simplifies to

$$\left.\begin{aligned} \Delta\Phi(\cdot,t) &= 0 && \text{in } \Omega(t),\\ \Phi(\cdot,t) &= \gamma_t\kappa(t) && \text{on } \Gamma(t),\\ V_n &= \partial_n\Phi(\cdot,t) - \Delta_{\Gamma(t)}\gamma_t && \text{on } \Gamma(t). \end{aligned}\right\} \tag{1.2}$$

This problem generalizes the well-known Hele–Shaw flow with surface tension regularization in the following way: Any solution represents a gradient flow with respect to the usual energy functional

$$E(\gamma,\Gamma) := \int_\Gamma \gamma \, d\Gamma,$$

where $\gamma > 0$ is now variable on Γ, and to the Riemannian metric g_Γ on the infinite-dimensional manifold $\mathcal{M}$ of surfaces Γ enclosing a fixed volume given by

$$g_\Gamma(v_1,v_2) := \int_\Omega \nabla\varphi_1 \nabla\varphi_2 \, dx, \tag{1.3}$$

where the φ_i, $i = 1,2$ are (weak) solutions of the Neumann problems

$$\Delta\varphi_i = 0 \text{ in } \Omega, \quad \partial_n\varphi_i = v_i \text{ on } \Gamma.$$

The functions v_i can be identified with tangent vectors of $\mathcal{M}$; note that the conservation of volume implies $\int_\Gamma v_i = 0 \, d\Gamma$. For more details and references, see [1, 7, 10]. Kinetic undercooling regularization corresponds to adding in (1.3) a boundary integral term $\beta \int_\Gamma v_1 v_2 \, d\Gamma$ with $\beta > 0$, this case is discussed in [10].

As mentioned already, we assume that the values of the function Γ are transported with the liquid particles: Introducing Lagrangian coordinates $x = x(\xi,t)$, $\xi \in \Gamma(0)$ corresponding to the velocity field via

$$\partial_t x(\xi,t) = \nabla\varphi\big(x(\xi,t),t\big) \text{ for } t \geq 0, \quad x(\xi,0) = \xi, \tag{1.4}$$

we obtain that $x = x(\cdot,t)$ is a diffeomorphism from $\Gamma(0)$ onto $\Gamma(t)$, and the transport law for γ_t takes the form

$$\gamma_t\big(x(\xi,t)\big) = \gamma_0(\xi), \quad \xi \in \Gamma(0), \ t \geq 0. \tag{1.5}$$

This assumption is reasonable, for example, when γ depends on temperature and heat diffusion is negligible compared to convection. While it certainly oversimplifies the physical situation in the case when, for example, surfactants play a role, it seems that the mathematical character of the problem is essentially the same there as in our case. Note, however, that the situation here is qualitatively different from other models like anisotropic Hele–Shaw flow (cf. [6]) because in our case the evolution is not determined solely by the shape of the evolving domain but there is a coupling with a transport problem in the moving boundary.

Our approach is based on reformulating (1.1) as a vector-valued evolution equation for a diffeomorphism mapping a fixed reference manifold to the moving boundary, i.e., we work with "Lagrangian cordinates" for the boundary. This approach which

might appear unnatural at first glance has two advantages: The transport problem for γ is simply solved by prescribing a fixed smooth positive function on this reference manifold and pushing it forward to the moving boundary. Moreover, the invariance of the problem with respect to reparametrizations of the boundary can be used to derive nonlocal analogues of Leibniz' rule of differentiation (3.10) which we will use in the course of our estimates. On the other hand, however, our approach implies that the resulting vector valued evolution equation (2.2), cannot be strictly parabolic, even for strictly positive γ. Instead, the (linearized) problem is degenerate in tangential directions, and although we have a third-order parabolic operator in normal direction, straightforward coercivity estimates for the vector valued problem are not available. Therefore we work with energy estimates with respect to a variable inner product (4.26) which is specifically constructed for this purpose. (This is similar to the usual treatment of symmetrizable hyperbolic systems.)

The paper is organized as follows.

After announcing our main results on short-time existence in section 2, we start the proofs in section 3 by investigating mapping properties of the occurring nonlocal operators in Sobolev scales. In particular, we derive flexible multilinear estimates for their Fréchet derivatives in low norms and extend them to higher norms by a generalized chain rule based on invariance properties. For related considerations concerning the analytic dependence of the Dirichlet–Neumann operator on the domain we refer to [5] and the references given there. Section 4 is devoted to the proof of the crucial estimate providing the semiboundedness of the evolution operator with respect to our variable inner product.

Technically, we use the natural decomposition of the right-hand side into a second-order operator mapping vectors to scalars and a first-order operator mapping scalars to vectors. Furthermore, we use the fact that the right-hand side is—in a sense to be made precise later—coercive with respect to the normal component. The semiboundedness enables us to invoke an abstract existence result based on Galerkin approximations and Rothe's method. This is done in section 5. In this way, we prove our main result (Theorems 2.1 and 2.2) on short-time wellposedness of the moving boundary problem (1.1), (1.4), (1.5). We will omit certain details as they are parallel to the discussion in [10]. However, the right-hand side of the evolution problem obtained there is of order two. As we are concerned here with an evolution equation whose right-hand side is of order three, we have to refine the construction of our variable inner product from [10] by including certain lower-order terms. Differing from the situation there, here we have to demand strict positivity of γ because its inverse γ^{-1} enters one of these terms.

Finally, in section 6 we investigate the evolution near the equilibrium solutions given by balls with constant γ. In this situation, Theorem 6.8 gives global existence in time and the evolving domain approaches a nontrivial equilibrium configuration depending on the given (nonconstant) γ and the initial domain. In contrast to the classical case of constant γ where the equilibria are given only by balls, any shape near a ball occurs as an equilibrium configuration for a certain function γ near the constant. Due to the degeneracy of our problem, the proof of long-time existence is more involved compared to the known proofs for the case of constant γ, cf. [4, 8, 13].

2. Statement of the local existence results. We begin by listing some notation. $C, C_1, \dots$ etc. denote generic constants; their dependencies on other quantities is only indicated if not obvious from the context. Let $E \subseteq \mathbb{R}^m$, $m \geq 2$ be a bounded domain with smooth boundary $S := \partial E$ and ν the outer unit normal on S. For

$M = S$ or $M = E$, we make constant use of the usual L^2-based Sobolev spaces $H^s(S)$, $H^s(S, \mathbb{R}^m)$ of order s with values in $\mathbb{R}$ and $\mathbb{R}^m$, respectively. The norms of these spaces will be denoted by $\|\cdot\|_s^M$; for $M = S$ the upper index M is dropped in most cases. When Fréchet derivatives of operator-valued mappings are considered, the additional arguments describing the variations are written in accolades ($\{\}$).

Now, as already mentioned in the introduction, we reformulate the moving boundary problem (1.1)—(1.5) by describing $\Gamma(t)$ as an embedding $u(\cdot, t) : S \to \mathbb{R}^m$ such that the curves $t \mapsto u(y, t)$ for fixed $y \in S$ are trajectories belonging to the velocity field and γ_t is constant along these curves. This approach enables us to consider γ_t as a known function during the evolution at the cost of describing the moving boundary by m functions. To do so, let

$$U := \big\{u : S \to \mathbb{R}^m \;\big|\; u = w|_S \text{ with } w \in \mathrm{Diff}(\bar{E}, \Omega_u \cup \Gamma_u)\big\}, \tag{2.1}$$

where

$$\Omega_u = w(E) \quad \text{and} \quad \Gamma_u = \partial\Omega_u = u(S).$$

Throughout this paper, we use the abbreviation

$$U_s := U \cap H^s(S, \mathbb{R}^m).$$

Now, (1.1)—(1.5) is reduced to the following Cauchy problem, which will be investigated afterwards. For a given $u_0 \in U_s$, s sufficiently large, we look for $T > 0$ and a mapping $[0, T] \ni t \mapsto u(t) \in U_s$, such that

$$u'(t) = \mathcal{F}\big(u(t)\big), \quad t \in [0, T], \tag{2.2}$$

$$u(0) = u_0. \tag{2.3}$$

Thereby, for $u \in U$, we have set

$$\mathcal{F}(u) := F(u)\mathcal{G}(u) \quad \text{with} \quad \mathcal{G}(u) := H(u) + G(u), \tag{2.4}$$

where, for any given function f on S,

$$F(u)f := \nabla\varphi(u, f) \circ u \tag{2.5}$$

and $\varphi = \varphi(u, f)$ denotes the solution of the Dirichlet problem

$$\Delta\varphi = 0 \text{ in } \Omega_u, \quad \varphi = f \circ u^{-1} \text{ on } \Gamma_u. \tag{2.6}$$

Further, $H(u)$, $G(u)$ are given by

$$H(u) := \gamma(\kappa_{\Gamma_u} \circ u), \quad G(u) := -A(u)\Delta(u)\gamma. \tag{2.7}$$

Here $\gamma \in C^\infty(S)$ is a fixed and given positive function, κ_{Γ_u} denotes the mean curvature of Γ_u with sign and scaling conventions as above, and

$$\Delta(u)w := \Delta_{\Gamma_u}(w \circ u^{-1}) \circ u \tag{2.8}$$

is the pullback to S of the Laplace–Beltrami operator Δ_{Γ_u} on Γ_u and

$$A(u)f := \varphi_N\big(u, f\big) \circ u \tag{2.9}$$

is the Neumann–Dirichlet operator, i.e., $\varphi_N = \varphi_N(u, f)$ solves the Neumann problem

$$\Delta\varphi_N = 0 \text{ in } \Omega_u, \quad \partial_n\varphi_N = c + f \circ u^{-1} \text{ on } \Gamma_u, \quad \int_{\Gamma_u} \varphi_N \, dx = 0. \tag{2.10}$$

The constant $c = c(u, f) \in \mathbb{R}$ in (2.10) is determined by the solvability condition

$$\int_{\Gamma_u} (f \circ u + c) \, d\Gamma_u = 0; \tag{2.11}$$

clearly $c(u, f) = 0$ for $f = \Delta(u)\gamma$. For fixed smooth γ on S, the mapping $u \mapsto H(u)$ constitutes a quasilinear second order differential operator on S. Moreover, the solutions of the boundary value problems (2.6), (2.10) depend smoothly on the domain Ω_u, i.e., on $u \in H^s$, $s > (m+1)/2$ and $f \mapsto F(u)f$, $f \mapsto A(u)f$ represent pseudodifferential operators of order one and minus one, respectively. In particular, G is a pseudodifferential operator of lower order than H and may be considered as a correction term to ensure the gradient flow structure of the evolution problem. We will show later that

$$[u \mapsto \mathcal{F}(u)] \in C^\infty\big(U_s, H^{s-3}(S, \mathbb{R}^m)\big) \tag{2.12}$$

for $s > (m+3)/2$, $s \geq 3$. Now we are in position to formulate our main results.

THEOREM 2.1 (short-time existence and uniqueness). *Fix an even integer $s_0 > (m+7)/2$, $s_0 \geq 6$ and assume $\gamma \in C^\infty(S)$ strictly positive on S. Let $s \geq s_0$ be an even integer and $u_0 \in U_s$. Then there exist $T > 0$ and a unique solution*

$$u \in C\big([0, T], U_s\big) \cap C^1\big([0, T], H^{s-3}(S, \mathbb{R}^m)\big) \tag{2.13}$$

of the initial value problem (2.2), (2.3). Additionally, any given $\bar{u}_0 \in U_{s_0}$ has a suitable H^{s_0}-neighborhood K, such that for initial values u_0 varying in $K \cap H^s$, there are $T > 0$ and C independent of u_0 such that

$$\|u(t)\|_s \leq C(1 + \|u(0)\|_s) \text{ for all } t \in [0, T]. \tag{2.14}$$

THEOREM 2.2 (regularity and continuous dependence on initial values). *Under the assumptions of Theorem 2.1 let u be a any solution to (2.2) in the class (2.13) with some $T > 0$. Then there holds the following:*

(i) *$u(0) \in H^{s+1}(S, \mathbb{R}^m)$ implies $u(t) \in H^{s+1}(S, \mathbb{R}^m)$ for all $t \in [0, T]$.*

(ii) *Assume $u_0^n \to u_0$ in $H^s(S, \mathbb{R}^m)$ for $n \to \infty$. Then, for n sufficiently large, there exist solutions u_n of (2.2) in the class (2.13) with initial values $u_n(0) = u_0^n$ and there holds $u_n \to u$ in $C\big([0, T], H^s(S, \mathbb{R}^m)\big)$.*

The proof of both theorems is given in section 5.

Remark 2.3. The restriction to even integers s is due purely to the construction of our bilinear form involving integer powers of a generalized Laplacian. This restriction can be lifted afterwards by using the nonlinear interpolation result given in [2, Proposition A.1 and Remark A.2]. The dimension independent restriction $s_0 \geq 6$ is needed as we use dual estimates for elliptic boundary value problems in norms with negative index.

3. Smooth domain dependence of the nonlocal operators. We start by gathering some properties of the nonlocal operators F, A, and $\mathcal{G}$ defined by (2.4)–(2.11). The multilinear estimates for the Fréchet derivatives can be seen as counterparts to the product estimate

$$\|u_1 \dots u_k\|_t \leq C\|u_1\|_{s_1} \dots \|u_k\|_{s_k}$$

holding if $0 \le t \le s_i \le \sigma$, $\sigma > (m-1)/2$, $\sum_{i=1}^k s_i \ge t + (k-1)\sigma$. Here, however, we have to deal with nonlocal operators of various orders involving differentiations and the solution of elliptic BVP.

The statements and their proofs are essentially parallel to Corollary 4.4 and Lemma 4.5 in [10], therefore proofs will be omitted. Note, however, that $f \mapsto F(u)f$ is an operator of order one here as (2.6) is a Dirichlet problem. In fact, the normal component of F is given by the Dirichlet–Neumann operator while the tangential component is given by the tangential gradient of f.

Due to the variability in the choice of the s_i, the estimates will be flexible enough to control various lower order terms that will occur in what follows.

LEMMA 3.1. (i) *Let $s > (m+1)/2$ and $t \in [1, s]$ be given. Then*

$$F \in C^\infty\big(U_s, \mathcal{L}(H^t(S), H^{t-1}(S, \mathbb{R}^m))\big),$$
$$A \in C^\infty\big(U_s, \mathcal{L}(H^{t-1}(S), H^t(S, \mathbb{R}^m))\big),$$

and for any $u \in U_s$ and any choice of $s_1, \dots s_{k+1} \in [t, s]$ with $s_1 + \dots + s_{k+1} \ge t + ks$ there exists a constant $C > 0$ such that for all $f \in H^s(S)$ and all $u_1, \dots, u_k \in H^s(S, \mathbb{R}^m)$ there holds

$$\big\|F^{(k)}(u)\{u_1, \dots, u_k\}f\big\|_{t-1} \le C\|u_1\|_{s_1} \cdots \|u_k\|_{s_k}\|f\|_{s_{k+1}}, \tag{3.1}$$

$$\big\|A^{(k)}(u)\{u_1, \dots, u_k\}f\big\|_{t} \le C\|u_1\|_{s_1} \cdots \|u_k\|_{s_k}\|f\|_{s_{k+1}-1}. \tag{3.2}$$

(ii) *Let $s > (m+3)/2$ and $t \in [2, s]$ be given. Then*

$$\mathcal{G} \in C^\infty\big(U_s, H^{s-2}(S)\big)$$

and for any $u \in U_s$ and any choice of $s_1, \dots s_k \in [t, s]$ with $s_1 + \dots + s_{k+1} \ge t + ks$ there exists a constant $C > 0$ such that for all $u_1, \dots, u_k \in H^s(S, \mathbb{R}^m)$ there holds

$$\big\|\mathcal{G}^{(k)}(u)\{u_1, \dots, u_k\}\big\|_{t-1} \le C\|u_1\|_{s_1} \cdots \|u_k\|_{s_k}. \tag{3.3}$$

The constants may be chosen independently of u as u varies in bounded and weakly closed subsets of U_s.

Remark 3.2. Note that a bounded subset of $H^s(S)$ is weakly closed if and only if it is closed in $H^t(S)$ for some $t < s$. Then it is compact in all $H^t(S)$ with $t < s$.

Remark 3.3. The estimate (3.3) is not optimal as we do not use the quasilinear character of $\mathcal{G}$. For our purposes, however, it will be sufficient.

Note that Lemma 3.1 implies the smoothness assertion (2.12).

Next, we prove some related estimates in norms with negative index. The use of such norms implies a loss of flexibility. Essentially, these estimates are parallel to product estimates of the type

$$\|u_1 \dots u_k\|_t \le C\|u_1\|_s \dots \|u_{k-1}\|_s\|u_k\|_t,$$

$t \in [-s, s]$, $s > (m-1)/2$, which can be proved by duality arguments if $t < 0$.

LEMMA 3.4. *Assume $s > (m+1)/2$, $s \ge 4$, $t \in [-3, s-1]$. Then*

$$F \in C^\infty\big(U_s, \mathcal{L}(H^{t+1}(S), H^t(S))\big),$$
$$A \in C^\infty\big(U_s, \mathcal{L}(H^{t-1}(S), H^t(S))\big),$$

and for $u \in U_s$

$$\|F'(u)u_1\|_t \le C\|u_1\|_s\|f\|_{t+1}, \tag{3.4}$$

$$\|A(u)f\|_t \le C\|f\|_{t-1}, \tag{3.5}$$

$$\|A^{(k)}(u)\{u_1,\dots,u_k\}f\|_t \le C\|u_1\|_s\dots\|u_k\|_s\|f\|_{t-1}, \tag{3.6}$$

$$\|A^{(k)}(u)\{u_1,\dots,u_k\}f\|_t \le C\|u_1\|_s\dots\|u_{k-1}\|_s\|u_k\|_t\|f\|_s, \tag{3.7}$$

$u_1,\dots,u_k \in H^s(S,\mathbb{R}^m)$, $k \in \mathbb{N}$. *The constants* C *can be chosen independently of* u *as* u *varies in bounded, weakly closed subsets of* U_s.

Proof. We will restrict ourselves to the assertions concerning A. Fix $s_0 \in ((m+1)/2, s)$ and an extension operator $\mathcal{E} \in \mathcal{L}(H^t(S), H^{t+1/2}(E))$, $t > 0$. Pick $v \in U_s$ and choose an H^{s_0}-neighborhood $V_{s_0} \subset U_s$ and $u_0 \in C^\infty(\overline{E},\mathbb{R}^m)$ such that

$$\widetilde{u} := u_0 + \mathcal{E}(u - u_0) \in \mathrm{Diff}(\overline{E}, \overline{\Omega}_u).$$

This is possible by Lemma 4.1 in [10].

For $u \in V_{s_0}$, let the transformed operators $L(u)$ and $\mathcal{B}(u)$ be defined by

$$L(u)\psi := \partial_i(\sqrt{g}g^{ij}\partial_j\psi), \quad \mathcal{B}(u)\psi := \nu_i\sqrt{g}g^{ij}\partial_j\psi,$$

where $\sqrt{g}$, g^{ij} are the volume element and the (inverse) coefficients of the metric on E induced by $\widetilde{u}$, respectively, and ν is the outer unit normal on S. We consider the transformed boundary value problem

$$L(u)\psi = \Phi_1, \quad \mathcal{B}(u)\psi = \omega(u)(\Phi_2 + c), \quad \int_s \omega(u)(\Phi_2 + c)\,dS = \int_E \sqrt{g}\Phi_1\,dx, \tag{3.8}$$

$c = c(u,\Phi_1,\Phi_2) \in \mathbb{R}$. Here $\omega(u) = d\Gamma_u/dS$ is the surface element belonging to the transformation induced by u which is given by a nonlinear first-order differential operator in $\widetilde{u}$.

For $\tau > 0$ and $v \in L^2(E)$ define

$$\|v\|_\tau := \sup_{z\in H^\tau(E),\|z\|_\tau=1} \left|\int vz\,dx\right|.$$

(This differs from the usual norm in $H^{-\tau}(E) := (H_0^\tau(E))'$.) The BVP (3.8) is uniquely solvable and ψ satisfies an estimate

$$\|\psi\|_t + \|\psi\|^E_{t+1/2} \le C\big(\|\Phi_1\|^E_{t-3/2} + \|\Phi_2\|_{t-1}\big) \tag{3.9}$$

(cf. [9, Lemma 3.1]).

As $A(u)f$ is the trace of the solution ψ of (3.8) with $\Phi_1 = 0$, $\Phi_2 = f$, we get (3.5) immediately from (3.9).

Note that $A'(u)\{u_1\}f$ is given as the solution ψ' of

$$\begin{aligned} &L(u)\psi' = -L'(u)\{u_1\}\psi,\\ &\mathcal{B}(u)\psi' = -\mathcal{B}'(u)\{u_1\}\psi + \omega'(u)\{u_1\}(f + c(u,0,f)) + \omega(u)\partial_u c(u,0,f)\{u_1\}. \end{aligned}$$

As $f \mapsto c(u,0,f)$ and $v \mapsto \partial_u c(u,0,f)\{v\}$ are given by smoothing operators, to obtain (3.6) and (3.7) it is sufficient to use (3.9) and estimate either

$$\|L'(u)\{u_1\}\psi\|^\Omega_{t-3/2} \le C\big(\|\mathcal{E}u_1\|^\Omega_{t+1/2} + \|u_1\|_t\big)\|\psi\|_{s+1} \le C\|u_1\|_t\|f\|_s$$

or

$$\|L'(u)\{u_1\}\psi\|^{\Omega}_{t-3/2} \le C\big(\|\psi\|^{\Omega}_{t+\frac12} + \|\psi\|_t\big)\|u_1\|_s \le C\|u_1\|_s\|f\|_{t-1},$$

together with analogous estimates for $\|\mathcal{B}'(u)\psi\|_{t-1}$ and $\|\omega'(u)\{u_1\}f\|_{t-1}$. The general case now follows by induction over k, cf. [10, Lemma 4.5].

The estimate (3.4) can be obtained in a similar fashion, discussing a Dirichlet problem instead of (3.8).

Finally, the uniformity of the estimates follows from the fact that bounded, weakly closed subsets of U_s are compact in $H^{s_0}(S)$. □

We choose m smooth vector fields $D_1, \dots, D_m$ on S such that

$$\mathrm{Span}\{D_1, \dots, D_m\} = T_x \ \text{ for all } \ x \in S$$

and use the multi-index notation $D^\alpha = D_1^{\alpha_1} \cdots D_m^{\alpha_m}$, $\alpha = (\alpha_1, \dots, \alpha_m)$ for higher order derivatives; for simplicity we assume that $(D_1, \dots, D_m)$ coincides with the tangential gradient on S. Note that, for any integer $s \ge 0$, we can use

$$(u, v)_s = \sum_{|\alpha| \le s} (D^\alpha u, D^\alpha v)_{L^2(S)}$$

as a scalar product generating the norm in $H^s(S)$. Moreover, as an immediate consequence of the invariance properties

$$\big(F(u)f\big) \circ \tau = F(u \circ \tau)(f \circ \tau)$$

for any diffeomorphism τ on S, we have a differentiation rule which resembles Leibniz' rule at an abstract level, cf. [10]: For any multi-index α and $u \in U_s$, $f \in H^s(S)$, $s > |\alpha| + (m+1)/2$ there holds

$$D^\alpha F(u)f = \sum c_{\beta_1,\dots,\beta_{k+1}} F^{(k)}(u)\{D^{\beta_1}u, \dots, D^{\beta_k}u\}D^{\beta_{k+1}}f, \tag{3.10}$$

where the sum has to be extended over all integers k and systems of nonnegative multi-indices $\beta_1, \dots, \beta_{k+1}$ with

$$0 \le k \le |\alpha|, \quad 1 \le |\beta_1|, \dots, |\beta_k|, \quad \beta_1 + \cdots + \beta_{k+1} = \alpha. \tag{3.11}$$

The coefficients are nonnegative integers, in particular $c_\alpha = c_{\alpha,0} = 1$.

Combining the differentiation rule for F with the estimate of the derivatives in lower norms we obtain the following proposition.

PROPOSITION 3.5. (i) *Let $s \ge s_0 > (m+1)/2$, s integer, $u \in U_s$. Then*

$$\|F(u)f\|_{s-1} \le C\big(\|u\|_s\|f\|_{s_0} + \|f\|_s\big) \tag{3.12}$$

with a uniform constant as long as u varies in H^{s_0}-bounded and weakly H^{s_0}-closed subsets of U_s.

(ii) *Assume additionally that $s \ge s_0 + 2$ and let α be any multi-index with $|\alpha| = s$. Writing $D^\alpha = D^{\alpha_1} \cdots D^{\alpha_s}$ with $|\alpha_1| = \cdots = |\alpha_s| = 1$, we have*

$$D^\alpha F(u)f = F(u)D^\alpha f + F'(u)\{D^\alpha u\}f + \sum_{i=1}^{s} F'(u)\{D^{\alpha_i}u\}D^{\beta_i}f + R_\alpha(u)f, \tag{3.13}$$

where $\alpha = \alpha_i + \beta_i$ and the remainder term allows the estimate

$$\|R_\alpha(u)f\|_0 \le C\big(\|u\|_s\|f\|_{s_0+1} + \|f\|_{s-1}\big).$$

The constant can be chosen uniformly as u varies in H^{s_0+2}-bounded, weakly H^{s_0+2}-closed subsets of U_s.

Proof. We consider the more complicated situation (ii) only. According to (3.10), the remainder term has a representation as a sum of terms

$$I_\beta = F^{(k)}(u)\{D^{\beta_1}u, \dots, D^{\beta_k}u\}D^{\beta_{k+1}}f,$$

where the multi-indices satisfy (3.11) and additionally

$$|\beta_1|, \dots, |\beta_k| \le s-1, \quad |\beta_{k+1}| \le s-2.$$

Hence $k \ge 1$. For each of the terms I_β, we will choose numbers $\theta_1, \dots, \theta_{k+1} \in [0,1]$ such that $\theta_1 + \cdots + \theta_{k+1} = 1$ and set

$$s_i := (1-\theta_i)s_0 + \theta_i.$$

If $k = 1$ we choose θ_1, θ_2 such that $\theta_1 + \theta_2 = 1$ and $|\beta_2| = \theta_1 + \theta_2(s-2)$. If $k = 2$ and $|\beta_3| = 0$ we choose $\theta_i := (|\beta_i| - 1)/(s-2)$ for $i = 1, 2$ and $\theta_3 := 0$. If $k = 2$ and $|\beta_3| \ge 1$ or $k \ge 3$ we choose

$$\theta_i := (|\beta_i| - 1)/(s-3) \text{ for } i = 1,2,3, \quad \theta_i := |\beta_i|/(s-3) \text{ for } i \ge 4.$$

In all cases, we have

$$\left.\begin{aligned} |\beta_i| + s_i &\le (1-\theta_i)(s_0+2) + \theta_i s, && i = 1, \dots, k, \\ |\beta_{k+1}| + s_{k+1} &\le (1-\theta_{k+1})(s_0+1) + \theta_{k+1}(s-1). \end{aligned}\right\} \tag{3.14}$$

Set $\lambda := \theta_1 + \cdots + \theta_k$. Using (3.1) with $t = 1$, $s = s_0$, (3.14), norm convexity, and Young's inequality we get

$$\begin{aligned} \|I_\beta\|_0 &\le C\|u\|_{|\beta_1|+s_1} \dots \|u\|_{|\beta_1|+s_k}\|f\|_{|\beta_{k+1}|+s_{k+1}} \\ &\le C\|u\|_{s_0+2}^{k-1}\big(\|u\|_{s_0+2}\|f\|_{s-1}\big)^{1-\lambda}\big(\|u\|_s\|f\|_{s_0+1}\big)^{\lambda} \\ &\le C\|u\|_{s_0+2}^{k-1}\big(\|u\|_{s_0+2}\|f\|_{s-1} + \|u\|_s\|f\|_{s_0+1}\big), \end{aligned}$$

and the result follows. ∎

The following lemma provides an explicit characterization for the linearization of F, namely up to terms of order zero in v,

$$F'(u)f \approx -F(u)(v \cdot F(u)f).$$

This structure will be important later. It can be verified in an informal way by performing the variation on Ω_u itself instead of transforming the problem to the reference domain.

LEMMA 3.6. *Let $s > (m+3)/2$. Then for $u \in U_s$, $v \in H^s(S, \mathbb{R}^m)$, and $f \in H^s(S)$ there holds*

$$\|F'(u)\{v\}f + F(u)(v \cdot F(u)f)\|_0 \le C\|f\|_s\|v\|_0.$$

Proof. From (2.5) we get

$$F_i'(u)\{v\}f = \partial_i\phi' \circ u + v_j\partial_i\partial_j\phi \circ u$$

with $\phi = \phi(u,f)$ from (2.6) and $\phi' = \phi'(u,f)\{v\}$ given by

$$\phi'(u,f)\{v\}(x) = \partial_\varepsilon(\phi(u+\varepsilon v, f)(x))|_{\varepsilon=0}, \quad x \in \Omega_u.$$

The function ϕ' satisfies

$$\Delta\phi' = 0 \text{ in } \Omega_u, \quad \phi' = -\nabla\phi \cdot v \text{ on } \Gamma_u, \tag{3.15}$$

therefore $\partial_i\phi' \circ u = -F_i(u)(v \cdot F(u)f)$.

Parallel to the proof of Lemma 5.1 in [10] one obtains

$$\|v_j\partial_i\partial_j\phi \circ u\|_0 \le C\|v\|_0\|\phi(u,f)\|_{C^2(\overline{\Omega}_u)} \le C\|f\|_s\|v\|_0.$$

This proves the assertion. □

4. The main estimate. In this section we prove a priori estimates in H^s for the nonlinear operator $\mathcal{F}$ w.r.t. variable bilinear forms which we define in what follows. As already mentioned in the introduction, these estimates are the main ingredient in the existence proof.

To begin with, for $u \in U_s$, $s > (m+1)/2$ we define

$$P(u)v := v \cdot \big(n(u) \circ u\big), \quad N(u)w := w\,\big(n(u) \circ u\big), \tag{4.1}$$

$$\Lambda(u)w := \nabla_{\Gamma_u}(w \circ u^{-1}) \circ u \tag{4.2}$$

as the Euclidean scalar product and multiplication with outer normal $n(u)$ of Γ_u and pullback of tangential gradient ∇_{Γ_u} along Γ_u, respectively. Considered as operators in v and w, the coefficients of $P(u)$, $N(u)$ and $\Lambda(u)$ are smooth functions of u and its first derivatives. Thus,

$$P(u) \in \mathcal{L}\big(H^t(S,\mathbb{R}^m), H^t(S)\big), \quad N(u) \in \mathcal{L}\big(H^t(S), H^t(S,\mathbb{R}^m)\big), \tag{4.3}$$

$$\Lambda(u) \in \mathcal{L}\big(H^t(S), H^{t-1}(S,\mathbb{R}^m)\big) \tag{4.4}$$

depend smoothly on $u \in U_s$ for $|t| \le s-1$ and $|t-1| \le s-1$, respectively. Clearly, the operators P, N, Λ satisfy invariance properties as stated for F in [10]. As a consequence, the differentiation rule (3.10) is also true for P, N, Λ; we make use of that without explicit mention. Further more, recall that the pullback $\Delta(u)w$ of the Laplace–Beltrami operator Δ_{Γ_u} on Γ_u according to (2.8) and the operator $H(u)$ according to (2.7) may be expressed as

$$\Delta(u)w = \Lambda_i(u)\big(\Lambda_i(u)w\big), \quad H(u) = -\gamma\Lambda_i(u)\big(n_i(u) \circ u\big), \tag{4.5}$$

respectively.

In the further considerations of this section we fix s_0 to be the smallest integer such that $s_0 \ge 6$ and $s_0 > (m+7)/2$ and set

$$\widetilde{U}_s := U_s \cap K \text{ for all } s \ge s_0$$

with a fixed H^{s_0}-bounded and weakly H^{s_0}-closed subset $K \subseteq U_{s_0}$. Note that

$$1 \le C\|u\|_{s_0} \le C'\|u\|_s, \quad \|u\|_{C^3(S)} \le C$$

for all $u \in \widetilde{U}_s$, $s \ge s_0$.

Furthermore, we have the estimates

$$\|\mathcal{G}(u)\|_{s-2}, \|\mathcal{F}(u)\|_{s-3} \le C\|u\|_s \ \text{ for all } \ u \in \widetilde{U}_s, s \ge s_0,$$

and the operators defined in (4.3), (4.4) are bounded uniformly with respect to $u \in \tilde{U}_s$.

Due to our choice of the differential operators D_i, the Laplace–Beltrami operator on the compact reference manifold S is given by $\Delta_0 := D_i D_i$. It has an approximate inverse, i.e., there is an operator $\Delta_0^+ \in \mathcal{L}\big(H^\tau(S), H^{\tau+2}(S)\big)$, $\tau \in \mathbb{R}$, such that

$$\Delta_0\Delta_0^+ = \Delta_0^+\Delta_0 = \mathrm{id} + Q_0,$$

with a smoothing operator Q_0 simply given by orthogonal projection in $L^2(S)$ onto the subspace of functions which are constant on each connectivity component of S; in particular, $Q_0 \in \mathcal{L}(H^\tau(S), H^\sigma(S))$ for any $\sigma, \tau \in \mathbb{R}$. In the same manner, we define the approximate inverse $\Delta^+(u)$ for $\Delta(u)$. In this case we have

$$[u \mapsto \Delta^+(u)] \in C^\infty\big(U_s, \mathcal{L}\big(H^t(S), H^{t+2}(S)\big)\big), \quad t \in [0, s-2]$$

and

$$\Delta(u)\Delta(u)^+ = \Delta(u)^+\Delta(u) = \mathrm{id} + Q(u), \tag{4.6}$$

where $Q(u) \in \mathcal{L}\big(H^\tau(S), H^\sigma(S)\big)$ for any $\sigma \in \mathbb{R}$, $\tau \ge 1 - s$, and the corresponding norms are bounded independently of $u \in \tilde{U}_s$.

LEMMA 4.1. *Let $s \ge s_0$ with $s = 2k$, $k \in \mathbb{N}$, and $u \in U_s$. Then we have*

$$\Delta_0^k \mathcal{F}(u) = \tilde{F}(u)\big(\tilde{G}(u)(\Delta_0^k u)\big) + F(u)\big(R_a(u)\big) + R_b(u), \tag{4.7}$$

where the abbreviations

$$\tilde{F}(u)f := F(u)f + F_0(u)f, \quad \tilde{G}(u)v := \gamma\Delta(u)(P(u)v) + G_1(u)v$$

have been used. Here $f \mapsto F_0(u)f$ and $v \mapsto G_1(u)v$ are operators of order zero and one, respectively,

$$F_0(u) \in \mathcal{L}\big(H^t(S), H^t(S, \mathbb{R}^m)\big), \quad G_1(u) \in \mathcal{L}\big(H^t(S, \mathbb{R}^m), H^{t-1}(S)\big), \tag{4.8}$$

$t \in [-1, s-1]$ and $t \in [-2, s-3]$, respectively, and the remainder terms R_a, R_b satisfy

$$\|R_a(u)\|_0 \le C\|u\|_{s-1}, \tag{4.9}$$

$$\|R_b(u)\|_0 \le C\big(\|\tilde{G}(u)(\Delta_0^k u)\|_{-1} + \|\mathcal{F}(u)\|_{s_0-3}\|u\|_s + \|u\|_{s-1}\big). \tag{4.10}$$

The constants are independent of u and the operator norms of $F_0(u)$ and $G_1(u)$ are bounded independently of u as long as u varies in a set $\tilde{U}_s$.

Proof. The operator $F(u)$ vanishes on constants and, by elliptic regularity,

$$\|f\|_{s_0-2} \le C\|F(u)f\|_{s_0-3}, \qquad u \in \tilde{U}_s \tag{4.11}$$

if f has zero mean value over S. We define

$$\tilde{\mathcal{G}}(u) := \mathcal{G}(u) - \frac{1}{|S|}\int_S \mathcal{G}(u)\, dS. \tag{4.12}$$

Using Proposition 3.5 (ii), we write $\Delta_0^k\mathcal{F}(u) = \Delta_0^k F(u)\tilde{\mathcal{G}}(u)$ in the form

$$\begin{aligned}&F(u)\big(\Delta_0^k\tilde{\mathcal{G}}(u)\big) + F'(u)\big\{\Delta_0^k u\big\}\tilde{\mathcal{G}}(u) + 2\sum_{j=0}^{k-1} F'(u)\big\{D_i u\big\}\Delta_0^j D_i\Delta_0^{k-1-j}\tilde{\mathcal{G}}(u) + R_1(u)\\ &= F(u)\big(\Delta_0^k\mathcal{G}(u)\big) + F'(u)\big\{\Delta_0^k u\big\}\tilde{\mathcal{G}}(u) + 2kF'(u)\big\{D_i u\big\}D_i\Delta_0^{k-1}\mathcal{G}(u)\\ &\quad + R_1(u) + R_2(u).\end{aligned}$$

According to this proposition and (4.11), $R_1(u)$ allows the estimate

$$\|R_1(u)\|_0 \le C\big(\|u\|_s\|\tilde{\mathcal{G}}(u)\|_{s_0-2} + \|\tilde{\mathcal{G}}(u)\|_{s-1}\big) \le C\big(\|u\|_s\|\mathcal{F}(u)\|_{s_0-3} + \|\mathcal{G}(u)\|_{s-1}\big).$$

For $R_2(u)$ we find from Lemma 3.1 (with $t = 1$) that

$$\begin{aligned}\|R_2(u)\|_0 &\le 2\sum_{j=0}^{k-1}\big\|F'(u)\big\{D_i u, [\Delta_0^j, D_i]\Delta_0^{k-1-j}\mathcal{G}(u)\big\}\big\|_0\\ &\le C\sum_{i,j}\big\|[\Delta_0^j, D_i]\Delta_0^{k-1-j}\mathcal{G}(u)\big\|_1 \le C\|\mathcal{G}(u)\|_{s-1}.\end{aligned}$$

Further, by Lemma 3.6 we have

$$F'(u)\big\{\Delta_0^k u\big\}\tilde{\mathcal{G}}(u) = -F(u)\big(\Delta_0^k u\cdot\mathcal{F}(u)\big) + R_3(u)$$

with

$$\|R_3(u)\|_0 \le C\|\Delta_0^k u\|_0\|\tilde{\mathcal{G}}(u)\|_{s_0-2} \le C\|u\|_s\|\mathcal{F}(u)\|_{s_0-3}.$$

Defining $F_0(u)$ by

$$F_0(u)v := 2kF'(u)\big\{D_i u\big\}D_i\Delta_0^+ v,$$

we get

$$\Delta_0^k\mathcal{F}(u) = \big(F(u) + F_0(u)\big)\big(\Delta_0^k\mathcal{G}(u) - \Delta_0^k u\cdot\mathcal{F}(u)\big) + R_4(u)$$

with a remainder term

$$R_4(u) = F_0(u)\big(\Delta_0^k u\cdot\mathcal{F}(u)\big) + R_1(u) + R_2(u) + R_3(u) - 2kF'(u)\{D_i u\}D_i Q_0\Delta_0^{k-1}\widetilde{\mathcal{G}}(u).$$

Hence, using

$$\big\|F_0(u)\big(\Delta_0^k u\cdot\mathcal{F}(u)\big)\big\|_0 \le C\|u\|_s\|\mathcal{F}(u)\|_{s_0-3}$$

and the above estimates for R_1, R_2, R_3, we obtain

$$\|R_4(u)\|_0 \le C\big(\|\mathcal{G}(u)\|_{s-1} + \|u\|_s\|\mathcal{F}(u)\|_{s_0-3}\big). \tag{4.13}$$

Recall that $\mathcal{G}(u)$ depends linearly on γ. Slightly abusing notation, we write $\mathcal{G}(u)\gamma$ etc. in the remaining part of this proof (see (6.6)). Note that in analogy to (3.10), we get, for sufficiently smooth u,

$$D^\alpha\mathcal{G}(u)\gamma = \sum c_{\beta_1,\dots,\beta_{k+1}}\mathcal{G}^{(k)}(u)\{D^{\beta_1}u,\dots,D^{\beta_k}u\}D^{\beta_{k+1}}\gamma$$

with (3.11) holding for k and the multi-indices $\beta_1, \dots, \beta_{k+1}$. Applying this differentiation rule, we get

$$\Delta_0^k \mathcal{G}(u)\gamma = \mathcal{G}'(u)\big\{\Delta_0^k u\big\}\gamma + \mathcal{G}_1(u)\gamma + R_5(u),$$

where $\mathcal{G}_1(u)$ contains all terms where derivatives of u of order $s-1$ and $s-2$ occur. Consequently, it is a sum of terms of the forms

$$\mathcal{G}'(u)\big\{\Delta_0^j D_i \Delta_0^{k-1-j} u\big\} D_i\gamma, \quad \mathcal{G}''(u)\big\{D_i u, \Delta_0^j D_i \Delta_0^{k-1-j} u\big\}\gamma,$$

with $j \in \{0, \dots, k-1\}$, and

$$\mathcal{G}'(u)\{z\}D_iD_l\gamma, \quad \mathcal{G}''(u)\{D_iu, z\}D_l\gamma, \quad \mathcal{G}''(u)\{D_iD_lu, z\}\gamma, \quad \mathcal{G}'''(u)\{D_iu, D_lu, z\}\gamma,$$

with $z := \Delta_0^j D_i \Delta_0^\mu D_l \Delta_0^{k-2-j-\mu} u$ and $j, \mu \in \{0, \dots, k-2\}$, $j + \mu \le k-2$. Using the estimates (3.3) and the assumption $s \ge s_0 > (m+7)/2$, one obtains from analogous arguments as in the proof of Proposition 3.5 (ii) that

$$\|R_5(u)\|_0 \le C\|u\|_{s-1}. \tag{4.14}$$

Writing in the above terms,

$$\Delta_0^{k-1-j} u \approx (\Delta_0^+)^{j+1} \Delta_0^k u, \quad \Delta_0^{k-2-j-\mu} u \approx (\Delta_0^+)^{j+\mu+2} \Delta_0^k u \tag{4.15}$$

up to smoothing remainder terms, we get

$$\Delta_0^k \mathcal{G}(u)\gamma = \mathcal{G}'(u)\big\{\Delta_0^k u\big\}\gamma + \mathcal{G}_2(u)\{\Delta_0^k u\}\gamma + R_6(u) \tag{4.16}$$

with a first-order operator $v \mapsto \mathcal{G}_2(u)\{v\}\gamma$ and a remainder term $R_6(u)$ satisfying the estimate (4.14) again. Hence, using that the linearization of the mean curvature $H(u)$ has $\Delta(u)(P(u)v)$ as its main part, i.e.,

$$\mathcal{G}'(u)\{v\}\gamma = \gamma\Delta(u)P(u)v + \mathcal{G}_3(u)\{v\}\gamma \tag{4.17}$$

with a first-order operator $v \mapsto \mathcal{G}_3(u)\{v\}\gamma$, we get the representation (4.7) with

$$G_1(u)v := \mathcal{G}_2(u)\{v\}\gamma + \mathcal{G}_3(u)\{v\}\gamma - v \cdot \mathcal{F}(u)$$

and with the remainder terms

$$R_a(u) := R_6(u), \quad R_b(u) := R_4(u) + F_0(u)\big(R_6(u)\big).$$

Now the estimate (4.9) of R_a coincides with (4.14), whereas the estimate (4.10) of R_b follows from

$$\begin{aligned}\|R_b(u)\|_0 &\le C\big(\|R_4(u)\|_0 + \|R_6(u)\|_0\big)\\ &\le C\big(\|\mathcal{G}(u)\|_{s-1} + \|u\|_s \|\mathcal{F}(u)\|_{s_0-3} + \|u\|_{s-1}\big)\end{aligned}$$

by (4.13), (4.14) and

$$\begin{aligned}\|\mathcal{G}(u)\|_{s-1} &\le C\big(\|\Delta_0^k \mathcal{G}(u)\|_{-1} + \|\mathcal{G}(u)\|_0\big)\\ &= C\big(\|\tilde{G}(u)(\Delta_0^k u) + R_6(u)\|_{-1} + \|\mathcal{G}(u)\|_0\big)\\ &\le C\big(\|\tilde{G}(u)(\Delta_0^k u)\|_{-1} + \|u\|_{s-1}\big).\end{aligned}$$

The statements (4.8) are consequences of Lemma 3.4 and (3.4). This becomes clear if G_1 is written out explicitly in terms of differential operators and Fréchet derivatives of A. □

Now fix $s \geq s_0$ with $s = 2k$, $k \in \mathbb{N}$. Taking F_0 and G_1 as in Lemma 4.1 we set for $u \in U_s$ that

$$\tilde{\mathcal{F}}(u)v := \big(F(u) + F_0(u)\big)\tilde{G}(u)v.$$

We are now ready to define the variable inner product for which we will prove the semiboundedness of $\mathcal{F}$. For $u \in U_s$ let $M(u)$ be the operator defined by

$$M(u)v := M_0(u)v + \tilde{M}_0(u)v, \quad \tilde{M}_0(u)v = M_1(u)P(u)v + N(u)M_2(u)v. \tag{4.18}$$

Here, the main part M_0 of M is given by

$$M_0(u)v := v - \Lambda(u)A(u)P(u)v \tag{4.19}$$

with A from (2.9) (cf. [10]), whereas the lower order terms are given by

$$M_1(u)w := -M_0(u)F_0(u)A(u)w, \tag{4.20}$$

$$M_2(u)v := \Delta(u)^+\big(\gamma^{-1}G_1(u)v\big). \tag{4.21}$$

From (3.5) and (4.8) we get

$$\begin{aligned} M_0(u) &\in \mathcal{L}\big(H^t(S,\mathbb{R}^m), H^t(S,\mathbb{R}^m)\big), \quad -4 \leq t \leq s-2,\\ M_1(u) &\in \mathcal{L}\big(H^t(S), H^{t+1}(S,\mathbb{R}^m)\big), \quad -2 \leq t \leq s-3,\\ M_2(u) &\in \mathcal{L}\big(H^t(S,\mathbb{R}^m), H^{t+1}(S)\big), \quad -2 \leq t \leq s-3. \end{aligned}$$

The operators depend smoothly on $u \in U_s$ and have uniformly bounded norms as u varies in $\widetilde{U}_s$.

Because of $P(u)\Lambda(u) = 0$, the operator $M_0(u)$ constitutes an isomorphism in $L^2(S,\mathbb{R}^m)$ with inverse

$$M_0(u)^{-1}v = v + \Lambda(u)A(u)P(u)v. \tag{4.22}$$

In particular, we have

$$c\|v\|_0 \leq \|M_0(u)v\|_0 \leq C\|v\|_0, \tag{4.23}$$

$$c\|v\|_0 - C\|v\|_{-1} \leq \|M(u)v\|_0 \leq C\|v\|_0 \tag{4.24}$$

with suitable positive constants c, C independent of $u \in \tilde{U}_s$ and $v \in L^2$.

The main motivation for our choice of M_0 is the easily checked identity

$$\big(M_0(u)F(u)f, M_0(u)v\big)_0 = \big(B(u)f, P(u)v\big)_0, \tag{4.25}$$

where $f \mapsto B(u)f := P(u)(F(u)f)$ is the Dirichlet–Neumann operator. Note that on the left we have an inner product for a vector-valued function while there is an inner product for scalar functions on the right; this will enable us to make use of the coercivity in the normal direction. The choice of M_1 and M_2 results from the need to control certain lower order terms that will appear in our estimates, in particular, (4.34).

For the sake of completeness, we gather some properties of B which we will need in what follows. We will use the commutator notation $[Q_1, Q_2] := Q_1Q_2 - Q_2Q_1$ for operators, in particular, if f is a function we will write $[f, Q]w := fQw - Q(fw)$. Note that property (b) is in fact the L^2-symmetry of $B(u)$ with respect to the measure induced from Γ_u.

LEMMA 4.2. *Assume $u \in \tilde{U}_s$, $f \in C^1(S)$, $w \in H^2(S)$, $v \in H^1(S)$. Then:*

(i) *If $f \geq \alpha > 0$, then*

$$\int_S fwB(u)w\,dS \geq c\|w\|_{1/2}^2 - C\|w\|_0^2$$

with $c = c(\alpha) > 0$, $C = C(\|f\|_{C^1})$. Moreover,

(ii)

$$\int_S w\,B(u)v\,dS = \int_S \omega(u)v\,B(u)(\omega(u)^{-1}w)\,dS,$$

(iii)

$$\|B(u)w\|_{-2} \leq C\|w\|_{-1},$$

(iv)

$$\|[f, B(u)]w\|_0 \leq C\|w\|_0$$

with $C = C(\|f\|_{C^1})$,

(v)

$$\|[\Lambda_i(u), B(u)]w\|_0 \leq C\|w\|_1.$$

All constants are independent of $u \in \tilde{U}_s$.

Proof. (i) As in the proof of Lemma 3.4 we extend u to a diffeomorphism from E to Ω_u and denote the coefficients of the corresponding induced metric by g^{ij} and the corresponding volume element by $\sqrt{g}$. Let ν denote the outer unit normal on S and let $\mathcal{E}$ denote the harmonic extension from S into E. Let ϕ be the solution of the Dirichlet problem

$$L(u)\phi := \partial_i(\sqrt{g}g^{ij}\partial_j\phi) = 0 \text{ in } E, \quad \phi|_S = w.$$

Then

$$B(u)w = \omega(u)^{-1}\nu_i\sqrt{g}g^{ij}\partial_j\phi,$$

and by integration by parts

$$\begin{aligned}
&\int_S fw\,B(u)w\,dS \\
&= \int_S f\phi\omega(u)^{-1}\nu_i\sqrt{g}g^{ij}\partial_j\phi\,dS = \int_E \partial_i(\mathcal{E}(f\omega(u)^{-1})\phi\sqrt{g}g^{ij}\partial_j\phi)\,dx \\
&= \int_E \mathcal{E}(f\omega(u)^{-1})\sqrt{g}g^{ij}\partial_i\phi\partial_j\phi\,dx + \int_E \partial_i(\mathcal{E}(f\omega(u)^{-1})\sqrt{g}g^{ij})\phi\partial_j\phi\,dx \\
&\geq c\|\phi\|_1^{E^2} - C\|\phi\|_1^E\|\phi\|_0^E \geq c\|\phi\|_1^{E^2} - C\|\phi\|_0^{E^2} \geq c\|w\|_{1/2}^2 - C\|w\|_0^2.
\end{aligned}$$

The uniformity of these estimates with respect to $u \in \tilde{U}_s$ follows by a compactness argument as in [10].

(ii) The assertion follows from transforming the integral to Γ_u, applying Green's formula and transforming back.

(iii) Using (ii), the assertion follows from a standard duality argument and the fact that $B(u) \in \mathcal{L}(H^2(S), H^1(S))$.

(iv) Maintaining the notation from the proof of (i), we have

$$[f, B(u)]w = \omega(u)^{-1}\nu_i\sqrt{g}g^{ij}(f\partial_j\phi - \partial_j\psi),$$

where ψ satisfies

$$L(u)\psi = 0 \text{ in } E, \quad \psi|_S = fw = f\phi|_S.$$

Therefore, by estimates parallel to Lemma 4.3 in [10],

$$\begin{aligned}\|[f, B(u)]w\|_{1/2} &\le \|\omega(u)^{-1}\nu_i\sqrt{g}g^{ij}\partial_j(\phi\mathcal{E}f - \psi)\|_{1/2} + C\|\phi\|_{1/2}\\ &\le C\left(\|L(u)(\phi\mathcal{E}f)\|_0^E + \|\phi\|_{1/2}\right) \le C\left(\|\phi\|_1^E + \|\phi\|_{1/2}\right) \le C\|w\|_{1/2}.\end{aligned}$$

As both multiplication by f and $B(u)$ are symmetric with respect to the L^2-inner product induced from Γ_u, we get by duality that

$$\|[f, B(u)]w\|_{-1/2} \le C\|w\|_{-1/2},$$

and the result follows by interpolation.

(v) We have, by the chain rule for the operators D_k,

$$\begin{aligned}[\Lambda_i(u), B(u)] &= [\alpha_k^i(u)D_k, B(u)] = [\alpha_k^i(u), B(u)]D_k + \alpha_k^i(u)[D_k, B(u)]\\ &= [\alpha_k^i(u), B(u)]D_k + \alpha_k^i(u)B'(u)\{D_ku\}.\end{aligned}$$

The result now follows from (ii) and the estimate

$$\|B'(u)\{D_ku\}w\|_0 \le C\|w\|_1,$$

which is a simple consequence of (3.1). □

The next lemma will be crucial in the proof of the main estimate as it will provide coercivity for the normal component.

LEMMA 4.3. *There are positive constants c, C such that*

$$(\Delta(u)^+(\gamma^{-1}w), B(u)w)_0 \le -c\|w\|_{-1/2}^2 + C\|w\|_{-2}^2$$

for all $u \in \tilde{U}_s$, $w \in H^1(S)$.

Proof. Set $z := \Delta(u)^+(\gamma^{-1}w)$. Then $w = \gamma\Delta(u)z + \gamma Q(u)(\gamma^{-1}z)$; see (4.6). By Lemma 4.2 (ii),

$$\begin{aligned}I &:= \left(\Delta(u)^+(\gamma^{-1}w), B(u)w\right)_0 \le \left(z, B(u)\gamma\Delta(u)z\right)_0 + \|z\|_1\|B(u)(\gamma Q(u)(\gamma^{-1}z))\|_{-1}\\ &\le \left(\omega(u)\gamma\Delta(u)z, B(u)(\omega(u)^{-1}z)\right)_0 + C\|w\|_{-1}^2.\end{aligned}$$

Setting $\tilde{z} := \omega(u)^{-1}z$, $\tilde{\gamma} := \omega(u)^2\gamma$ and using (4.5) we get

$$\begin{aligned}I &\le \left(\tilde{\gamma}\Delta(u)\tilde{z}, B(u)\tilde{z}\right)_0 + \left(\omega(u)\gamma[\omega(u), \Delta(u)]\tilde{z}, B(u)\tilde{z}\right)_0 + C\|w\|_{-1}^2\\ &\le \left(\tilde{\gamma}\Lambda_i(u)\Lambda_i(u)\tilde{z}, B(u)\tilde{z}\right)_0 + C\|z\|_1^2 + C\|w\|_{-1}^2.\end{aligned}$$

By integration by parts, one obtains an estimate

$$\left|\int_S \Lambda_i(u) f\, dS\right| \le C \int_S |f|\, dS,$$

cf. [10, (5.5)]. This yields

$$\begin{aligned} I &\le -(\Lambda_i(u)\tilde z, \Lambda_i(u)\tilde\gamma B(u)\tilde z)_0 + C\|z\|_1^2 + C\|w\|_{-1}^2 \\ &\le -(\Lambda_i(u)\tilde z, \tilde\gamma\Lambda_i(u) B(u)\tilde z)_0 + C\|z\|_1 \sum_i \|[\Lambda_i(u),\gamma]B(u)\tilde z\|_0 + C\|w\|_{-1}^2 \\ &\le -(\Lambda_i(u)\tilde z, \tilde\gamma B(u)\Lambda_i(u)\tilde z)_0 + C\|z\|_1 \sum_i \|[B(u),\Lambda_i(u)]\tilde z\|_0 + C\|w\|_{-1}^2 \\ &\le -c\|\Lambda_i(u)\tilde z\|_{1/2}^2 + C\|z\|_1^2 + C\|w\|_{-1}^2 \\ &\le -c\|\tilde z\|_{3/2}^2 + C\|w\|_{-1}^2 \le -c\|w\|_{-1/2}^2 + C\|w\|_{-2}^2, \end{aligned}$$

where parts (i) and (v) of Lemma 4.2 have been used together with interpolation in the scale $H^t(S)$. □

In view of (4.23), (4.24) for every fixed $u \in U_s$, $s \ge s_0$, $s = 2k$, $k \in \mathbb{N}$, and λ sufficiently large (independent of $u \in \widetilde{U}_s$),

$$(v,w)_{s,u} := \lambda\big(M_0(u)v, M_0(u)w\big)_0 + \big(M(u)\Delta_0^k v, M(u)\Delta_0^k w\big)_0 \tag{4.26}$$

defines a scalar product on $H^s(S,\mathbb{R}^m)$ which is equivalent to the usual one.

The next two lemmas provide properties of the inner product $(\cdot,\cdot)_{s,u}$, which will be used when we apply the abstract existence result of Theorem 5.2 to our situation. They are parallel to Lemmas 5.3 and 5.4 in [10], therefore the proofs are omitted here. Note the uniformity of all estimates with respect to $u \in \tilde U_s$.

LEMMA 4.4. *Assume $s \ge s_0$.*

(i) *There exists a $C > 0$ such that for all $v \in H^{s+3}(S,\mathbb{R}^m)$, $w \in H^s(S,\mathbb{R})$, $u \in \tilde U_s$*

$$(v,w)_{s,u} \le C\|v\|_{s+3}\|w\|_{s-3}.$$

(ii) *There exist λ_0, $c_0 > 0$ such that for all $v \in H^{s+6}(S,\mathbb{R}^m)$, $\lambda \ge \lambda_0$*

$$\big(v, (-\Delta_0^3 + \lambda)v\big)_{s,u} \ge c_0\|v\|_{s+3}^2.$$

As an immediate consequence of Lemma 4.4 (i) we get the existence of a continuous bilinear form $\langle\cdot,\cdot\rangle_{s,u}$ on $H^{s+3}(S,\mathbb{R}^m) \times H^{s-3}(S,\mathbb{R}^m)$ compatible with $(\cdot,\cdot)_{s,u}$, i.e., there holds $\langle v,w\rangle_{s,u} = (v,w)_{s,u}$ for all $v,w \in H^{s+3}(S,\mathbb{R}^m)$. Further, we put for $\varepsilon \in (0,1]$

$$\langle v,w\rangle_{s,u}^{\varepsilon} := \langle v,w\rangle_{s_0,u} + \varepsilon^2\langle v,w\rangle_{s,u}. \tag{4.27}$$

LEMMA 4.5. *We assume as above that $s \ge s_0$, $\varepsilon \in (0,1]$.*

(i) *For fixed $u \in U_s$, the mapping $\langle\cdot,\cdot\rangle_{s,u}^{\varepsilon} : H^{s+3}(S,\mathbb{R}^m) \times H^{s-3}(S,\mathbb{R}^m) \to \mathbb{R}$ constitutes a continuous, nondegenerate bilinear form whose restriction to $H^{s+3}(S,\mathbb{R}^m) \times H^{s+3}(S,\mathbb{R}^m)$ is symmetric.*

(ii) *With constants $C > 0$ independent of ε, u, v, w, one has for $u, w \in \tilde U_s$ and $v \in H^{s+3}(S,\mathbb{R}^m)$:*

$$C^{-1}\big(\|v\|_{s_0}^2 + \varepsilon^2\|v\|_s^2\big) \le \langle v,v\rangle_{s,u}^{\varepsilon} \le C\big(\|v\|_{s_0}^2 + \varepsilon^2\|v\|_s^2\big), \tag{4.28}$$

$$\langle v,v\rangle_{s,u}^{\varepsilon} \le \langle v,v\rangle_{s,w}^{\varepsilon}\big(1 + C\|u-w\|_{s_0-3}\big). \tag{4.29}$$

(iii) *Weak convergences* $u_n \rightharpoonup u$ *in* H^s, $w_n \rightharpoonup w$ *in* H^{s-3} *imply*

$$\langle v, w_n\rangle^{\varepsilon}_{s,u_n} \to \langle v, w\rangle^{\varepsilon}_{s,u}$$

for all $v \in H^{s+3}$.

Now we are prepared to formulate and prove the following a priori estimates for $\mathcal{F}$ w.r.t. the bilinear forms $(\cdot,\cdot)_{s,u}$.

PROPOSITION 4.6. *Let* $s \geq s_0$ *be an even integer. Then*

$$\big(u, \mathcal{F}(u)\big)_{s,u} \leq C\|u\|_s^2 \tag{4.30}$$

for all $u \in \widetilde{U}_s \cap C^\infty(S, \mathbb{R}^m)$ *with a constant* C *independent of* u.

Proof. For later use, we prove the estimate in the following stronger form: For every $\varepsilon > 0$ there exists a constant $C(\varepsilon)$ such that

$$\big(u, \mathcal{F}(u)\big)_{s,u} \leq C\|u\|_s\big((\varepsilon + \|\mathcal{F}(u)\|_{s_0-3})\|u\|_s + C(\varepsilon)\|u\|_{s-1}\big). \tag{4.31}$$

Setting $v := \Delta_0^k u$ and using the notations of Lemma 4.1 we have

$$\big(M(u)\Delta_0^k u, M(u)\Delta_0^k \mathcal{F}(u)\big)_0 = I(u)v^2 + J(u) + \big(M(u)v, M(u)R_b(u)\big)_0$$

with

$$I(u)v^2 := \big(M(u)v, M(u)\tilde{\mathcal{F}}(u)v\big)_0 \quad J(u) := \big(M(u)v, M(u)F(u)R_a(u)\big)_0.$$

From (4.10) we obtain

$$\big(M(u)v, M(u)R_b(u)\big)_0 \leq C\|u\|_s\big(\|\tilde{G}(u)v\|_{-1} + \|\mathcal{F}(u)\|_{s_0-3}\|u\|_s + \|u\|_{s-1}\big).$$

To estimate $J(u)$, we write this term as $J_1(u) + \cdots + J_4(u)$ with

$$J_1(u) = \big(M_0(u)v, M_0(u)F(u)R_a(u)\big)_0, \quad J_2(u) = \big(\tilde{M}_0(u)v, M_0(u)F(u)R_a(u)\big)_0,$$
$$J_3(u) = \big(M_0(u)v, \tilde{M}_0(u)F(u)R_a(u)\big)_0, \quad J_4(u) = \big(\tilde{M}_0(u)v, \tilde{M}_0(u)F(u)R_a(u)\big)_0.$$

Using (4.25) and (4.9), we obtain for J_1

$$J_1(u) = \big(B(u)R_a(u), P(u)v\big)_0 \leq C\|R_a(u)\|_0\|P(u)v\|_1 \leq C\|u\|_{s-1}\|P(u)v\|_1.$$

As

$$P(u)v = \Delta(u)^+\big(\gamma^{-1}(\tilde{G}(u)v - G_1(u)v)\big) - Q(u)(P(u)v) \tag{4.32}$$

we see from (4.8) that

$$\|P(u)v\|_1 \leq C\big(\|\tilde{G}(u)v\|_{-1} + \|v\|_0\big), \tag{4.33}$$

and consequently

$$J_1(u) \leq C\|u\|_{s-1}\big(\|\tilde{G}(u)v\|_{-1} + \|u\|_s\big).$$

For J_2 we have

$$J_2(u) \leq C\|\tilde{M}_0(u)v\|_1\|M_0(u)F(u)R_a(u)\|_{-1} \leq C\|v\|_0\|R_a(u)\|_0 \leq C\|u\|_s\|u\|_{s-1},$$

and the same estimates are valid for J_3, J_4, thus

$$J(u) \le C\|u\|_{s-1}\big(\|\tilde G(u)v\|_{-1} + \|u\|_s\big).$$

Further, we decompose I according to $I(u)v^2 = I_1(u)v^2 + \cdots + I_4(u)v^2$ in the same manner as J, i.e.,

$$\begin{aligned} I_1(u)v^2 &= \big(M_0(u)v, M_0(u)\tilde{\mathcal{F}}(u)v\big)_0, \quad I_2(u)v^2 = \big(\tilde M_0(u)v, M_0(u)\tilde{\mathcal{F}}(u)v\big)_0,\\ I_3(u)v^2 &= \big(M_0(u)v, \tilde M_0(u)\tilde{\mathcal{F}}(u)v\big)_0, \quad I_4(u)v^2 = \big(\tilde M_0(u)v, \tilde M_0(u)\tilde{\mathcal{F}}(u)v\big)_0 \end{aligned}$$

and each term is estimated separately. We start with the leading order term $I_1(u)v^2$. Using (4.25) again, it can be written as

$$\begin{aligned} I_1(u)v^2 &= \big(M_0(u)v, M_0(u)F(u)\tilde G(u)v\big)_0 + \big(M_0(u)v, M_0(u)F_0(u)\tilde G(u)v\big)_0\\ &= \big(P(u)v, B(u)\tilde G(u)v\big)_0 + \big(M_0(u)v, M_0(u)F_0(u)\tilde G(u)v\big)_0. \end{aligned}$$

In the first summand we insert (4.32) and use

$$\big|\big(Q(u)P(u)v, B(u)\tilde G(u)v\big)_0\big| \le C\|v\|_0\|\tilde G(u)v\|_{-1},$$

and by Lemma 4.3

$$\big(\Delta(u)^+(\gamma^{-1}\tilde G(u)v), B(u)\tilde G(u)v\big)_0 \le -c_0\|\tilde G(u)v\|^2_{-1/2} + C\|\tilde G(u)v\|^2_{-1}.$$

Remembering the definitions (4.20), (4.21) of M_1 and M_2, we have

$$\begin{aligned} \big(M_0(u)v, M_0(u)F_0(u)\tilde G(u)v\big)_0 &= -\big(M_0(u)v, M_1(u)B(u)\tilde G(u)v\big)_0,\\ \big(\Delta(u)^+(\gamma^{-1}G_1(u)v), B(u)\tilde G(u)v\big)_0 &= \big(M_2(u)v, B(u)\tilde G(u)v\big)_0, \end{aligned}$$

and consequently we arrive at

$$\begin{aligned} I_1(u)^2v \le &-\big(M_2(u)v, B(u)\tilde G(u)v\big)_0 - \big(M_0(u)v, M_1(u)B(u)\tilde G(u)v\big)_0\\ &- c_0\|\tilde G(u)v\|^2_{-1/2} + C\big(\|v\|_0 + \|\tilde G(u)v\|_{-1}\big)\|\tilde G(u)v\|_{-1}. \end{aligned} \tag{4.34}$$

This coercive estimate of $I_1(u)v^2$ with respect to $\tilde G(u)v$ is the crucial "leading order estimate" here and will be used below to control the lower order terms.

From Lemma 4.2 (iii) we get $F \in \mathcal{L}(H^{-1}(S), H^{-2}(S,\mathbb{R}^m))$ and therefore for I_4 we have the estimate

$$\begin{aligned} |I_4(u)v^2| &\le C\|\tilde M_0(u)v\|_1\|\tilde M_0(u)\tilde{\mathcal{F}}(u)v\|_{-1}\\ &\le C\|v\|_0\|(F(u)+F_0(u))\tilde G(u)v\|_{-2} \le C\|v\|_0\|\tilde G(u)v\|_{-1}, \end{aligned} \tag{4.35}$$

where (4.8) has been applied again. Further, concerning I_3 we have

$$I_3(u)v^2 = \big(M_0(u)v, \tilde M_0F(u)\tilde G(u)v\big)_0 + \big(M_0(u)v, \tilde M_0F_0(u)\tilde G(u)v\big)_0,$$

where the last summand allows the estimate

$$\big|\big(M_0(u)v, \tilde M_0(u)F_0(u)\tilde G(u)v\big)_0\big| \le C\|v\|_0\|\|\tilde G(u)v\|_{-1}.$$

Remembering $\tilde M_0(u) = M_1(u)P(u) + N(u)M_2(u)$, the first summand is written

$$\begin{aligned} &\big(M_0(u)v, \tilde M_0(u)F(u)\tilde G(u)v\big)_0\\ &\qquad = \big(M_0(u)v, M_1(u)B(u)\tilde G(u)v\big)_0 + \big(P(u)v, M_2(u)F(u)\tilde G(u)v\big)_0. \end{aligned}$$

Using (4.33) we get

$$\begin{aligned}\big|\big(P(u)v, M_2(u)F(u)\tilde G(u)v\big)_0\big| &\le C\|P(u)v\|_1\|\tilde G(u)v\|_{-1}\\ &\le C\big(\|v\|_0+\|\tilde G(u)v\|_{-1}\big)\|\tilde G(u)v\|_{-1},\end{aligned}$$

and consequently

$$I_3(u)v^2 \le \big(M_0(u)v, M_1(u)B(u)\tilde G(u)v\big)_0 + C\big(\|v\|_0+\|\tilde G(u)v\|_{-1}\big)\|\tilde G(u)v\|_{-1}. \tag{4.36}$$

Arguing along the same lines for I_2 we obtain

$$I_2(u)v^2 = \big(\tilde M_0(u)v, M_0(u)F(u)\tilde G(u)v\big)_0 + \big(\tilde M_0(u)v, M_0(u)F_0(u)\tilde G(u)v\big)_0,$$

where again

$$\big|\big(\tilde M_0(u)v, M_0F_0(u)\tilde G(u)v\big)_0\big| \le C\|v\|_0\|\tilde G(u)v\|_{-1}$$

and

$$\begin{aligned}&\big(\tilde M_0(u)v, M_0(u)F(u)\tilde G(u)v\big)_0\\ &\qquad = \big(M_2(u)v, B(u)\tilde G(u)v\big)_0 + \big(M_1(u)P(u)v, M_0(u)F(u)\tilde G(u)v\big)_0\end{aligned}$$

with

$$\begin{aligned}&\big|\big(M_1(u)P(u)v, M_0(u)F(u)\tilde G(u)v\big)_0\big|\\ &\qquad\le \|M_1(u)P(u)v\|_2\|M_0(u)F(u)\tilde G(u)v\|_{-2} \le C\|P(u)v\|_1\|\tilde G(u)v\|_{-1}\\ &\qquad\le C\big(\|v\|_0+\|\tilde G(u)v\|_{-1}\big)\|\tilde G(u)v\|_{-1}.\end{aligned}$$

Thus we have

$$I_2(u)v^2 \le \big(M_2(u)v, B(u)\tilde G(u)v\big)_0 + C\big(\|v\|_0+\|\tilde G(u)v\|_{-1}\big)\|\tilde G(u)v\|_{-1}. \tag{4.37}$$

Summarizing, we get

$$\begin{aligned}(u,\mathcal F(u))_{s,u} \le -c_1\|\tilde G(u)v\|^2_{-1/2} + C_1\big(&\|\tilde G(u)v\|^2_{-1} + \|u\|_s(\|\tilde G(u)v\|_{-1}\\ &+\|\mathcal F(u)\|_{s_0-3}\|u\|_s + \|u\|_{s-1})\big)\end{aligned}$$

and, estimating further,

$$\begin{aligned}\|u\|_s\|\tilde G(u)v\|_{-1} &\le \frac{\varepsilon}{C_1}\|u\|^2_s + C_2(\varepsilon)\|\tilde G(u)v\|^2_{-1},\\ \|\tilde G(u)v\|^2_{-1} &\le \frac{c_1}{C_1(1+C_2(\varepsilon))}\|\tilde G(u)v\|^2_{-1/2} + C(\varepsilon)\|\tilde G(u)v\|^2_{-3},\\ \|\tilde G(u)v\|^2_{-3} &\le C\|u\|^2_{s-1}\end{aligned}$$

we obtain (4.31). □

Remark 4.7. Reinspecting the estimates in the previous proofs it is straightforward to check that for fixed $s \ge s_0$ the occurring constants, in particular in (4.30), (4.31), are independent of γ as long as γ varies in some fixed set

$$\big\{\gamma \in C^\infty(S) \mid \gamma \ge \gamma^* > 0, \|\gamma\|_{s_1} \le M\big\} \tag{4.38}$$

with some sufficiently large $s_1 = s_1(s)$.

Using Lemma 3.6, we write for $u \in \tilde{U}_s$, $v \in H^s$, $s \geq s_0$

$$\mathcal{F}'(u)v = F(u)\big(\gamma\Delta(u)(P(u)v) + G_2(u)v\big) + R(u)v, \tag{4.39}$$

where $v \mapsto G_2(u)v := \mathcal{G}_3(u)v - \mathcal{F}(u) \cdot v$ and $v \mapsto R(u)v$ are operators of order one and zero, respectively. Note that $\mathcal{F}'(u)$ coincides with $\tilde{\mathcal{F}}(u)$ if G_1 is replaced by G_2 and F_0 is replaced by 0. Hence, by defining (cf. (4.18)–(4.21))

$$M_3(u) := \Delta(u)^+(\gamma^{-1}G_2(u)v), \quad \tilde{M}(u) := M_0(u) + M_3(u),$$

we find that $\tilde{M}$ has the same properties as M above, and we obtain, parallel to (4.31), the following estimate which will be used in the uniqueness proof.

LEMMA 4.8. *Let $s \geq s_0$. Then there exists a constant C such that for all $u \in \tilde{U}_s$, $v \in H^s$ we have*

$$\big(\tilde{M}(u)\mathcal{F}'(u)v, \tilde{M}(u)v\big)_0 \leq C\|v\|_0^2.$$

5. Proof of short time existence and uniqueness. We are now ready to prove our main results as announced in Theorems 2.1 and 2.2. As the existence proof is in some respects analogous to the corresponding considerations in [10], we restrict ourselves to an outline and refer to that paper for details.

Fix an even integer $s_0 > (m+7)/2$, $s_0 \geq 6$, and let $s \geq s_0$ be an even integer as well. Let $\tilde{U}_s$ be defined as above. The notations $C_w([0,T],X)$ and $C_w^1([0,T],X)$ will denote the spaces of weakly continuous and weakly continuously differentiable functions, respectively, with values in some subset X of a normed space.

At first an estimate which provides uniqueness and Lipschitz continuous dependence on the initial value in the L^2-norm is given.

PROPOSITION 5.1. *Let $u, v \in C_w([0,T],\tilde{U}_{s_0}) \cap C_w^1([0,T],H^{s_0-3}(S,\mathbb{R}^m))$ be two solutions of* (2.2). *Then*

$$\|v(t) - u(t)\|_0 \leq C\|v(0) - u(0)\|_0 \tag{5.1}$$

with C depending only on $\tilde{U}_{s_0}$ and on T.

Proof. Set $w(t) := v(t) - u(t)$ and note that

$$w(t) \in C([0,T],H^\sigma(S,\mathbb{R}^m)) \cap C^1([0,T],H^{\sigma-3}(S,\mathbb{R}^m))$$

for $\sigma < s_0$. In particular, the map

$$t \mapsto g(t) := \|\tilde{M}(u(t))w(t)\|_0^2$$

is differentiable and has the derivative

$$\begin{aligned} g'(t) = {} & 2\big(\tilde{M}'(u(t))\{u'(t)\}w(t), \tilde{M}(u(t))w(t)\big)_0 \\ & + 2\big(\tilde{M}(u(t))(\mathcal{F}(v(t)) - \mathcal{F}(u(t))), \tilde{M}(u(t))w(t)\big)_0. \end{aligned}$$

To estimate the first term we note that, parallel to the estimates in Lemma 3.1,

$$\begin{aligned} \|\tilde{M}'(u(t))\{u'(t)\}w(t)\|_0 &\leq C\|u'(t)\|_{s_0-3}\|w(t)\|_0 \\ &\leq C\|\mathcal{F}(u(t))\|_{s_0-3}\|w(t)\|_0 \leq C\|w(t)\|_0. \end{aligned}$$

The second term can be estimated by using

$$\mathcal{F}(v(t)) - \mathcal{F}(u(t)) = \mathcal{F}'(u(t))w(t) + R,$$

where

$$R := \int_0^1 \int_0^\tau \mathcal{F}''(\theta v(t) + (1-\theta)u(t))\{w(t), w(t)\}\, d\theta d\tau$$

allows an estimate

$$\|R\|_0 \le C\|w(t)\|_3 \|w(t)\|_{s_0-3} \le C\|w(t)\|_{s_0}\|w(t)\|_0 \le C\|w(t)\|_0,$$

where estimates on $\mathcal{F}''$ parallel to Lemma 3.1 and norm convexity have been used. Thus, by Lemma 4.8,

$$\begin{aligned} g'(t) &\le 2\big(\tilde{M}(u(t))\mathcal{F}'(u(t))w(t), \tilde{M}(u(t))w(t)\big)_0 + C\|w(t)\|_0^2 \\ &\le \|w(t)\|_0^2 \le Cg(t). \end{aligned}$$

Therefore, by Gronwall's inequality,

$$\|v(t) - u(t)\|_0^2 \le Cg(t) \le Cg(0) \le C\|v(0) - u(0)\|_0^2. \qquad \square$$

To prove existence, we will rely on an abstract existence theorem whose proof has been given in [10]. It generalizes an existence theorem concerning evolution equations with semibounded operators by Kato and Lai [11] to the case of variable bilinear forms. The setting is the following:

(H) Let $X \subseteq Y \subseteq Z$ be real and separable Banach spaces with dense and continuous embeddings and $\mathcal{U} \subseteq Y$ be open. For every $u \in \mathcal{U}$, let $\langle \cdot, \cdot \rangle_u : X \times Z \to \mathbb{R}$ be a continuous and nondegenerate bilinear form, such that with fixed constants $C \ge 1$, $M \ge 0$:

- (H1) $\langle v, w\rangle_u = \langle w, v\rangle_u$ for all $v, w \in X$;
- (H2) $C^{-1}\|v\|_Y^2 \le \langle v, v\rangle_u \le C\|v\|_Y^2$ for all $v \in X$, $u \in \mathcal{U}$;
- (H3) $\langle v, v\rangle_u \le \langle v, v\rangle_w\big(1 + M\|u - w\|_Z\big)$ for all $v \in X$, $u, w \in \mathcal{U}$;
- (H4) weak convergences $u_n \rightharpoonup u$ in Y, $u_n, u \in \mathcal{U}$, and $w_n \rightharpoonup w$ in Z imply $\langle v, w_n\rangle_{u_n} \to \langle v, w\rangle_u$ for all $v \in X$.

Assuming (H) holds, by the dense embedding $X \subseteq Y$ and

$$\big|\langle v, w\rangle_u\big|^2 \le \langle v, v\rangle_u \langle w, w\rangle_u \le C^2\|v\|_Y^2\|w\|_Y^2 \quad \text{for} \quad v, w \in X$$

there exists to each $u \in \mathcal{U}$ a scalar product $(\cdot, \cdot)_u$ on Y, which is compatible with $\langle \cdot, \cdot \rangle_u$, i.e., we have

$$(v, w)_u = \langle v, w\rangle_u \quad \text{for} \quad v \in X, w \in Y.$$

Moreover, for $u_n, u \in \mathcal{U}$, $u_n \rightharpoonup u$, $w_n \rightharpoonup w$ in Y implies

$$(v, w_n)_{u_n} \to (v, w)_u \quad \text{for all} \quad v \in X.$$

For the sake of brevity we put

$$\|v\|_u = (v, v)_u^{1/2}, \quad |||u||| = (u, u)_u^{1/2}\ .$$

THEOREM 5.2. *Assume* (H) *is satisfied with some ball*

$$\mathcal{U} = B := \big\{u \in Y \,\big|\, \|u\|_Y < R\big\}, \quad R > 0,$$

and $\mathcal{F} : B \to Z$ *is a weakly sequentially continuous mapping such that*

$$2\langle u, \mathcal{F}(u)\rangle_u + M\,\|\mathcal{F}(u)\,\|_Z |||u||| \le \beta\big(|||u|||^2\big) \ \text{ for all } \ u \in X \cap B \tag{5.2}$$

with a C^1*-function* $\beta : \mathbb{R}_+ \to \mathbb{R}_+ = [0,\infty)$. *Let* $u_0 \in B$,

$$|||u_0||| < r := R/(2C^3)^{1/2},$$

and $T > 0$ *such that the solution* ρ *of the scalar Cauchy problem*

$$d\rho/dt = \beta\big(\rho(t)\big), \quad \rho(0) = |||u_0|||^2 \tag{5.3}$$

exists on $[0,T]$ *and satisfies* $\rho(t) < r^2$ *there. Then the Cauchy problem*

$$u'(t) = \mathcal{F}(u(t)), \qquad u(0) = u_0 \tag{5.4}$$

possesses a solution $u \in C_w([0,T],\mathcal{U}) \cap C^1_w([0,T],Z)$ *for which additionally*

$$|||u(t)|||^2 \le \rho(t) \ \text{ for all } \ t \in [0,T],$$
$$u(t) \to u_0 \ \text{ in } \ Y \ \text{ for } \ t \to +0.$$

Proof of Theorems 2.1, 2.2 *(outline):* Instead of (2.2), (2.3) we consider the Cauchy problem

$$\left.\begin{array}{rcl} v'(t) &=& \hat{\mathcal{F}}(v) := \mathcal{F}(v + w_0), \\ v(0) &=& u_0 - w_0, \end{array}\right\} \tag{5.5}$$

where w_0 is smooth and near u_0.

To apply Theorem 5.2, we set for $\varepsilon \in (0,1]$

$$\begin{array}{ll} X := H^{s+3}(S,\mathbb{R}^m), & \|\cdot\|_X := \|\cdot\|_{s_0+3} + \varepsilon\|\cdot\|_{s+3}, \\ Y := H^{s}(S,\mathbb{R}^m), & \|\cdot\|_X := \|\cdot\|_{s_0} + \varepsilon\|\cdot\|_{s}, \\ Z := H^{s-3}(S,\mathbb{R}^m), & \|\cdot\|_X := \|\cdot\|_{s_0-3} + \varepsilon\|\cdot\|_{s-3}. \end{array}$$

For $u \in \tilde{U}_{s_0}$, let $\langle\cdot,\cdot\rangle^\varepsilon_u$ be the bilinear form compatible to the inner product on Y given by

$$(v,w)^\varepsilon_u := (v,w)_{s_0,u} + \varepsilon^2 (v,w)_{s,u}$$

with $(v,w)_{s_0,u}$, $(v,w)_{s,u}$ given by (4.26). Lemma 4.5 ensures that this bilinear form satisfies the assumptions (H), with constants independent of ε. Thus Theorem 5.2 yields existence of a solution

$$u \in C_w([0,T],\tilde{U}_s) \cap C^1_w([0,T],H^{s-3}(S,\mathbb{R}^m))$$

and an estimate

$$\|u(t)\|_s \le C(1 + \|u_0\|_s) \tag{5.6}$$

with C independent of u_0 and t.

The uniqueness result from Proposition 5.1 enables us to define an evolution operator T_t by setting $T_t u_0 := u(t)$. By a nonlinear interpolation result given in [2, Proposition A.1 and Remark A.2], the estimates (5.1) and (5.6) imply H^τ-continuity of T_t for $\tau \in [0, s)$, uniformly in $t \in [0, T]$. Approximation of the initial value u_0 by $u_0^n \in H^{s+1}$ and of the solution u by the corresponding solutions $u_0^n \in C([0,T], \tilde{U}_s) \cap C^1([0,T], H^{s-3}(S, \mathbb{R}^m))$ then yields

$$u \in C([0,T], \tilde{U}_s) \cap C^1([0,T], H^{s-3}(S, \mathbb{R}^m))$$

by uniform convergence. Finally, the existence time T can be shown to be independent of s by standard continuation arguments. For further details we refer to [10].

6. Nontrivial equilibria and long-time existence. In this section we will investigate the existence of equilibrium points and the long-time dynamic of the evolution problem (2.2). Our considerations are restricted to situations near trivial equilibria; i.e., we will assume that the domain is near a ball and γ is near a constant. Therefore in the following we specialize the reference domain to

$$E := \{x \in \mathbb{R}^m \mid |x| < 1\}, \quad S := \partial E = \{x \in \mathbb{R}^m \mid |x| = 1\}.$$

First we show that for any given domain Ω_u near a ball there exists a corresponding γ such that (u, γ) yields an equilibrium point for (2.2). Of course, the opposite question is more interesting for our evolution: Given a surface energy density, find a corresponding class of equilibrium shapes, and for given initial shape show global existence of the solution in time and convergence to some member of this class. Our proof of this is organized as follows. Using the refined semiboundedness estimate of Proposition 4.6 we obtain weak exponential growth of a solution in higher Sobolev norms provided the solution remains near the trivial equilibrium with respect to some lower norms. This enables us to show that the scalar function $f(t)$, which is defined by (6.11) below, controls the evolution. Then a simple discussion of the spectral properties of the evolution equation for f yields global existence.

We start by stating some simple integral identities needed later on; in particular, assertion (ii) of the following lemma together with volume conservation implies that the center of gravity remains fixed during evolutions under consideration.

LEMMA 6.1. (i) *We have*

$$\int_S \omega(u) N_i(u) u_j \, dS = \delta_{ij} |\Omega_u|, \tag{6.1}$$

where δ_{ij} denotes the Kronecker symbol, and

$$\int_S \omega(u) N(u) \big(\mathcal{G}(u)\gamma\big) \, dS = 0. \tag{6.2}$$

(ii) *For any solution $u = u(t)$ of (2.2) the vector of first moments*

$$M(t) := \int_{\Omega_{u(t)}} x \, dx \tag{6.3}$$

is independent of t.

Proof. (i) After retransformation onto Γ_u with outer normal n, (6.1) follows from

$$\int_{\Gamma_u} n_i x_j \, d\Gamma_u = \int_{\Omega_u} \partial_i x_j \, dx = \delta_{ij} |\Omega_u|,$$

whereas (6.2) reads

$$\int_{\Gamma_u} n(\gamma\kappa - \psi) \, d\Gamma_u = 0,$$

where ψ is harmonic in Ω_u with Neumann boundary condition $\partial_n \psi = \Delta_{\Gamma_u}\gamma$ on Γ_u. By Green's formula we get

$$\int_{\Gamma_u} n\psi \, d\Gamma_u = \int_{\Gamma_u} x \partial_n \psi \, d\Gamma_u;$$

hence, writing $n\kappa = \Delta_{\Gamma_u} x$ on Γ_u we obtain

$$\int_{\Gamma_u} n(\gamma\kappa - \psi) \, d\Gamma_u = \int_{\Gamma_u} \big(\gamma \Delta_{\Gamma_u} x - x \Delta_{\Gamma_u}\gamma\big) \, d\Gamma_u = 0 \tag{6.4}$$

by an integration by parts.
(ii) Consider the solution to (1.1) corresponding to u. We have, using Green's formula and (6.4),

$$\begin{aligned} \dot{M}(t) &= \int_{\Gamma(t)} x V_n \, d\Gamma(t) = \int_{\Gamma(t)} x \, \partial_n \phi \, d\Gamma(t) \\ &= \int_{\Gamma(t)} \partial_n x \, \phi \, d\Gamma(t) = \int_{\Gamma(t)} n(\gamma\kappa - \psi) \, d\Gamma(t) = 0, \end{aligned}$$

which is the assertion. ∎

Remark 6.2. Note that the presence of the correction term G is crucial not only for the generalized gradient flow property but also for the validity of Lemma 6.1.

In further considerations we assume $s \geq s_0$ with $s_0 \in \mathbb{N}$ fixed as in section 4, set

$$\tilde{U}_s := \big\{ u \in H^s(S, \mathbb{R}^m) \;\big|\; \|u - w_0\|_{s_0} \leq \delta_0 \big\}, \quad w_0(x) := x \;\text{ for }\; x \in S \tag{6.5}$$

and assume $\delta_0 > 0$ sufficiently small, whenever necessary. Moreover, to stress the dependency on γ, we consider now $\mathcal{F}(u)$ and $\mathcal{G}(u)$ as linear operators defined by

$$\mathcal{F}(u)v := F(u)\big(\mathcal{G}(u)v\big), \quad \mathcal{G}(u)v := -v\Lambda_i(u)\big(n_i(u) \circ u\big) - A(u)\Delta(u)v; \tag{6.6}$$

for $s \geq s_0$ und $2 \leq t \leq s$ the operators

$$\mathcal{G}(u) \in \mathcal{L}\big(H^t(S), H^{t-1}(S)\big), \quad \mathcal{F}(u) \in \mathcal{L}\big(H^t(S), H^{t-2}(S)\big) \tag{6.7}$$

depend smoothly on $u \in \tilde{U}_s$. For a given surface energy density $\gamma \in C^\infty(S)$ and $u_0 \in \tilde{U}_s$ the Cauchy problem (2.2), (2.3) reappears as

$$\dot{u} = \mathcal{F}(u)\gamma, \quad u(0) = u_0. \tag{6.8}$$

We call a function γ on S an *equilibrium surface energy density* for a given $u \in \tilde{U}_s$ if and only if

$$\mathcal{F}(u)\gamma = 0 \;\text{ on }\; S \quad \text{or equivalently} \quad \mathcal{G}(u)\gamma = \text{const. on } S. \tag{6.9}$$

The latter condition yields a nonlocal first-order elliptic equation for the determination of an equilibrium surface energy density γ. By straightforward perturbation arguments and expansion into spherical harmonics in case of $u = w_0$, the next lemma ensures the existence of a solution of this equation, uniquely determined up to a scaling factor and a linear combination of m functions.

LEMMA 6.3. *Assume $\delta_0 > 0$ sufficiently small. Then for any given $u \in \tilde{U}_s$, $s \geq s_0$ there exists a uniquely determined positive function $\gamma(u) \in H^{s-1}(S)$ such that*

$$\mathcal{G}(u)\gamma(u) = -1 \ \text{ on } \ S, \qquad \int_S N(u)\gamma(u)\,dS = 0.$$

Proof. By elliptic regularity it suffices to consider the case $s = s_0$. If $a \in \mathbb{R}^m$, $u \in \tilde{U}_s$, and $\gamma \in H^{s-1}(S)$ such that

$$\mathcal{G}(u)\gamma = -1 + a \cdot u \ \text{ on } \ S,$$

then (6.2) implies

$$\int_S \omega(u)N(u)(a \cdot u)\,dS = 0,$$

and further $a_1 = \cdots = a_m = 0$ by (6.1). Hence it suffices to show the invertibility of the operator

$$[(\gamma, a) \mapsto L(u)(\gamma, a)] \in \mathcal{L}\big(H^{s-1}(S) \times \mathbb{R}^m, H^{s-2}(S) \times \mathbb{R}^m\big)$$

given by

$$L(u)(\gamma, a) := \big(\mathcal{G}(u)\gamma - a \cdot u, c\big), \quad c := \int_S N(u)\gamma\,dS.$$

As $L(u)$ depends smoothly on $u \in \tilde{U}_s$ it remains to show the existence of

$$L(w_0)^{-1} \in \mathcal{L}\big(H^{s-2}(S) \times \mathbb{R}^m, H^{s-1}(S) \times \mathbb{R}^m\big). \tag{6.10}$$

In this case we have

$$\mathcal{G}(w_0)\gamma = -(m-1)\gamma - A_S\Delta_S\gamma,$$

where A_S and Δ_S denote the Neumann–Dirichlet operator and the Laplace–Beltrami operator on the unit sphere S, respectively. Hence, if we expand

$$\gamma = \sum_{l=0}^{\infty} \gamma_l, \quad \gamma_1 = b \cdot x,$$

where γ_l is a spherical harmonic of degree l and $b \in \mathbb{R}^m$, it follows from

$$\Delta_S\gamma_l = -l(l+m-2)\gamma_l, \qquad A_S\gamma_l = l^{-1}\gamma_l \quad (l > 0)$$

that

$$L(w_0)(\gamma, a) = \Big(-(m-1)\gamma_0 - a \cdot x + \sum_{l=2}^{\infty}(l-1)\gamma_l, \frac{|S|}{m}b\Big).$$

This immediately gives (6.10). Clearly, the equilibrium surface energy density belonging to w_0 is the constant function

$$\overline{\gamma} := \gamma(w_0) = L(w_0)^{-1}(-1,0) = -1/\kappa_0 \ \text{ on } \ S$$

with the curvature $\kappa_0 = -(m-1)$ on S. The proof is complete. ∎

In the following considerations, to derive a priori estimates independent of the existence interval $[0,T]$, let

$$u \in C^0\big([0,T],\tilde{U}_{s_0+4}\big) \cap C^1\big([0,T],H^{s_0+1}(S,\mathbb{R}^m)\big)$$

be any solution of (6.8). Thereby, without explicit mentioning, we always assume that

$$\|u_0\|_{s_0+4} = \|u(0)\|_{s_0+4} \le M$$

and γ is taken from some set of the form (4.38) with fixed positive constants γ^*, M and with a sufficiently large $s_1 \in \mathbb{N}$, such that, in view of Remark 4.7, the constants in the estimates of section 4 are independent of γ for $s \le s_0+4$. Further, we define (cf. (4.12))

$$f(t) := \tilde{\mathcal{G}}\big(u(t)\big)\gamma = \mathcal{G}\big(u(t)\big)\gamma - \frac{1}{|S|}\int_S \mathcal{G}\big(u(t)\big)\gamma\, dS. \tag{6.11}$$

Note that $F\big(u(t)\big)v \equiv 0$ for $v \equiv \text{const}$ implies

$$\dot{u}(t) = \mathcal{F}\big(u(t)\big)\gamma = F\big(u(t)\big)f(t). \tag{6.12}$$

Lemmas 6.4 and 6.5 show in which sense the evolution of u can be controlled by $f(t)$, $t \in [0,T]$ with constants independent of the existence time T. As the main step, we find from the improved semiboundedness estimate (4.31) that $\|u(t)\|$ has only slow exponential growth; more precisely, we have the following lemma.

LEMMA 6.4. *Let $\varepsilon > 0$, $a > 0$, $c > 0$ be given. There are constants $C = C(a,c,\varepsilon)$, $\delta = \delta(\varepsilon) > 0$ such that each of the assumptions*

(i) $\|f(t)\|_{s_0-2} \le \delta$, $t \in [0,T]$, *or*

(ii) $\|f(t)\|_{s_0-2} \le ce^{-at}$, $t \in [0,T]$,

implies

$$\|u(t)\|_{s_0+4} \le Ce^{\varepsilon t} \ \text{ for all } \ t \in [0,T]. \tag{6.13}$$

The constants C and δ may be chosen independently of u and T.

Proof. In view of Theorem 2.2 it is sufficient to prove (6.13) for any sufficiently regular solution $u = u(t)$ of (6.8). In particular, the mapping $t \mapsto g(t) := \big(u(t),u(t)\big)_{s,u(t)}$ with $s := s_0+4$ may be assumed to be differentiable. From Proposition 4.6, estimate (4.31) we get for any given $\varepsilon > 0$,

$$\big(u(t),\mathcal{F}(u(t))\gamma\big)_{s,u(t)} \le \Big(C\|\mathcal{F}(u(t))\gamma\|_{s_0-3} + \frac{\varepsilon}{2}\Big)\, g(t) + C(\varepsilon)$$

and, using $D(u)\{w\}v^2$ as abbreviation for the derivative of the mapping $u \mapsto (v,v)_{s,u}$,

$$\big|D(u(t))\{\mathcal{F}(u(t))\gamma\}u(t)^2\big| \le C\|\mathcal{F}(u(t))\gamma\|_{s_0-3}\|u(t)\|_s^2.$$

Consequently, by differentiating g, we have

$$\begin{aligned} g'(t) &= 2\big(u(t),\dot u(t)\big)_{s,u(t)} + D\big(u(t)\big)\big\{\dot u(t)\big\}u(t)^2 \\ &\le \Big(C\|\mathcal{F}(u(t))\gamma\|_{s_0-3} + \frac{\varepsilon}{2}\Big)\, g(t) + C(\varepsilon) \le \Big(C_1\|f(t)\|_{s_0-2} + \frac{\varepsilon}{2}\Big)\, g(t) + C(\varepsilon). \end{aligned}$$

Defining

$$\alpha(t) := \frac{\varepsilon}{2}t + C_1\int_0^t \|f(s)\|_{s_0-2}\,ds$$

and noting that under the assumptions (i) or (ii) we have

$$0 \le \alpha(t) \le \varepsilon t + C(a,c)$$

and we get from Gronwall's inequality that

$$\begin{aligned} \|u(t)\|_s^2 \le Cg(t) &\le Ce^{\alpha t}\left(g(0) + C(\varepsilon)\int_0^t e^{-\alpha(s)}\,ds\right) \\ &\le C(\varepsilon)e^{\varepsilon t}(g(0)+t) \le C(\varepsilon)e^{2\varepsilon t}(g(0)+1) \le C(\varepsilon)(M^2+1)e^{2\varepsilon t}. \end{aligned}$$

This implies the assertion. ∎

LEMMA 6.5. *Let $\varepsilon > 0$ and $a > 0$ be given. Then there exists $\delta > 0$ such that*

$$\|u(0) - w_0\|_{s_0} \le \delta, \quad \|f(t)\|_{s_0+1} \le \delta e^{-at} \ \text{ for all } \ t \in [0,T] \tag{6.14}$$

imply

$$\|u(t) - w_0\|_{s_0} \le \varepsilon \ \text{ for all } \ t \in [0,T].$$

The constant δ may be chosen independently of T and u.

Proof. We assume according to Lemma 6.4 that

$$\|u(t)\|_{s_0+1} \le Ce^{at/2} \ \text{ for } \ t \in [0,T].$$

Define $g(t) := \|u(t) - w_0\|_{s_0}^2$. Then

$$\begin{aligned} g'(t) = 2\big(\mathcal{F}(u(t))\gamma, u(t) - w_0\big)_{s_0} &\le C\|\mathcal{F}(u(t))\gamma\|_{s_0} \\ &= C\|F(u(t))f(t)\|_{s_0}, \end{aligned}$$

and, consequently,

$$g'(t) \le C\|u(t)\|_{s_0+1}\|f(t)\|_{s_0+1} \le C'\delta e^{at/2}e^{-at}.$$

Hence, for δ sufficiently small,

$$g'(t) \le \tfrac{1}{4}a\varepsilon^2 e^{-at/2} \quad \text{and} \quad g(0) \le \tfrac{1}{2}\varepsilon^2.$$

This implies

$$g(t) \le \tfrac{1}{2}\varepsilon^2(2 - e^{-at/2}) \le \varepsilon^2,$$

which is the assertion. ∎

Now, to obtain estimates of $f(t)$ we derive the evolution equation satisfied by f. Differentiation of (6.11) with respect to t gives

$$\dot f(t) = \mathcal{G}\big(u(t)\big)\big\{\dot u(t)\big\}\gamma - \frac{1}{|S|}\int_S \mathcal{G}\big(u(t)\big)\big\{\dot u(t)\big\}\gamma\,dS,$$

hence inserting (6.12) we obtain

$$\dot f(t) = \mathcal{H}\big(u(t),\gamma\big)f(t), \tag{6.15}$$

where the operator $\mathcal{H}$ is given by

$$\mathcal{H}(u,\gamma)w := \mathcal{G}'(u)\{F(u)w\}\gamma - \frac{1}{|S|}\int_S \mathcal{G}'(u)\{F(u)w\}\gamma\,dS. \tag{6.16}$$

The operator $\mathcal{H}$ is negative semibounded in L^2 in the following sense.

LEMMA 6.6. *For $\|u-w_0\|_{s_0}$ and $\|\gamma-\overline{\gamma}\|_{s_0}$ sufficiently small we have with a positive constant c independent of u and γ:*

$$\big(\mathcal{H}(u,\gamma)w,w\big)_0 \le -c\|w\|_{3/2}^2 \tag{6.17}$$

for all $w\in C^\infty(S)$ with

$$\int_S w\,dS = 0, \quad \int_S \omega(u)N_i(u)w\,dS = 0.$$

Proof. Instead of (6.17) we prove the estimate in the form

$$\big(\mathcal{H}(u,\gamma)w,w\big)_0 \le -c_1\|w\|_{3/2}^2 + c_2R(u,w) \tag{6.18}$$

for *all* $w\in C^\infty(S)$ with some constants $c_1,c_2>0$, where

$$R(u,w) := \left(\int_S w\,dS\right)^2 + \sum_{i=1}^m\left(\int_S \omega(u)N_i(u)w\,dS\right)^2.$$

Further, by perturbation arguments using

$$\|\mathcal{H}(u,\gamma)w - \mathcal{H}(w_0,\bar\gamma)w\|_{-3/2} \le C\|w\|_{3/2}\big(\|u-w_0\|_{s_0} + \|\gamma-\overline{\gamma}\|_{s_0}\big),$$

it suffices to show (6.18) for $u=w_0$, $\gamma=\bar\gamma$. For $\gamma=\bar\gamma$ we have

$$\mathcal{G}(u)\overline{\gamma} = \overline{\gamma}H(u),$$

therefore the linearization of the mean curvature at a sphere yields

$$\mathcal{G}'(w_0)\{v\}\overline{\gamma} = \overline{\gamma}\big((m-1)(x\cdot v) + \Delta_S(x\cdot v)\big).$$

As $[w\mapsto x\cdot F(w_0)w] = B_S$ is the Dirichlet–Neumann operator on the unit sphere S, this implies (6.18). ∎

LEMMA 6.7. *There exists $a>0$ with the property that for any $\varepsilon>0$ there exists $\delta>0$ such that $\|u(0)-w_0\|_{s_0}\le\delta$ and $\|\gamma-\bar\gamma\|_{s_0}\le\delta$ imply*

$$\|f(t)\|_{s_0+1} \le \varepsilon e^{-at} \quad \textit{for all } t\in[0,T].$$

δ may be chosen independently of u and $T>0$.

Proof. First, note that by definition of $f(t)$ and Lemma 6.1 (i) we have

$$\int_S f(t)\,dS = 0, \quad \int_S \omega\big(u(t)\big)N_i\big(u(t)\big)f(t)\,dS = 0.$$

Consequently (6.15) and Lemma 6.6 imply

$$\frac{d}{dt}\big(\|f(t)\|_0^2\big) = 2\big(\mathcal{H}(u(t),\gamma)f(t), f(t)\big)_0 \le -c\|f(t)\|_0^2,$$

with some $c > 0$, hence

$$\|f(t)\|_0 \le e^{-ct}\|f(0)\|_0.$$

Further, as

$$\|f(t)\|_{s_0-2} \le C\big(\|u(t) - w_0\|_{s_0} + \|\gamma - \bar\gamma\|_{s_0}\big) \le C(\delta_0 + \delta),$$

we get from Lemma 6.4 (i) by assuming δ and the constant δ_0 in the definition (6.5) of $\tilde U_s$ sufficiently small,

$$\|u(t)\|_{s_0+4} \le Ce^{\mu t}, \quad \mu := c/(2(s_0+1))$$

and, moreover, using the estimate (2.14),

$$\|f(t)\|_{s_0+2} = \|\mathcal{G}(u(t))\gamma\|_{s_0+2} \le C\|u(t)\|_{s_0+4} \le C'e^{\mu t}.$$

Now we have by interpolation

$$\|f(t)\|_{s_0+1} \le C\big(e^{\mu t}\big)^{\frac{s_0+1}{s_0+2}}\big(e^{-ct}\|f(0)\|_0\big)^{\frac{1}{s_0+2}} = C\|f(0)\|_0^{\frac{1}{s_0+2}}e^{-at}$$

with $a = c/(2(s_0+2))$. This implies the assertion. ∎

Now we are in position to formulate our main result about the long-time existence and convergence to an equilibrium configuration for $t \to \infty$.

THEOREM 6.8. *Let $M > 0$ be given. Then there exists an $\varepsilon > 0$ such that for $\gamma \in C^\infty(S)$ with $\|\gamma\|_{s_1} \le M$, $\|\gamma - \bar\gamma\|_{s_0} \le \varepsilon$ and for any initial value $u_0 \in H^{s_0+4}$ with $\|u_0\|_{s_0+4} \le M$, $\|u_0 - w_0\|_{s_0} \le \varepsilon$ the solution of the Cauchy problem (6.8) exists for all $t > 0$. Moreover, $u(t)$ converges exponentially to some $u^\star = u^\star(u_0,\gamma)$ in H^s, $s < s_0 + 4$ for $t \to \infty$, i.e.,*

$$\|u(t) - u^\star\|_s \le Ce^{-at} \tag{6.19}$$

with suitable $C, a > 0$ (depending on $s < s_0 + 4$). Finally, we have $\mathcal{G}(u^\star)\gamma = \text{const}$ on S, i.e., γ is an equilibrium surface energy density for $u^\star$.

Proof. First, choose $\delta \in (0, \delta_0)$ and $T > 0$ such that, by our local existence theorems, initial values $u_0 \in H^{s_0+4}$ with $\|u_0 - w_0\|_{s_0} \le \delta$ guarantees the (unique) solvability of (6.8) on the time interval $[0,T]$ with $u(t) \in \tilde U_{s_0+4}$, $t \in [0,T]$. Then, for $\varepsilon > 0$ sufficiently small, Lemmas 6.5 and 6.7 ensure $\|u(T) - w_0\|_{s_0} \le \delta$, hence the solution can be continued to the interval $[T, 2T]$ with $u(t) \in \tilde U_{s_0+4}$, $t \in [0, 2T]$. Applying now Lemma 6.5 and 6.7 to the time interval $[0, 2T]$ (note the independence of the constants in these lemmas of the time interval length), we obtain $\|u(2T) - w_0\|_{s_0} \le$

δ_0 again and the solution can be continued to the interval $[2T, 3T]$. Repeating these arguments yields global existence of the solution. Moreover, Lemma 6.7 implies

$$\|f(t)\|_{s_0+1} \le Ce^{-at} \text{ for all } t \ge 0$$

with $a > 0$. Consequently, for any $\varepsilon > 0$ there exists a constant $C(\varepsilon)$ such that

$$\|u(t)\|_{s_0+4} \le C(\varepsilon)e^{\varepsilon t} \text{ for all } t \ge 0 \tag{6.20}$$

by Lemma 6.4 (ii). Further, for $0 \le t \le \theta < \infty$,

$$\|u(t) - u(\theta)\|_0 \le \int_t^\theta \|F(u(\tau))f(\tau)\|_0 \, d\tau \le C \int_t^\theta \|f(\tau)\|_{s_0+1} \, d\tau \le Ce^{-at},$$

and by interpolation using (6.20) with ε sufficiently small,

$$\|u(t) - u(\theta)\|_s \le C_s e^{-a_s t} \text{ for } s < s_0 + 4$$

with suitable constants $C_s, a_s > 0$. This implies convergence of $u(t)$ to some $u^\star$ in $H^s(S)$, $s < s_0 + 4$ as $t \to \infty$ and the estimate (6.19). The final statement follows from letting $t \to \infty$ in (6.11). □

Remark 6.9. It is not hard to see the following regularity property of $u^\star$: if the initial value u_0 belongs additionally to H^s with some $s > s_0 + 4$, then $u^\star(u_0, \gamma) \in H^{s'}$ for $s' < s$ (recall that we have always assumed $\gamma \in C^\infty$). The question whether or not $u^\star$ belongs to H^s remains open and requires more sophisticated estimates.

To illustrate possible equilibrium shapes $\Gamma_{u^\star}$ according to Theorem 6.8 we have performed several numerical test calculations for $m = 2, 3$. In a 2D situation, starting from a circle $S = \Gamma_{w_0}$ and a surface energy density of form

$$\gamma(x) = 1 + 0.8\cos(6\varphi), \quad x = (\cos\varphi, \sin\varphi) \in S, \ 0 \le \varphi < 2\pi,$$

we obtain an equilibrium shape $\Gamma_{u^\star}$ as pictured by the solid line in Figure 1. In contrast, if the correction term $G(u)$, which ensures the gradient flow structure of the evolution problem, is dropped in the definition (2.4), then the resulting shape $\Gamma_{u^\star}$ is given by the dotted line in Figure 1. Clearly, in the latter situation the center of gravity remains fixed due to the symmetries of the chosen initial values and every equilibrium configuration (u, γ) is characterized by $\gamma\kappa = \text{const}$ on Γ_u, hence Γ_u must be convex. (This is similar to a Hele–Shaw evolution where the values of γ are transported only in normal direction to the moving boundary, as this also leads to a dropping of the term $G(u)$.) As Figure 1 shows, this convexity is not true for the full problem. The second example concerns an axisymmetric situation in 3D. Here the evolution starts from the unit sphere $S = \Gamma_{w_0}$ with the surface energy density

$$\gamma(x) = 1.0 + 0.8x_1(4.0x_1^2 - 3.0), \quad x = (x_1, x_2, x_3) \in S$$

and results in a equilibrium shape as shown in Figure 2. To indicate the length scale we have added grid lines with distance 0.25 in each direction.

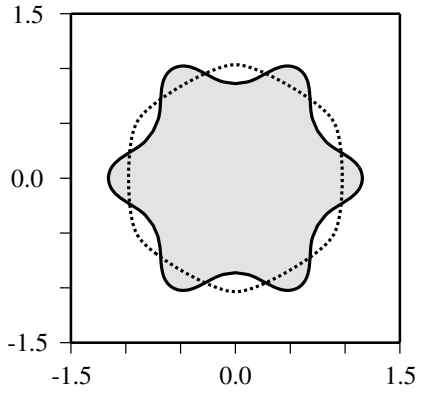


FIG. 1. 2*D examples.*

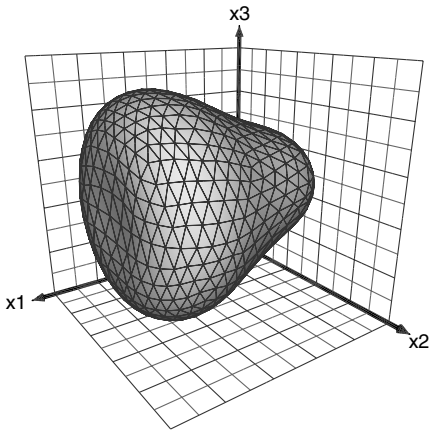


FIG. 2. 3*D example.*

REFERENCES

[1] R. ALMGREN, *Singularity formation in Hele–Shaw bubbles*, Phys. Fluids, 8 (1996), pp. 344–352.
[2] K. BEYER AND M. GÜNTHER, *On the Cauchy problem for a capillary drop,* I: *Irrotational motion*, Math. Meth. Appl. Sci., 21 (1998), pp. 1149–1183.
[3] B. A. BEZUGLYĬ AND N. A. IVANOVA, *Gas bubbles in a Hele–Shaw cell manipulated by a light beam*, Tech. Phys. Lett., 28 (2002), pp. 828–829.
[4] P. CONSTANTIN AND M. PUGH, *Global solutions for small data to the Hele–Shaw problem*, Nonlinearity, 6 (1993), pp. 393–415.
[5] W. CRAIG AND D. P. NICHOLLS, *Travelling two and three dimensional capillary gravity water waves*, SIAM J. Math. Anal., 32 (2000), pp. 323–359.
[6] K. DECKELNICK AND C. ELLIOTT, *Local and global existence results for anisotropic Hele–Shaw flows*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 265–294.
[7] L. GIACOMELLI AND F. OTTO, *Variational formulation for the lubrication approximation of the Hele–Shaw flow*, Calc. Var. Partial Differential Equations, 3 (2001), pp. 377–403.
[8] J. ESCHER AND G. SIMONETT, *A center manifold analysis for the Mullins-Sekerka model*, J. Differential Equations, 143 (1998), pp. 267–292.
[9] M. GÜNTHER AND G. PROKERT, *On Stokes flow with variable and degenerate surface tension coefficient*, NoDEA Nonlinear Differential Equations Appl., 12 (2005), pp. 21–60.
[10] M. GÜNTHER AND G. PROKERT, *On a Hele–Shaw type domain evolution with convected surface energy density*, SIAM J. Math. Anal., 37 (2005), pp. 372–410.
[11] T. KATO AND C. Y. LAI, *Nonlinear evolution equations and the Euler flow*, J. Funct. Anal., 56 (1984), pp. 15–28.
[12] S. R. MARUVADA, C. W. PARK, AND D. Y. YOON, *The influence of surfactant on the bubble motion in Hele–Shaw cells*, Phys. Fluids, 6 (1994), pp. 3267–3275.
[13] G. PROKERT *Existence results for Hele–Shaw flow driven by surface tension*, European J. Appl. Math., 9 (1998), pp. 195–221.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1186–1213

LOW MACH NUMBER FLOWS AND COMBUSTION*

THOMAS ALAZARD†

Abstract. We prove uniform existence results for the full Navier–Stokes equations for time intervals which are independent of the Mach number, the Reynolds number, and the Péclet number. We consider general equations of state and give an application for the low Mach number limit combustion problem introduced by Majda in [*Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, NewYork, 1984].

Key words. low Mach limit, uniform stability, combustion equation

AMS subject classifications. 35B25, 35B40, 35B65

DOI. 10.1137/050644100

1. Introduction.

For a fluid with density ϱ, velocity v, pressure P, temperature T, internal energy e, Lamé coefficients ζ, η, and coefficient of thermal conductivity k, the full Navier–Stokes equations, written in a nondimensional way, are

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho v) = 0, \\ \partial_t(\rho v) + \operatorname{div}(\rho v \otimes v) + \dfrac{\nabla P}{\varepsilon^2} = \mu\big(2\operatorname{div}(\zeta D v) + \nabla(\eta \operatorname{div} v)\big), \\ \partial_t(\rho e) + \operatorname{div}(\rho v e) + P \operatorname{div} v = \kappa \operatorname{div}(k \nabla T) + Q, \end{cases} \tag{1.1}$$

where $\varepsilon \in (0,1]$, $(\mu,\kappa) \in [0,1]^2$, and Q is a given source term (see [11, 16, 18]). In order to be closed, the system is supplemented with a thermodynamic closure law, so that ρ, P, e, T are completely determined by only two of these variables. Also, it is assumed that ζ, η, and k are smooth functions of the temperature.

This paper is devoted to the asymptotic limit where the Mach number ε tends to 0. We are interested in proving results independent of the Reynolds number $1/\mu$ and the Péclet number $1/\kappa$. Our main result asserts that the classical solutions of (1.1) exist and are uniformly bounded on a time interval independent of ε, μ, and κ.

This is a continuation of our previous work [1], where the study was restricted to perfect gases and small source terms Q of size $O(\varepsilon)$. We refer to the introduction of [1] for references and a short historical survey of the background of these problems (see also the survey papers of Danchin [9], Desjardins and Lin [10], Gallagher [13], Schochet [24], and Villani [26]).

The case of perfect gases is interesting in its own: first, perfect gases are widely studied in the physical literature; and second, it contains the important analysis of the singular terms. Yet, modeling real gases requires general equations of state (see [4, 19]). Moreover, we shall see that it is interesting to consider large source terms Q for it allows us to answer a question addressed by Majda in [18] concerning the combustion equations.

*Received by the editors November 2, 2005; accepted for publication (in revised form) May 22, 2006; published electronically December 15, 2006.
http://www.siam.org/journals/sima/38-4/64410.html
†MAB, Université de Bordeaux I, 33405 Talence Cedex, France (thomas.alazard@math.u-bordeaux1.fr).

1.1. The equations. To be more precise, we begin by rewriting the equations under the form $L(u,\partial_t,\partial_x)u+\varepsilon^{-1}S(u,\partial_x)u=0$, which is the classical framework of a singular limit problem.

Before we proceed, three observations are in order. First, for the low Mach number limit problem, the point is not so much to use the conservative form of the equations, but instead to balance the acoustics components. This is one reason it is interesting to work with the unknowns P, v, T (see [18]). Second, the general case must allow for large density and temperature variations as well as very large acceleration of order of the inverse of the Mach number (see section 5 in [16]). Since $\partial_t v$ is of order $\varepsilon^{-2}\nabla P$, this suggests that we seek P under the form $P=\text{const}+O(\varepsilon)$. As in [20], since P and T are positive functions, it is pleasant to set

$$P=\underline{P}e^{\varepsilon p},\quad T=\underline{T}e^{\theta}, \tag{1.2}$$

where $\underline{P}$ and $\underline{T}$ are given positive constants, say the reference states at spatial infinity. Finally, the details of the following computations are given in the appendix.

From now on, the unknown is (p,v,θ) with values in $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$. We are interested in the general case where p and θ are uniformly bounded in ε (so that $\nabla T=O(1)$ and $\partial_t v=O(\varepsilon^{-1})$).

By assuming that ρ and e are given smooth functions of (P,T), it is found that, for smooth solutions of (1.1), (P,v,T) satisfies a system of the form

$$\begin{cases}\alpha(\partial_t P+v\cdot\nabla P)+\operatorname{div}v=\kappa\beta\operatorname{div}(k\nabla T)+\beta Q,\\ \rho(\partial_t v+v\cdot\nabla v)+\dfrac{\nabla P}{\varepsilon^2}=\mu\big(2\operatorname{div}(\zeta Dv)+\nabla(\eta\operatorname{div}v)\big),\\ \gamma(\partial_t T+v\cdot\nabla T)+\operatorname{div}v=\kappa\delta\operatorname{div}(k\nabla T)+\delta Q,\end{cases} \tag{1.3}$$

where the coefficients α, β, γ, and δ are smooth functions of (P,T). Then, by writing $\partial_{t,x}P=\varepsilon P\partial_{t,x}p$, $\partial_{t,x}T=T\partial_{t,x}\theta$ and redefining the functions k, ζ, and η, it is found that (p,v,θ) satisfies a system of the form

$$\begin{cases}g_1(\phi)(\partial_t p+v\cdot\nabla p)+\dfrac{1}{\varepsilon}\operatorname{div}v=\dfrac{\kappa}{\varepsilon}\chi_1(\phi)\operatorname{div}(k(\theta)\nabla\theta)+\dfrac{1}{\varepsilon}\chi_1(\phi)Q,\\ g_2(\phi)(\partial_t v+v\cdot\nabla v)+\dfrac{1}{\varepsilon}\nabla p=\mu B_2(\phi,\partial_x)v,\\ g_3(\phi)(\partial_t\theta+v\cdot\nabla\theta)+\operatorname{div}v=\kappa\chi_3(\phi)\operatorname{div}(k(\theta)\nabla\theta)+\chi_3(\phi)Q,\end{cases} \tag{1.4}$$

where $\phi:=(\theta,\varepsilon p)$ and $B_2(\phi,\partial_x)=\chi_2(\phi)\operatorname{div}(\zeta(\theta)D\cdot)+\chi_2(\phi)\nabla(\eta(\theta)\operatorname{div}\cdot)$.

We are now in position to explain the main differences between ideal gases and general gases. First, we note that the source term Q introduces an arbitrary unsigned large term of order $1/\varepsilon$ in the equations. Second, to emphasize the role of the thermodynamics, we suppose now that $Q=0$ and mention that, for perfect gases, the coefficient $\chi_1(\phi)$ is a function of εp alone (see Proposition A.8). Hence, for perfect gases, the limit constraint is linear in the sense that it reads $\operatorname{div}v_e=0$ with $v_e=v-\kappa\chi_1(0)k(\theta)\nabla\theta$. By contrast, for general equations of state, the limit constraint is nonlinear.

1.2. Assumptions. To avoid confusion, we denote by $(\vartheta,\wp)\in\mathbb{R}^2$ the place holder of the unknown $(\theta,\varepsilon p)$. Hereafter, it is assumed that:

(H1) The functions ζ, η, and k are C^∞ functions of $\vartheta\in\mathbb{R}$, satisfying $k>0$, $\zeta>0$, and $\eta+2\zeta>0$.

(H2) The functions g_i and χ_i $(i=1,2,3)$ are C^∞ positive functions of $(\vartheta,\wp)\in\mathbb{R}^2$. Moreover,

$$\chi_1<\chi_3,$$

and there exist F, G such that $(\vartheta,\wp)\mapsto(F(\vartheta,\wp),\wp)$ and $(\vartheta,\wp)\mapsto(\vartheta,G(\vartheta,\wp))$ are C^∞ diffeomorphisms from $\mathbb{R}^2$ onto $\mathbb{R}^2$, $F(0,0)=G(0,0)=0$, and

$$g_1\frac{\partial F}{\partial\vartheta}=-g_3\frac{\partial F}{\partial\wp}>0,\qquad g_1\chi_3\frac{\partial G}{\partial\vartheta}=-g_3\chi_1\frac{\partial G}{\partial\wp}<0.$$

Remark 1.1. Assumption (H2) is used to prove various energy estimates. The main hypothesis is the inequality $\chi_1<\chi_3$. In the appendix, it is proved that the inequality $\chi_1<\chi_3$ holds whenever the density ρ and the energy e are C^∞ functions of $(P,T)\in(0,+\infty)^2$, such that $\rho>0$ and

$$P\frac{\partial\rho}{\partial P}+T\frac{\partial\rho}{\partial T}=\rho^2\frac{\partial e}{\partial P},\quad \frac{\partial\rho}{\partial P}>0,\quad \frac{\partial\rho}{\partial T}<0,\quad \frac{\partial e}{\partial T}\frac{\partial\rho}{\partial P}>\frac{\partial e}{\partial P}\frac{\partial\rho}{\partial T}. \tag{1.5}$$

1.3. Main result. We are interested in the case without smallness assumption and consider general initial data, general equations of state, and large source terms Q. To get around the above mentioned nonlinear features of the penalization operator, we establish a few new qualitative properties. These properties are enclosed in various uniform stability results, which assert that the classical solutions of (1.4) exist and are uniformly bounded for a time independent of ε, μ, and κ. We concentrate below on the whole space problem or the periodic case and work in the Sobolev spaces H^σ endowed with the norms $\|u\|_{H^\sigma}:=\|(I-\Delta)^{\sigma/2}u\|_{L^2}$.

The following result is the core of all our other uniform stability results. On the technical side, it contains the idea that one can prove uniform estimates without uniform control of the L^2_x norm of the velocity v.

THEOREM 1.2. *Let $d=1$ or $d\geqslant 3$ and $\mathbb{N}\ni s>1+d/2$. For all source term $Q=Q(t,x)\in C^\infty_0(\mathbb{R}\times\mathbb{R}^d)$ and all $M_0>0$, there exist $T>0$ and $M>0$ such that for all $(\varepsilon,\mu,\kappa)\in(0,1]\times[0,1]\times[0,1]$ and all initial data $(p_0,v_0,\theta_0)\in H^{s+1}(\mathbb{R}^d)$ satisfying*

$$\|(\nabla p_0,\nabla v_0)\|_{H^{s-1}}+\|(\theta_0,\varepsilon p_0,\varepsilon v_0)\|_{H^{s+1}}\leqslant M_0, \tag{1.6}$$

the Cauchy problem for (1.4) *has a unique solution (p,v,θ) in $C^0([0,T];H^{s+1}(\mathbb{R}^d))$ such that*

$$\sup_{t\in[0,T]}\|(\nabla p(t),\nabla v(t))\|_{H^{s-1}}+\|(\theta(t),\varepsilon p(t),\varepsilon v(t))\|_{H^s}\leqslant M. \tag{1.7}$$

A refined statement is proved in section 3.

A notable corollary of Theorem 1.2 is Theorem 4.1, which is the requested result for application to the low Mach number limit. Detailed discussions of the periodic case and the combustion equations are included in sections 5 and 6. The assumption $d\neq 2$ is explained in Remark 2.6.

2. Preliminaries. In order not to interrupt the proofs later on, we collect here some estimates. The main result of this section is Proposition 2.4, which complements the Friedrichs'-type estimate

$$\|\nabla v\|_{H^s}\leqslant\|\operatorname{div}v\|_{H^s}+\|\operatorname{curl}v\|_{H^s}, \tag{2.1}$$

which is immediate using Fourier transform. We prove a variant where div v is replaced by $\operatorname{div}(\rho v)$ where ρ is a positive weight.

Notation. The symbol $\lesssim$ stands for $\leqslant$ up to a positive, multiplicative constant, which depends only on parameters that are considered fixed.

2.1. Nonlinear estimates. Throughout the paper, we will make intensive and often implicit uses of the following estimates.

For all $\sigma \geqslant 0$, there exists K such that, for all $u, v \in L^\infty \cap H^\sigma(\mathbb{R}^d)$,

$$\|uv\|_{H^\sigma} \leqslant K \|u\|_{L^\infty} \|v\|_{H^\sigma} + K \|u\|_{H^\sigma} \|v\|_{L^\infty}. \tag{2.2}$$

For all $s > d/2$, $\sigma_1 \geqslant 0$, $\sigma_2 \geqslant 0$ such that $\sigma_1 + \sigma_2 \leqslant 2s$, there exists a constant K such that, for all $u \in H^{s-\sigma_1}(\mathbb{R}^d)$ and $v \in H^{s-\sigma_2}(\mathbb{R}^d)$,

$$\|uv\|_{H^{s-\sigma_1-\sigma_2}} \leqslant K \|u\|_{H^{s-\sigma_1}} \|v\|_{H^{s-\sigma_2}}. \tag{2.3}$$

For all $s > d/2$ and for all C^∞ function F vanishing at the origin, there exists a smooth function C_F such that, for all $u \in H^s(\mathbb{R}^d)$,

$$\|F(u)\|_{H^s} \leqslant C_F(\|u\|_{L^\infty}) \|u\|_{H^s}. \tag{2.4}$$

2.2. Estimates in $\mathbb{R}^3$. Consider the Fourier multiplier $\nabla\Delta^{-1}$ with symbol $-i\xi/|\xi|^2$. This operator is, at least formally, a right inverse for the divergence operator. The only thing we will use below is that $\nabla\Delta^{-1}u$ is well defined whenever $u = u_1 u_2$ with $u_1, u_2 \in L^\infty \cap H^\sigma(\mathbb{R}^d)$ for some $\sigma \geqslant 0$.

PROPOSITION 2.1. *Given $d \geqslant 3$ and $\sigma \in \mathbb{R}$, the Fourier multiplier $\nabla\Delta^{-1}$ is well defined on $L^1(\mathbb{R}^d) \cap H^\sigma(\mathbb{R}^d)$ with values in $H^{\sigma+1}(\mathbb{R}^d)$. Moreover, there exists a constant K such that, for all $u \in L^1(\mathbb{R}^d) \cap H^\sigma(\mathbb{R}^d)$,*

$$\left\|\nabla\Delta^{-1}u\right\|_{H^{\sigma+1}} \leqslant K \|u\|_{L^1} + K \|u\|_{H^\sigma}. \tag{2.5}$$

Proof. Set $\langle\xi\rangle := (1 + |\xi|^2)^{1/2}$. It is enough to check that the L^2-norm of $(\langle\xi\rangle^{\sigma+1}/|\xi|)\,|\widehat{u}(\xi)|$ is estimated by the right-hand side of (2.5). To do that we write

$$\int_{|\xi|\leqslant 1} \frac{\langle\xi\rangle^{2\sigma+2}}{|\xi|^2} |\widehat{u}(\xi)|^2 \, d\xi \lesssim \|u\|_{L^1}^2, \quad \int_{|\xi|\geqslant 1} \frac{\langle\xi\rangle^{2\sigma+2}}{|\xi|^2} |\widehat{u}(\xi)|^2 \, d\xi \lesssim \|u\|_{H^\sigma}^2,$$

where we used $1/|\xi|^2 \in L^1(\{|\xi| \leqslant 1\})$ for all $d \geqslant 3$. ∎

The next proposition is well known. Its corollary is a special case of a general estimate established in [5].

PROPOSITION 2.2. *Given $d \geqslant 3$ and $s > d/2$, there exists a constant K such that, for all $u \in H^s(\mathbb{R}^d)$,*

$$\|u\|_{L^\infty} \leqslant K \|\nabla u\|_{H^{s-1}}. \tag{2.6}$$

Proof. Since $H^s(\mathbb{R}^d) \hookrightarrow L^\infty(\mathbb{R}^d)$, it suffices to prove the result for u in the Schwartz class $\mathcal{S}(\mathbb{R}^d)$. Now, starting from the Fourier inversion theorem, the Cauchy–Schwarz inequality yields the desired estimate:

$$\|u\|_{L^\infty} \leqslant \left(\int \frac{d\xi}{|\xi|^2 \langle\xi\rangle^{2(s-1)}}\right)^{1/2} \left(\int \langle\xi\rangle^{2(s-1)} |\xi\widehat{u}(\xi)|^2 \, d\xi\right)^{1/2} \lesssim \|\nabla u\|_{H^{s-1}}. \qquad ∎$$

COROLLARY 2.3. *Given $d \geqslant 3$ and $\mathbb{N} \ni s > d/2$, there exists a constant K such that, for all $u_1, u_2 \in H^s(\mathbb{R}^d)$,*

$$\|u_1 u_2\|_{H^s} \leqslant K \|\nabla u_1\|_{H^{s-1}} \|u_2\|_{H^s}. \tag{2.7}$$

Proof. One has to estimate the L^2-norm of $\partial_x^\alpha(u_1 u_2)$, where $\alpha \in \mathbb{N}^d$ satisfies $|\alpha| \leqslant s$. Rewrite this term as $u_1 \partial_x^\alpha u_2 + [\partial_x^\alpha, u_1]u_2$. Since the commutator is a sum of terms of the form $\partial_x^\beta u_1 \partial_x^\gamma u_2$ with $\beta > 0$, the product rule (2.3) implies that

$$\|[\partial_x^\alpha, u_1]u_2\|_{L^2} \lesssim \|\nabla u_1\|_{H^{s-1}} \|u_2\|_{H^s}. \tag{2.8}$$

Moving to the estimate of the first term, we write

$$\|u_1 \partial_x^\alpha u_2\|_{L^2} \leqslant \|u_1\|_{L^\infty} \|u_2\|_{H^s} \lesssim \|\nabla u_1\|_{H^{s-1}} \|u_2\|_{H^s}. \qquad \square$$

2.3. Friedrichs' lemma. With these preliminaries established, we are prepared to prove the following.

PROPOSITION 2.4. *Let $d \geqslant 3$ and $\mathbb{N} \ni s > d/2$. There exists a function $\mathcal{C}$ such that, for all $\varphi \in H^{s+1}(\mathbb{R}^d)$ and all vector field $v \in H^{s+1}(\mathbb{R}^d)$,*

$$\|\nabla v\|_{H^s} \leqslant C \|\mathrm{div}(e^\varphi v)\|_{H^s} + C \|\mathrm{curl}\, v\|_{H^s}, \tag{2.9}$$

where $C := (1 + \|\varphi\|_{H^{s+1}})\mathcal{C}(\|\varphi\|_{H^s}, \|\nabla\varphi\|_{L^\infty})$.

Proof. For this proof, we use the notation

$$R = \|\mathrm{div}(e^\varphi v)\|_{H^s} + \|\mathrm{curl}\, v\|_{H^s},$$

and we denote by C_φ various constants depending only on $\|\varphi\|_{H^s} + \|\nabla\varphi\|_{L^\infty}$.

All the computations given below are meaningful since it is sufficient to prove (2.9) for C^∞ functions with compact supports. We begin by setting

$$\widetilde{v} = v + \nabla\Delta^{-1}\big(\nabla\varphi \cdot v\big).$$

The reason to introduce $\widetilde{v}$ is that

$$e^\varphi \,\mathrm{div}\,\widetilde{v} = \mathrm{div}(e^\varphi v), \quad \mathrm{curl}\,\widetilde{v} = \mathrm{curl}\, v.$$

Hence, by using (2.1), we have

$$\|\nabla\widetilde{v}\|_{H^s} \leqslant \big\|e^{-\varphi}\,\mathrm{div}(e^\varphi v)\big\|_{H^s} + \|\mathrm{curl}\, v\|_{H^s} \leqslant C_\varphi R. \tag{2.10}$$

The proof of (2.9) thus reduces to estimating $v_1 := v - \widetilde{v}$, which satisfies

$$\mathrm{div}(e^\varphi v_1) = -e^\varphi \nabla\varphi \cdot \widetilde{v}, \quad \mathrm{curl}\, v_1 = 0.$$

Again, to estimate v_1 we introduce $\widetilde{v}_1 := v_1 + \nabla\Delta^{-1}\big(\nabla\varphi \cdot v_1\big)$, which solves

$$\mathrm{div}\,\widetilde{v}_1 = -\nabla\varphi \cdot \widetilde{v}, \quad \mathrm{curl}\,\widetilde{v}_1 = 0.$$

The estimate (2.1) implies that $\|\nabla\widetilde{v}_1\|_{H^s} \leqslant \|\nabla\varphi \cdot \widetilde{v}\|_{H^s}$. By using (2.10) and the product rule (2.7), applied with $u_1 = \widetilde{v}$ and $u_2 = \nabla\varphi$, we find that

$$\|\nabla\widetilde{v}_1\|_{H^s} \leqslant \|\nabla\varphi \cdot \widetilde{v}\|_{H^s} \lesssim \|\varphi\|_{H^{s+1}} \|\nabla\widetilde{v}\|_{H^{s-1}} \leqslant \|\varphi\|_{H^{s+1}} C_\varphi R. \tag{2.11}$$

Hence, it remains only to estimate $v_2 = v_1 - \widetilde{v}_1$, which satisfies

$$\operatorname{div}(e^{\varphi} v_2) = -e^{\varphi} \nabla\varphi \cdot \widetilde{v_1} \quad \text{and} \quad \operatorname{curl} v_2 = 0.$$

To estimate v_2 the key point is the estimate

$$\|\nabla\varphi \cdot \widetilde{v}_1\|_{L^{d^\star}} \leqslant C_\varphi R \quad \text{with } d^\star = 2d/(d+2). \tag{2.12}$$

Let us assume (2.12) for a moment and continue the proof.

The constraint $\operatorname{curl} v_2 = 0$ implies that $v_2 = \nabla\Psi$, for some Ψ satisfying

$$\operatorname{div}(e^{\varphi}\nabla\Psi) = -e^{\varphi}\nabla\varphi \cdot \widetilde{v_1}.$$

This allows us to estimate $\nabla\Psi$ by a duality argument. We denote by $\langle\, \cdot, \cdot \,\rangle$ the scalar product in L^2 and write

$$\langle\, e^{\varphi}\nabla\Psi\,, \nabla\Psi \,\rangle = \langle\, e^{\varphi}\nabla\varphi \cdot \widetilde{v}_1\,, \Psi\,\rangle.$$

Denote by $\overline{d}$ the conjugate exponent of $d^\star$, $\overline{d} = d^\star/(d^\star - 1) = 2d/(d-2)$. The Hölder's inequality yields

$$\langle\, e^{\varphi}\nabla\Psi\,, \nabla\Psi\,\rangle \leqslant \|e^{\varphi}\nabla\varphi \cdot \widetilde{v}_1\|_{L^{d^\star}} \, \|\Psi\|_{L^{\overline{d}}}.$$

The first factor is estimated by means of the claim (2.12). In view of the Sobolev's inequality $\|\Psi\|_{L^{\overline{d}}} \lesssim \|\nabla\Psi\|_{L^2}$, we obtain

$$\langle\, e^{\varphi}\nabla\Psi\,, \nabla\Psi\,\rangle \leqslant C_\varphi R \, \|\nabla\Psi\|_{L^2}.$$

By using the elementary estimate $\|\nabla\Psi\|_{L^2}^2 \leqslant \|e^{-\varphi}\|_{L^\infty} \langle\, e^{\varphi}\nabla\Psi\,, \nabla\Psi\,\rangle$, we get

$$\|v_2\|_{L^2} = \|\nabla\Psi\|_{L^2} \leqslant C_\varphi R. \tag{2.13}$$

The end of the proof is straightforward. We write

$$\Delta\Psi = e^{-\varphi}\operatorname{div}(e^{\varphi}\nabla\Psi) - \nabla\varphi\cdot\nabla\Psi = -\nabla\varphi\cdot\widetilde{v}_1 - \nabla\varphi\cdot\nabla\Psi,$$

to obtain, for all $\sigma \in [0, s-1]$,

$$\|\nabla\Psi\|_{H^{\sigma+1}} \lesssim \|\nabla\Psi\|_{L^2} + \|\Delta\Psi\|_{H^\sigma} \lesssim \|\nabla\varphi\cdot\widetilde{v}_1\|_{H^\sigma} + (1 + \|\varphi\|_{H^s})\,\|\nabla\Psi\|_{H^\sigma}.$$

To estimate the first term on the right-hand side, we verify that the analysis establishing (2.7) also yields

$$\|\nabla\varphi\cdot\widetilde{v}_1\|_{H^{s-1}} \lesssim \|\varphi\|_{H^s}\,\|\nabla\widetilde{v}_1\|_{H^{s-1}} \leqslant C_\varphi R,$$

hence, by induction on σ,

$$\|\nabla\Psi\|_{H^s} \leqslant C_\varphi R + C_\varphi\,\|\nabla\Psi\|_{L^2}.$$

Exactly as above, one has

$$\begin{aligned}
&\|\nabla\Psi\|_{H^{s+1}} \lesssim \|\nabla\Psi\|_{L^2} + \|\Delta\Psi\|_{H^s} \lesssim \|\nabla\varphi\cdot\widetilde{v}_1\|_{H^s} + (1 + \|\varphi\|_{H^{s+1}})\,\|\nabla\Psi\|_{H^s}\\
&\|\nabla\varphi\cdot\widetilde{v}_1\|_{H^s} \lesssim \|\varphi\|_{H^{s+1}}\,\|\nabla\widetilde{v}_1\|_{H^{s-1}} \leqslant \|\varphi\|_{H^{s+1}}\,C_\varphi R.
\end{aligned}$$

As a consequence, we end up with

$$\|\nabla\Psi\|_{H^{s+1}} \leqslant \|\varphi\|_{H^{s+1}} (C_\varphi R + C_\varphi \|\nabla\Psi\|_{L^2}).$$

Therefore, the L^2 estimate (2.13) implies that

$$\|v_2\|_{H^{s+1}} = \|\nabla\Psi\|_{H^{s+1}} \leqslant \|\varphi\|_{H^{s+1}} C_\varphi R.$$

By combining this estimate with (2.11), we find that

$$\|\nabla v_1\|_{H^s} \leqslant \|\varphi\|_{H^{s+1}} C_\varphi R.$$

From the definition of v_1 and (2.10), we obtain the desired bound (2.9).

We now have to establish the claim (2.12).

With $\overline{d} = 2d/(d-2)$ as above, the Sobolev's inequality and (2.10) imply that

$$\|\widetilde{v}\|_{L^{\overline{d}}} \lesssim \|\nabla\widetilde{v}\|_{L^2} \leqslant C_\varphi R.$$

On the other hand, the Hölder's inequality yields

$$\|\nabla\varphi\cdot\widetilde{v}\|_{L^\delta} \lesssim \|\nabla\varphi\|_{L^2} \|\widetilde{v}\|_{L^{\overline{d}}} \quad \text{with } \delta = \frac{2\overline{d}}{2+\overline{d}} = \frac{d}{d-1}\cdot$$

By interpolating this estimate with $\|\nabla\varphi\cdot\widetilde{v}\|_{L^{\overline{d}}} \lesssim \|\nabla\varphi\|_{L^\infty} \|\widetilde{v}\|_{L^{\overline{d}}}$, we obtain

$$\forall p \in [\delta, \overline{d}], \quad \|\nabla\varphi\cdot\widetilde{v}\|_{L^p} \lesssim \|\nabla\varphi\|_{L^2\cap L^\infty} \|\widetilde{v}\|_{L^{\overline{d}}} \leqslant C_\varphi R.$$

Because $\operatorname{curl} v_1 = 0$, one can write $v_1 = \nabla\Psi_1$ for some function Ψ_1 satisfying $\Delta\Psi_1 = -\nabla\varphi\cdot\widetilde{v}$. Hence, the Calderon–Zygmund inequality and the previous estimate imply that

$$\|\nabla v_1\|_{L^\delta} = \big\|\nabla^2\Psi_1\big\|_{L^\delta} \lesssim \|\Delta\Psi_1\|_{L^\delta} \leqslant C_\varphi R.$$

Therefore, the Sobolev's inequality yields

$$\|v_1\|_{L^D} \leqslant C_\varphi R \quad \text{with } D = \frac{\delta d}{d-\delta} = \frac{d}{d-2},$$

hence, exactly as above, the Hölder's inequality gives

$$\forall p \in [\underline{d}, \overline{d}], \quad \|e^\varphi\nabla\varphi\cdot\widetilde{v}_1\|_{L^p} \leqslant C_\varphi R \quad \text{with } \underline{d} = \frac{2D}{2+D} = \frac{2d}{3d-4}\cdot \tag{2.14}$$

The key estimate (2.12) is now a consequence of the previous one. Indeed, the estimate (2.14) applies with $p = d^\star = 2d/(d+2)$ since

$$\forall d \geqslant 3, \qquad \underline{d} = \frac{2d}{3d-4} \leqslant \frac{2d}{d+2} \leqslant \frac{2d}{d-2} = \overline{d}.$$

This completes the proof of (2.9). □

For later references, we will need the following version of (2.9).

COROLLARY 2.5. *Let $d = 1$ or $d \geqslant 3$ and $\mathbb{N} \ni s > d/2$. There exists a function $\mathcal{C}$ such that, for all $\varphi \in H^{s+1}(\mathbb{R}^d)$ and all vector field $v \in H^{s+1}(\mathbb{R}^d)$,*

$$\|\nabla v\|_{H^s} \leqslant \mathcal{C}(\|\varphi\|_{H^{s+1}})\big(\|\operatorname{div} v\|_{H^s} + \|\operatorname{curl}(e^\varphi v)\|_{H^s}\big). \tag{2.15}$$

Proof. The case $d = 1$ is obvious. If $d \geqslant 3$, Proposition 2.4 (applied with (φ, v) replaced with $(-\varphi, e^{\varphi}v)$) yields

$$\|\nabla(e^{\varphi}v)\|_{H^s} \leqslant \mathcal{C}(\|\varphi\|_{H^{s+1}})\big(\|\operatorname{div} v\|_{H^s} + \|\operatorname{curl}(e^{\varphi}v)\|_{H^s}\big).$$

Hence, to prove (2.15) we need only prove that

$$\|\nabla v\|_{H^s} \leqslant \mathcal{C}(\|\varphi\|_{H^{s+1}})\,\|\nabla(e^{\varphi}v)\|_{H^s}\,. \tag{2.16}$$

To do that we write $\partial_i v = e^{-\varphi}\partial_i(e^{\varphi}v) - (e^{-\varphi}\partial_i\varphi)(e^{\varphi}v)$. The usual product rule (2.3) implies that the H^s norm of the first term is estimated by the right-hand side of (2.16). Moving to the second term, we use the product rule (2.7) to obtain $\|(e^{-\varphi}\partial_i\varphi)(e^{\varphi}v)\|_{H^s} \lesssim (1 + \|\varphi\|_{H^s})\,\|\partial_i\varphi\|_{H^s}\,\|\nabla(e^{\varphi}v)\|_{H^{s-1}}$. This proves the desired bound (2.16). ∎

Remark 2.6. The fact that Theorem 1.2 precludes the case $d = 2$ is a consequence of the fact that we do not know if (2.15) holds for $d = 2$.

3. Uniform stability. In this section, we prove Theorem 1.2. We follow closely the approach given in [1] and recall the scheme of the analysis and indicate the points at which the argument must be adapted.

Hereafter, we use the notations

$$a := (\varepsilon, \mu, \kappa) \in A := (0,1] \times [0,1] \times [0,1], \qquad \nu := \sqrt{\mu + \kappa},$$
$$\|u\|_{H^{\sigma+1}_{\alpha}} := \|u\|_{H^{\sigma}} + \alpha\,\|u\|_{H^{\sigma+1}} \qquad (\alpha \geqslant 0,\ \sigma \in \mathbb{R}).$$

Step 1: A refined statement. We first give our main result a refined form where the solutions satisfy the same estimates as the initial data. Also, to prove estimates independent of μ and κ, an important point is to seek the solutions in spaces which take into account an extra damping effect for the penalized terms.

Definition 3.1. *Let $T > 0$, $a = (\varepsilon, \mu, \kappa) \in [0,1]^3$ and set $\nu = \sqrt{\mu + \kappa}$. The space $\mathcal{X}^s_a(T)$ consists of these $(p, v, \theta) \in C^0([0,T]; H^s(\mathbb{R}^d))$ such that*

$$\nu(p, v, \theta) \in C^0([0,T]; H^{s+1}(\mathbb{R}^d)), \quad (\mu v, \kappa\theta) \in L^2(0,T; H^{s+2}(\mathbb{R}^d)).$$

The space $\mathcal{X}^s_a(T)$ is given the norm

$$\begin{aligned}\|(p, v, \theta)\|_{\mathcal{X}^s_a(T)} := {} & \|(\nabla p, \nabla v)\|_{L^\infty_T(H^{s-1})} + \|(\theta, \varepsilon p, \varepsilon v)\|_{L^\infty_T(H^{s+1}_\nu)} \\ & + \sqrt{\mu}\,\|\nabla v\|_{L^2_T(H^{s+1}_{\varepsilon\nu})} + \sqrt{\kappa}\,\|\nabla\theta\|_{L^2_T(H^{s+1}_\nu)} \\ & + \sqrt{\mu+\kappa}\,\|\nabla p\|_{L^2_T(H^s)} + \sqrt{\kappa}\,\|\operatorname{div} v\|_{L^2_T(H^s)}\,,\end{aligned}$$

with $\|\cdot\|_{L^p_T(X)}$ denoting the norm in $L^p(0,T;X)$.

For the study of nonlinear problems, it is important to relax the assumption that $Q \in C^\infty_0$. We will consider source terms Q in the following spaces.

Definition 3.2. *The space F^s consists of the function Q such that, for all $\mathbb{N} \ni m \leqslant s$, $\partial^m_t Q \in C^0_b(\mathbb{R}; H^{s+1-2m}(\mathbb{R}^d))$, where C^0_b stands for $C^0 \cap L^\infty$.*

Given a normed space X, we set $B(X; M) = \{x \in X : \|x\| \leqslant M\}$.

Theorem 3.3. *Assume that $d = 1$ or $d \geqslant 3$ and let $\mathbb{N} \ni s > 1 + d/2$. Given $M_0 > 0$ and $Q \in F^s$, there exist $T > 0$ and $M > 0$ such that, for all $a = (\varepsilon, \mu, \kappa) \in A$ and all initial data $(p_0, v_0, \theta_0) \in H^{s+1}(\mathbb{R}^d)$ satisfying*

$$\|(\nabla p_0, \nabla v_0)\|_{H^{s-1}} + \|(\theta_0, \varepsilon p_0, \varepsilon v_0)\|_{H^{s+1}} \leqslant M_0, \tag{3.1}$$

the Cauchy problem for (1.4) *has a unique solution* $(p,v,\theta)\in B(\mathcal{X}^s_a(T);M)$.

This theorem implies Theorem 1.2.

Remark 3.4. (i) The hybrid norm $\|\cdot\|_{H^{s+1}_{\varepsilon\nu}}$ was already used by Danchin in [8].

(ii) A close inspection of the proof indicates that Theorem 3.3 remains valid with (3.1) replaced by

$$\|(p_0,v_0,\theta_0)\|_{\mathcal{X}^s_a(0)}:=\|(\nabla p_0,\nabla v_0)\|_{H^{s-1}}+\|(\theta_0,\varepsilon p_0,\varepsilon v_0)\|_{H^{s+1}_\nu}\leqslant M_0.$$

Step 2: Local well posedness. We explain here how to reduce matters to proving uniform bounds. To do so, our first task is to establish the local well posedness of the Cauchy problem for fixed $a=(\varepsilon,\mu,\kappa)\in A$.

LEMMA 3.5. *Let* $d\geqslant 1$, $s>1+d/2$, *and* $a\in A$. *For all initial data* $U_0=(p_0,v_0,\theta_0)\in H^s(\mathbb{R}^d)$, *there exists a positive time* T *such that the Cauchy problem for* (1.4) *has a unique solution* $U=(p,v,\theta)\in C^0([0,T];H^s)$ *such that* $U(0)=U_0$. *Moreover, the interval* $[0,T^\star)$, *with* $T^\star<+\infty$, *is a maximal interval of* H^s *existence if and only if* $\limsup_{t\to T^\star}\|U(t)\|_{W^{1,\infty}(\mathbb{R}^d)}=+\infty$.

Lemma 3.5 is a special case of Proposition 4.5 established in section 4.

As in [1, 20], on account of the previous local existence result for fixed $a\in A$, Theorem 1.2 is a consequence of the following uniform estimates.

PROPOSITION 3.6. *Let* $d=1$ *or* $d\geqslant 3$, $\mathbb{N}\ni s>1+d/2$, *and* $M_0>0$. *Set* $H^\infty(\mathbb{R}^d):=\cap_{\sigma\geqslant 0}H^\sigma(\mathbb{R}^d)$. *There exist a constant* C_0 *and a nonnegative function* $C(\cdot)$ *such that, for all* $T\in(0,1]$ *and all* $a\in A$, *if* $(p,v,\theta)\in C^\infty([0,T];H^\infty(\mathbb{R}^d))$ *is a solution of* (1.4) *with initial data satisfying* (3.1), *then the norm* $\Omega_a(T):=\|U\|_{\mathcal{X}^s_a(T)}$ *satisfies*

$$\Omega_a(T)\leqslant C_0\exp\big((\sqrt{T}+\varepsilon)C(\Omega_a(T))\big). \tag{3.2}$$

To prove Proposition 3.6, as usual, a key step is to study the linearized system. This is the purpose of Theorem 3.10. With this result in hand, to establish the desired nonlinear estimates (3.2), the analysis is divided into four steps. This happens for two reasons. First, on the technical side, most of the work concerns the separation of the estimates into high and low frequency components, where the division occurs at frequencies of order of the inverse of ε (since the second-derivative terms with $O(1)$ coefficients and the first-derivative terms with $O(\varepsilon^{-1})$ coefficients balance there). Second, there is a division into terms whose evolution is estimated directly by eliminating large terms of size $O(\varepsilon^{-1})$ (see Lemmas 3.18 and 3.19), and terms whose size is estimated by means of Theorem 3.10 and the special structure of the equations (see Lemma 3.16).

This scheme of estimates has two useful properties. First, it avoids estimating the L^2 norm of p and v (to obtain a closed set of estimates, we will use the preliminary estimates from section 2). Second, it allows us to overcome the factor $1/\varepsilon$ in front of the source term Q. Indeed, the linear estimate in Theorem 3.10 is applied only to high-frequencies and weighted time derivatives $(\varepsilon\partial_t)^m$. Hence, the fact that the source term is assumed to be neither of high frequency nor have rapid time oscillations allows us to recover the lost factor of ε in the nonlinear estimates. Note that, in the combustion case, the assumptions on the source term Q may be verified directly from the equations. Also, we mention that the L^2 norm of (p,v) will later be estimated in section 4 under an additional hypothesis.

NOTATION 3.7. *From now on, we consider an integer* $s>1+d/2$, *a fixed time* $0<T\leqslant 1$, *a fixed triple of parameters* $a=(\varepsilon,\mu,\kappa)\in A$, *a bound* M_0, *a fixed smooth*

solution $U = (p, v, \theta) \in C^\infty([0,T]; H^\infty(\mathbb{R}^d))$ *of* (1.4) *with initial data satisfying* (3.1), *and we set*

$$\Omega := \|U\|_{\mathcal{X}^s_a(T)}\,.$$

With these notations, Proposition 3.6 *can be formulated concisely as follows: if* $d \neq 2$, *there exist constants* C_0 *depending only on* M_0 *and* C *depending only on* Ω *such that*

$$\Omega \leqslant C_0 e^{(\sqrt{T}+\varepsilon)C}.$$

Hereafter, we use the notations $\phi := (\theta, \varepsilon p)$ *and* $\nu := \sqrt{\mu+\kappa}$.

NOTATION 3.8. *For later application to the nonlinear case when* $Q = F(Y)$ *for some unknown function* Y, *we also give precise estimates in terms of norms of* Q. *For our purposes, the requested norm is the following:*

$$\Sigma := \sum_{0\leqslant m\leqslant s} \big\|(I-(\varepsilon\nu)^2\Delta)^{-m/2}\big(\varepsilon(\partial_t + v\cdot\nabla)\big)^m Q\big\|_{L^\infty(0,T;H^{s+1-m}_\nu)}. \tag{3.3}$$

Remark 3.9. To use nonlinear estimates, it is easier to work in Banach algebras. If $d \geqslant 3$, Proposition 2.2 shows that we can supplement the $\mathcal{X}^s_a$ estimates with L^∞ estimates for the velocity: it suffices to prove (3.2) with $C(\Omega_a(T))$ replaced by $C(\Omega^+_a(T))$, where $\Omega^+_a(T) := \Omega_a(T) + \|v\|_{L^\infty((0,T)\times\mathbb{R}^d)}$. Similarly, if $d \geqslant 3$, all the estimates involving the source term Q remain valid with Σ replaced by

$$\sum_{0\leqslant m\leqslant s} \big\|(I-(\varepsilon\nu)^2\Delta)^{-m/2}(\varepsilon\partial_t)^m Q\big\|_{L^\infty(0,T;H^{s+1-m}_\nu)}.$$

Step 3: An energy estimate for linearized equations. A key step in the analysis is to estimate the solution $(\widetilde{p},\widetilde{v},\widetilde{\theta})$ of linearized equations. As alluded to above, a notable fact is that we can see unsigned large terms $\varepsilon^{-1}f^\varepsilon(t,x)$ in the equations for p and v as source terms provided that: (1) they do not convey fast oscillations in time $\partial_t f^\varepsilon = O(1)$; (2) it does not imply a loss of derivatives. To be more precise, in the nonlinear estimates, we will see the term $\varepsilon^{-1}\chi_1(\phi)Q$ as a source term. Similarly, we can see terms of the form $\varepsilon^{-1}F(\varepsilon p, \theta, \sqrt{\kappa}\nabla\theta)$ as source terms. As a result, it is sufficient to consider the following linearized system:

$$\begin{cases} g_1(\phi)(\partial_t\widetilde{p} + v\cdot\nabla\widetilde{p}) + \dfrac{1}{\varepsilon}\operatorname{div}\widetilde{v} - \dfrac{\kappa}{\varepsilon}\operatorname{div}(k_1(\phi)\nabla\widetilde{\theta}) = F_1, \\ g_2(\phi)(\partial_t\widetilde{v} + v\cdot\nabla\widetilde{v}) + \dfrac{1}{\varepsilon}\nabla\widetilde{p} - \mu B_2(\phi,\partial_x)\widetilde{v} = F_2, \\ g_3(\phi)(\partial_t\widetilde{\theta} + v\cdot\nabla\widetilde{\theta}) + G(\phi,\nabla\phi)\cdot\widetilde{v} + \operatorname{div}\widetilde{v} - \kappa\chi_3(\phi)\operatorname{div}(k(\phi)\nabla\widetilde{\theta}) = F_3, \end{cases} \tag{3.4}$$

where the unknown $(\widetilde{p},\widetilde{v},\widetilde{\theta})$ is a smooth function of $(t,x) \in [0,T]\times\mathbb{R}^d$.

The following result establishes estimates on

$$\begin{aligned} \|(\widetilde{p},\widetilde{v},\widetilde{\theta})\|_{a,T} := {} & \|(\widetilde{p},\widetilde{v})\|_{L^\infty_T(H^1_{\varepsilon\nu})} + \|\widetilde{\theta}\|_{L^\infty_T(H^1_\nu)} \\ & + \sqrt{\kappa}\|\nabla\widetilde{\theta}\|_{L^2_T(H^1_\nu)} + \sqrt{\mu}\,\|\nabla\widetilde{v}\|_{L^2_T(H^1_{\varepsilon\nu})} \\ & + \sqrt{\mu+\kappa}\,\|\nabla\widetilde{p}\|_{L^2_T(L^2)} + \sqrt{\kappa}\,\|\operatorname{div}\widetilde{v}\|_{L^2_T(L^2)}\,, \end{aligned} \tag{3.5}$$

in terms of the norm $\|(\widetilde{p},\widetilde{v},\widetilde{\theta})\|_{a,0} := \|(\widetilde{p},\widetilde{v})(0)\|_{H^1_{\varepsilon\nu}} + \|\widetilde{\theta}(0)\|_{H^1_\nu}$ of the data.

THEOREM 3.10. *Let $d \geqslant 1$ and assume that G, k_1, and k_3 are C^∞ functions such that, for all $(\vartheta, \wp) \in \mathbb{R}^2$, $0 < k_1(\vartheta, \wp) < \chi_3(\vartheta, \wp)k(\vartheta)$. Set*

$$R_0 := \|\phi(0)\|_{H^{s-1}}, \quad R := \sup_{t\in[0,T]} \big\|(\phi, \partial_t\phi + v\cdot\nabla\phi, \nabla\phi, \nu\nabla^2\phi, \nabla v)(t)\big\|_{H^{s-1}}.$$

There exist constants C_0 depending only on R_0 and C depending only on R such that,

$$\|(\widetilde{p}, \widetilde{v}, \widetilde{\theta})\|_{a,T} \leqslant C_0 e^{TC} \|(\widetilde{p}_0, \widetilde{v}_0, \widetilde{\theta}_0)\|_{a,0} + C\int_0^T \|(F_1, F_2)\|_{H^1_{\varepsilon\nu}} + \|F_3\|_{H^1_\nu}\, dt.$$

In [1] we established the previous theorem with R_0 and R replaced by

$$R_0' = \|\phi(0)\|_{L^\infty}, \quad R' = \sup_{t\in[0,T]} \big\|(\phi, \partial_t\phi, v, \nabla\phi, \nu\nabla^2\phi, \nabla v)(t)\big\|_{L^\infty}.$$

To prove the above variant, we need only check two facts. First, in the proof of Theorem 4.3 in [1], the terms $\partial_t\phi$ and v always come together within terms involving the convective derivative $\partial_t\phi + v\cdot\nabla\phi$.

Second, we have to verify that the $L^\infty_{t,x}$ norms of the coefficients $(g_i(\phi), \dots)$ are estimated by constants of the form C_0e^{TC}. In [1] we used the estimate

$$\sup_{t\in[0,T]} \|F(\phi(t))\|_{L^\infty} \leqslant \|F(\phi(0))\|_{L^\infty} + T\sup_{t\in[0,T]} \|\partial_t F(\phi(t))\|_{L^\infty} \leqslant C_0' + TC',$$

for some constants depending only on R_0' and R'. Here, based on an usual estimate for hyperbolic equations, we can prove a similar bound.

LEMMA 3.11. *Let $F \in C^\infty(\mathbb{R}^2)$ be such that $F(0) = 0$. There exist constants C_0 depending only on R_0 and C depending only on R such that, for all $t \in [0,T]$, $\|F(\phi(t))\|_{H^{s-1}} \leqslant C_0e^{TC}$.*

Proof. Since $s - 1 > d/2$, the Moser's estimates (2.2) and (2.4) imply that there exists a function $\mathcal{C}$ depending only on the function F such that

$$\begin{aligned}\|(\partial_t + v\cdot\nabla)F(\phi)\|_{H^{s-1}} &\leqslant \big(1 + \|F'(\phi) - F'(0)\|_{H^{s-1}}\big)\, \|(\partial_t + v\cdot\nabla)\phi(t)\|_{H^{s-1}},\\ &\leqslant \mathcal{C}(\|(\phi, \partial_t\phi + v\cdot\nabla\phi)\|_{H^{s-1}}) \leqslant \mathcal{C}(R),\end{aligned}$$

and $\|F(\phi(0))\|_{H^{s-1}} \leqslant \mathcal{C}(\|\phi(0)\|_{H^{s-1}})$.

Hence, the desired estimate follows from the following estimate: there exists a constant V depending only on $\|\nabla v\|_{L^\infty_T H^{s-1}}$ such that

$$\sup_{t\in[0,T]} \|F(\phi(t))\|_{H^{s-1}} \leqslant \|F(\phi(0))\|_{H^{s-1}} + TV\sup_{t\in[0,T]} \|(\partial_t + v\cdot\nabla)F(\phi(t))\|_{H^{s-1}}.$$

To prove this result we set $\widetilde{u} := \partial_x^\alpha F(\phi)$, where $\alpha \in \mathbb{N}^d$ is such that $|\alpha| \leqslant s-1$. Then $\widetilde{u}$ solves

$$\partial_t u + v\cdot\nabla u = f := \partial_x^\alpha\big((\partial_t + v\cdot\nabla)F(\phi)\big) + [v, \partial_x^\alpha]\cdot\nabla F(\phi).$$

Since $s - 1 > d/2$, the product rule (2.3) implies that

$$\begin{aligned}\|[v, \partial_x^\alpha]\cdot\nabla F(\phi)\|_{L^2} &\lesssim \sum_{\beta+\gamma=\alpha,\ \beta>0} \|\partial_x^\beta v\partial_x^\gamma\nabla F(\phi)\|_{L^2}\\ &\lesssim \sum_{\beta+\gamma=\alpha,\ \beta>0} \|\partial_x^\beta v\|_{H^{s-1-(|\beta|-1)}}\, \|\partial_x^\gamma\nabla F\|_{H^{s-1-(|\gamma|+1)}},\end{aligned}$$

hence, $\|f\|_{L^2} \lesssim \|(\partial_t + v\cdot\nabla)F(\phi)\|_{H^{s-1}} + \|\nabla v\|_{H^{s-1}}\, \|F(\phi)\|_{H^{s-1}}$.

We next use an integration by parts argument yielding

$$\frac{d}{dt}\|\widetilde{u}\|_{L^2}^2 \leqslant (1+\|\operatorname{div} v\|_{L^\infty})\,\|\widetilde{u}\|_{L^2}^2 + \|f\|_{L^2}^2\,.$$

Gronwall's lemma concludes the proof. □

Step 4: High frequency estimates. We begin by estimating the high frequency component

$$\Omega^{\mathrm{HF}} := \|(I-J_{\varepsilon\nu})U\|_{\mathcal{X}_a^s(T)}\,,$$

where $\{J_h \,|\, h\in[0,1]\}$ is a Friedrichs' mollifier: $J_h = \jmath(hD_x)$ is the Fourier multiplier with symbol $\jmath(h\xi)$, where $\jmath$ is a C^∞ function of $\xi\in\mathbb{R}^d$ satisfying

$$0\leqslant \jmath\leqslant 1,\quad \jmath(\xi)=1 \text{ for } |\xi|\leqslant 1,\quad \jmath(\xi)=0 \text{ for } |\xi|\geqslant 2,\quad \jmath(\xi)=\jmath(-\xi)\,.$$

PROPOSITION 3.12. *Let $d\geqslant 1$. There exist constants C_0 depending only on M_0 and C depending only on Ω, such that*

$$\Omega^{\mathrm{HF}} \leqslant C_0 e^{\sqrt{T}C} + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}\,. \tag{3.6}$$

Proof. Introduce $P := (I-J_{\varepsilon\nu})\Lambda^s$ and $\widetilde{U} := (Pp, Pv, P\theta)$. Then, $\widetilde{U}$ satisfies system (3.4) with

$$k_1(\phi) := \chi_1(\phi)k(\theta),\quad G(\phi,\nabla\phi) := g_3(\phi)\nabla\theta,$$

and $F=(F_1,F_2,F_3)^T := f_{\mathrm{HF}}+f_Q+f_\chi$, where

$$f_Q := \begin{pmatrix} \varepsilon^{-1}P\big(\chi_1(\phi)Q\big)\\ 0\\ P\big(\chi_3(\phi)Q\big)\end{pmatrix},\quad f_\chi := \begin{pmatrix} -\kappa\varepsilon^{-1}\nabla\chi_1(\phi)\cdot\big(k(\theta)\nabla\widetilde{\theta}\big)\\ 0\\ 0\end{pmatrix},$$

and f_{HF} is given by

$$\begin{aligned}
f_{1,\mathrm{HF}} &= \big[g_1(\phi),P\big](\partial_t+v\cdot\nabla)p + g_1(\phi)\big[v,P\big]\cdot\nabla p \;-\frac{\kappa}{\varepsilon}\big[B_1(\phi,\partial_x),P\big]\theta,\\
f_{2,\mathrm{HF}} &= \big[g_2(\phi),P\big](\partial_t+v\cdot\nabla)v + g_2(\phi)\big[v,P\big]\cdot\nabla v \;-\mu\big[B_2(\phi,\partial_x),P\big]v,\\
f_{3,\mathrm{HF}} &= \big[g_3(\phi),P\big](\partial_t+v\cdot\nabla)\theta + g_3(\phi)\big\{v;P\big\}\cdot\nabla\theta - \kappa\big[B_3(\phi,\partial_x),P\big]\theta,
\end{aligned}$$

where $B_i(\phi,\partial_x)=\chi_i(\phi)\operatorname{div}(k(\theta)\nabla\cdot)$ $(i=1,3)$, $[A,B]=AB-BA$, and

$$\big\{v;P\big\}\cdot\nabla\theta := v\cdot\nabla P\theta + (Pv)\cdot\nabla\theta - P(v\cdot\nabla\theta).$$

Estimate for f_{HF}. We use the following analogue of Lemma 5.3 in [1]: there exists a constant $K=K(d,s)$ such that

$$\begin{aligned}
\big\|\big[f,P\big]u\big\|_{H^1_{\varepsilon\nu}} &\leqslant \varepsilon\nu K\,\|\nabla f\|_{L^\infty}\,\|u\|_{H^s} + \varepsilon\nu K\,\|\nabla f\|_{H^s}\,\|u\|_{L^\infty},\\
\big\|\big[f,P\big]u\big\|_{H^1_{\nu}} &\lesssim \nu K\,\|\nabla f\|_{L^\infty}\,\|u\|_{H^s} + \nu K\,\|\nabla f\|_{H^s}\,\|u\|_{L^\infty}.
\end{aligned}$$

The fact that the right-hand side only involves ∇f follows from the most simple of all the sharp commutator estimates established in [17]: for all $s>1+d/2$ and all Fourier

multiplier $A(D_x) \in \operatorname{Op} S^s_{1,0}$, there exists a constant K such that, for all $f \in H^s(\mathbb{R}^d)$ and all $u \in H^s(\mathbb{R}^d)$,

$$\|[f, A(D_x)]u\|_{L^2} \leqslant K \|\nabla f\|_{L^\infty} \|u\|_{H^{s-1}} + K \|\nabla f\|_{H^{s-1}} \|u\|_{L^\infty}. \tag{3.7}$$

As in [1], from this and the usual nonlinear estimates (2.2) and (2.4), it can be verified that there exists a generic function $\mathcal{C}$ (depending only on parameters that are considered fixed) such that,

$$\begin{aligned}
\|f_{1,\mathrm{HF}}\|_{H^1_{\varepsilon\nu}} &\leqslant \mathcal{C}\big(\|(\theta, \varepsilon p, \varepsilon v)\|_{H^{s+1}_\nu}\big)\{1 + \|\varepsilon(\partial_t + v\cdot\nabla)p\|_{H^s_\nu} + \kappa \|\theta\|_{H^{s+2}}\},\\
\|f_{2,\mathrm{HF}}\|_{H^1_{\varepsilon\nu}} &\leqslant \mathcal{C}\big(\|(\theta, \varepsilon p, \varepsilon v)\|_{H^{s+1}_\nu}\big)\{1 + \|\varepsilon(\partial_t + v\cdot\nabla)v\|_{H^s_\nu} + \mu \|\varepsilon v\|_{H^{s+2}}\},\\
\|f_{3,\mathrm{HF}}\|_{H^1_\nu} &\leqslant \mathcal{C}\big(\|(\theta, \varepsilon p, \varepsilon v)\|_{H^{s+1}_\nu}\big)\{1 + \|(\partial_t + v\cdot\nabla)\theta\|_{H^s_\nu} + \kappa \|\theta\|_{H^{s+2}}\}.
\end{aligned}$$

Set $\psi = (\theta, \varepsilon p, \varepsilon v)$. The key point is that

$$\begin{aligned}
&\|(\partial_t + v\cdot\nabla)\psi\|_{H^s_\nu}\\
&\quad\leqslant \mathcal{C}(\|\psi\|_{H^{s+1}_\nu})\big\{1 + \big\|(\nu\nabla p, \nu \operatorname{div} v, \varepsilon\mu\nabla^2 v, \kappa\nabla^2\theta)\big\|_{H^s} + \|Q\|_{H^s_\nu}\big\}.
\end{aligned} \tag{3.8}$$

This estimate differs from the one that appears in Lemma 5.14 in [1] in that the right-hand side does not involve v itself but only its derivatives. Yet, as the reader can verify, the same proof applies since we do not estimate $\partial_t\psi$, but instead $\partial_t\psi + v\cdot\nabla\psi$.

Estimate for f_Q and f_χ. By using the elementary estimate

$$\|(I - J_{\varepsilon\nu})u\|_{H^{\sigma+1}_{\varepsilon\nu}} \lesssim \varepsilon\nu \|u\|_{H^{\sigma+1}},$$

we find that

$$\frac{1}{\varepsilon}\|P(\chi_1(\phi)Q)\|_{H^1_{\varepsilon\nu}} + \|P(\chi_3(\phi)Q)\|_{H^1_\nu} \leqslant \|\chi_1(\phi)Q\|_{H^{s+1}_\nu} + \|\chi_3(\phi)Q\|_{H^{s+1}_\nu}.$$

The tame estimates (2.2) and (2.4) (see also Lemmas 5.5 and 5.6 in [1]) imply that

$$\|\chi_i(\phi)Q\|_{H^{s+1}_\nu} \lesssim (1 + \|\chi_i(\phi) - \chi_i(0)\|_{H^{s+1}_\nu})\,\|Q\|_{H^{s+1}_\nu} \lesssim \mathcal{C}(\|\phi\|_{H^{s+1}_\nu})\,\|Q\|_{H^{s+1}_\nu}$$

so that $\|f_{1,Q}\|_{L^\infty_T(H^1_{\varepsilon\nu})} + \|f_{3,Q}\|_{L^\infty_T(H^1_\nu)} \leqslant C\|Q\|_{L^\infty_T(H^{s+1}_\nu)}$. The technique for estimating f_χ is similar; we find that $\|f_{1,\chi}\|_{L^\infty_T(H^1_{\varepsilon\nu})} \leqslant C$.

By definition of $\|\cdot\|_{\mathcal{X}^s_a(T)}$, the previous estimates imply that there exists a constant C depending only on Ω such that

$$\begin{aligned}
\int_0^T \|(F_1, F_2)\|_{H^1_{\varepsilon\nu}} + \|F_3\|_{H^1_\nu}\,dt &\leqslant \sqrt{T}\bigg(\int_0^T \|(F_1, F_2)\|^2_{H^1_{\varepsilon\nu}} + \|F_3\|^2_{H^1_\nu}\,dt\bigg)^{1/2}\\
&\leqslant \sqrt{T}C + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}.
\end{aligned}$$

From here we can parallel the rest of the argument of section 5 in [1], to prove that $\|(Pp, Pv, P\theta)\|_{a,T} \leqslant C_0 \exp(\sqrt{T}C) + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}$, where the norm $\|\cdot\|_{a,T}$ is as defined in (3.5). Since $\Omega^{\mathrm{HF}} \lesssim \|(Pp, Pv, P\theta)\|_{a,T}$, this completes the proof. ∎

Step 5: Low frequency estimates. The following step is to estimate the low frequency part of the fast components:

$$\Omega^{\mathrm{LF}} := \|\operatorname{div} J_{\varepsilon\nu} v\|_{L^\infty_T(H^{s-1})} + \nu \|\operatorname{div} J_{\varepsilon\nu} v\|_{L^2_T(H^s)} + \|\nabla J_{\varepsilon\nu} p\|_{L^\infty_T(H^{s-1})} + \nu \|\nabla J_{\varepsilon\nu} p\|_{L^2_T(H^s)}.$$

PROPOSITION 3.13. *Let $d \geqslant 1$. There exist constants C_0 depending only on M_0, C depending only on Ω, and C' depending only on $\Omega + \Sigma$, such that*

$$\Omega^{\mathrm{LF}} \leqslant C_0 e^{(\sqrt{T}+\varepsilon)C} + \sqrt{T}C'. \tag{3.9}$$

By contrast with the high frequency regime, the estimate (3.9) cannot be obtained from the L^2 estimates by an elementary argument using differentiation of the equations (see [20, 24]). To overcome this problem, we first give estimates for the time derivatives, and next we use the special structure of the equations to estimate the spatial derivatives.

For the case of greatest physical interest ($d = 3$), the proof given in [1] applies with only minor changes. Indeed, as alluded to in Remark 3.9, it suffices to check that all the estimates involving $\|v\|_{H^s}$ remain valid with $\|v\|_{H^s}$ replaced by $\|v\|_{L^\infty} + \|\nabla v\|_{H^{s-1}}$. Yet, if $d \leqslant 2$, because of the lack of L^2 estimates for the velocity, we cannot use the time derivatives. For this problem, we use an idea introduced by Secchi in [25]. Namely, we replace ∂_t by the convective derivative

$$D_v := \partial_t + v \cdot \nabla.$$

For the reader's convenience, we indicate how to adapt the three main calculus inequalities in [1] when ∂_t is replaced by D_v.

First, to localize in the low frequency region we use the following commutator estimate. The thing of interest is the gain of an extra factor ε.

LEMMA 3.14. *Given $s > 1 + d/2$, there exists a constant K such that for all $\varepsilon \in [0,1]$, all $\nu \in [0,2]$, all $T > 0$, all $m \in \mathbb{N}$ such that $1 \leqslant m \leqslant s$, and all f, u, and v in $C^\infty([0,T]; H^\infty(\mathbb{D}))$,*

$$\big\| [f, J_{\varepsilon\nu}(\varepsilon D_v)^m] u \big\|_{H^{s-m+1}_{\varepsilon\nu}} \leqslant K\varepsilon \Big\{ \|f\|_{H^s} + \sum_{\ell=0}^{m-1} \|\Lambda^{-\ell}_{\varepsilon\nu} (\varepsilon D_v)^\ell D_v f\|_{H^{s-1-\ell}} \Big\} \times \Big\{ \|\Lambda^{-m}_{\varepsilon\nu} (\varepsilon D_v)^m u\|_{H^{s-m}_\nu} + \sum_{\ell=0}^{m-1} \|\Lambda^{-\ell}_{\varepsilon\nu} (\varepsilon D_v)^\ell u\|_{H^{s-1-\ell}} \Big\},$$

where $\Lambda^\sigma_{\varepsilon\nu} := (I - (\varepsilon\nu)^2 \Delta)^{\sigma/2}$.

To apply the previous lemma, we need estimates of the coefficients f and $D_v f$. Since, for system (1.4), the coefficients are functions of the slow variable $(\theta, \varepsilon p, \varepsilon v)$, the main estimates are the following.

LEMMA 3.15. *Let $s > 1 + d/2$ be an integer. There exists a function $\mathcal{C}(\cdot)$ such that, for all $a = (\varepsilon, \mu, \kappa) \in A$, all $T > 0$, and all smooth solution $(p, v, \theta) \in C^\infty([0,T]; H^\infty(\mathbb{D}))$ of (1.4), if $\nu \in [(\mu + \kappa)/2, 2]$, then the function Ψ defined by*

$$\Psi := \big(\psi, D_v\psi, \nabla\psi\big) \quad \text{where} \quad \psi := (\theta, \varepsilon p, \varepsilon v)$$

satisfies

$$\sum_{0\leqslant\ell\leqslant s}\|\Lambda_{\varepsilon\nu}^{-\ell}(\varepsilon D_v)^\ell\Psi\|_{H^{s-\ell-1}}\leqslant\mathcal{C}\big(\|\Psi\|_{H^{s-1}}+\Sigma\big),\tag{3.10}$$

$$\sum_{0\leqslant\ell\leqslant s}\|\Lambda_{\varepsilon\nu}^{-\ell}(\varepsilon D_v)^\ell\Psi\|_{H^{s-\ell}_\nu}\leqslant\mathcal{C}\big(\|\Psi\|_{H^{s-1}}+\Sigma\big)\,\|\Psi\|_{H^s_\nu},\tag{3.11}$$

where Σ *is as defined in* (3.3).

Once this is granted, we are in position to estimate the commutator of the equations (1.4) and $\mathcal{P}:=J_{\varepsilon\nu}(\varepsilon D_v)^s$:

$$\begin{aligned}
f_{1,\mathrm{LF}}&=\big[g_1(\phi),\mathcal{P}\big]D_vp+g_1(\phi)\big[v,\mathcal{P}\big]\cdot\nabla p-\frac{\kappa}{\varepsilon}\big[B_1(\phi,\partial_x),\mathcal{P}\big]\theta,\\
f_{2,\mathrm{LF}}&=\big[g_2(\phi),\mathcal{P}\big]D_vv+g_2(\phi)\big[v,\mathcal{P}\big]\cdot\nabla v-\mu\big[B_2(\phi,\partial_x),\mathcal{P}\big]v,\\
f_{3,\mathrm{LF}}&=\big[g_3(\phi),\mathcal{P}\big]D_v\theta+g_1(\phi)\big[v,\mathcal{P}\big]\cdot\nabla\theta-\kappa\big[B_3(\phi,\partial_x),\mathcal{P}\big]\theta.
\end{aligned}$$

It is found that

$$\|f_{1,\mathrm{LF}}\|_{H^1_{\varepsilon\nu}}+\|f_{2,\mathrm{LF}}\|_{H^1_{\varepsilon\nu}}+\|f_{3,\mathrm{LF}}\|_{H^1_\nu}\leqslant(1+\|\Psi\|_{H^s_\nu})\mathcal{C}(\|\Psi\|_{H^{s-1}}+\Sigma).$$

Note that Ψ is estimated by means of (3.8).

As in the high frequency regime, we have to estimate source terms of the form $\varepsilon^{-1}\mathcal{P}F(\Psi,Q)$. The fact that these large source terms cause no difficulty comes from the fact that $\varepsilon^{-1}J_{\varepsilon\nu}(\varepsilon D_v)^sF(\Psi,Q)=J_{\varepsilon\nu}(\varepsilon D_v)^{s-1}D_vF(\Psi,Q)$ together with $D_vF(\Psi,Q)=O(1)$ (the norm Σ introduced in (3.3) is the requested norm to give this statement a precise meaning).

With these results in hand, one can estimate $J_{\varepsilon\nu}(\varepsilon D_v)^s(p,v,\theta)$ by means of Theorem 3.10. Next, we estimate for $\operatorname{div}J_{\varepsilon\nu}v$ and $\nabla J_{\varepsilon\nu}p$ from the estimate of $J_{\varepsilon\nu}(\varepsilon D_v)^s(p,v,\theta)$ by means of the following induction argument.

LEMMA 3.16. *Set* $\|u\|_{\mathcal{K}^\sigma_\nu(T)}:=\|u\|_{L^\infty_T(H^{\sigma-1})}+\nu\,\|u\|_{L^2_T(H^\sigma)}$.

Let $\widetilde{U}:=(\widetilde{p},\widetilde{v},\widetilde{\theta})$ *solve*

$$\begin{cases}
g_1(\phi)(\partial_t\widetilde{p}+v\cdot\nabla\widetilde{p})+\varepsilon^{-1}\operatorname{div}\widetilde{v}-\kappa\varepsilon^{-1}\chi_1(\phi)\operatorname{div}(k(\theta)\nabla\widetilde{\theta})=f_1,\\
g_2(\phi)(\partial_t\widetilde{v}+v\cdot\nabla\widetilde{v})+\varepsilon^{-1}\nabla\widetilde{p}-\mu B_2(\phi,\partial_x)\widetilde{v}=f_2,\\
g_3(\phi)(\partial_t\widetilde{\theta}+v\cdot\nabla\widetilde{\theta})+\operatorname{div}\widetilde{v}-\kappa\chi_3(\phi)\operatorname{div}(k(\theta)\nabla\widetilde{\theta})=f_3.
\end{cases}\tag{3.12}$$

If support of the Fourier transform of $\widetilde{U}$ *is included in the ball* $\{|\xi|\leqslant 2/\varepsilon\nu\}$, *then there exist constant* C_0 *depending only on* M_0 *and* C *depending only on* Ω *such that, for all* $\sigma\in[1,s]$,

$$\begin{aligned}
&\|\nabla\widetilde{p}\|_{\mathcal{K}^\sigma_\nu(T)}+\|\operatorname{div}\widetilde{v}\|_{\mathcal{K}^\sigma_\nu(T)}\\
&\quad\leqslant\widetilde{C}\,\|(\varepsilon D_v)\widetilde{p}\,\|_{\mathcal{K}^\sigma_\nu(T)}+\widetilde{C}\,\|(\varepsilon D_v)\operatorname{div}\widetilde{v}\|_{\mathcal{K}^{\sigma-1}_\nu(T)}\\
&\qquad+\widetilde{C}\,\|\nabla\widetilde{p}\,\|_{L^\infty_T(L^2)}+\widetilde{C}\,\|\widetilde{\theta}(0)\|_{H^{\sigma+1}_\nu}+\varepsilon C\,\|\mu\widetilde{v}\|_{\mathcal{K}^{\sigma+1}_\nu(T)}\\
&\qquad+\varepsilon C\,\|(f_1,f_2)\|_{\mathcal{K}^\sigma_\nu(T)}+\nu\widetilde{C}\,\|f_3\|_{L^2_T(H^\sigma)},
\end{aligned}\tag{3.13}$$

where $\widetilde{C}:=C_0e^{(\sqrt{T}+\varepsilon)C}$.

Step 6: Estimates for the slow components. To complete the proof of (3.2), it remains to estimate $\operatorname{curl} v$ and θ. Yet, this is not straightforward. Following Métivier and Schochet [20], we begin by estimating $\operatorname{curl}(\gamma v)$ for some appropriate positive weight $\gamma = \Gamma(\theta, \varepsilon p)$.

LEMMA 3.17. *Let $d \geqslant 1$. There exist constants C_0 depending only on M_0 and C depending only on Ω, and there exists a function $\Gamma \in C^\infty(\mathbb{R}^2)$ such that, with $\gamma = \Gamma(\theta, \varepsilon p)$, there holds*

$$\|\operatorname{curl}(\gamma v)\|_{L^\infty_T(H^{s-1})} + \sqrt{\mu}\,\|\operatorname{curl}(\gamma v)\|_{L^2_T(H^s)} \leqslant C_0 e^{\sqrt{T}C} + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}.$$

LEMMA 3.18. *Let $d \geqslant 1$. There exist constants C_0 depending only on M_0 and C depending only on Ω, such that*

$$\|J_{\varepsilon\nu}\theta\|_{L^\infty_T(H^{s+1}_\nu)} + \sqrt{\kappa}\,\|J_{\varepsilon\nu}\theta\|_{L^2_T(H^{s+2}_\nu)} \leqslant C_0 e^{\sqrt{T}C} + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}.$$

The proofs of Lemmas 3.17 and 3.18 follow from a close inspection of the proofs of Lemmas 6.25 and 6.26 in [1]. We mention that this is where we use the function F of Assumption (H2) in section 1.2 (γ is related to the fluid entropy).

LEMMA 3.19. *Assume $d \geqslant 3$. There exist constants C_0 depending only on M_0 and C depending only on Ω such that, with $\gamma_0 = \Gamma(\theta_0, \varepsilon p_0)$, where Γ is as above, there holds*

$$\|\operatorname{curl}(\gamma_0 v)\|_{L^\infty_T(H^{s-1})} + \sqrt{\mu}\,\|\operatorname{curl}(\gamma_0 v)\|_{L^2_T(H^s)} \leqslant C_0 e^{\sqrt{T}C} + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}.$$

Proof. Set $\widetilde{\gamma} := \gamma - \gamma_0$. By Lemma 3.17, all we need to prove is that

$$\|\operatorname{curl}(\widetilde{\gamma} v)\|_{L^\infty_T(H^{s-1})} + \sqrt{\mu}\,\|\operatorname{curl}(\widetilde{\gamma} v)\|_{L^2_T(H^s)} \leqslant \sqrt{T}C + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}. \tag{3.14}$$

To do so, we claim that $\widetilde{\gamma}$ is small for small times:

$$\|\widetilde{\gamma}\|_{L^\infty_T(H^s)} + \nu\,\|\widetilde{\gamma}\|_{L^2_T(H^{s+1})} \leqslant \sqrt{T}C + \sqrt{T}C\,\|Q\|_{L^\infty_T(H^{s+1}_\nu)}. \tag{3.15}$$

Let us assume this and continue the proof.

We have to estimate $\operatorname{curl}(\widetilde{\gamma} v) = \widetilde{\gamma}\operatorname{curl} v + (\nabla\widetilde{\gamma}) \times v$. By combining the Cauchy–Schwarz estimate with the usual product rule (2.3) and the product rule (2.7), we find that

$$\begin{aligned}
\|\widetilde{\gamma}\operatorname{curl} v\|_{L^\infty_T(H^{s-1})} &\leqslant \|\widetilde{\gamma}\|_{L^\infty_T(H^{s-1})}\,\|\nabla v\|_{L^\infty_T(H^{s-1})},\\
\sqrt{\mu}\,\|\widetilde{\gamma}\operatorname{curl} v\|_{L^2_T(H^s)} &\leqslant \|\widetilde{\gamma}\|_{L^\infty_T(H^s)}\,\|\sqrt{\mu}\nabla v\|_{L^2_T(H^s)},\\
\|\nabla\widetilde{\gamma} \times v\|_{L^\infty_T(H^{s-1})} &\leqslant \|\widetilde{\gamma}\|_{L^\infty_T(H^s)}\,\|\nabla v\|_{L^\infty_T(H^{s-1})},\\
\sqrt{\mu}\,\|\nabla\widetilde{\gamma} \times v\|_{L^2_T(H^s)} &\leqslant \|\sqrt{\mu}\widetilde{\gamma}\|_{L^2_T(H^{s+1})}\,\|\nabla v\|_{L^\infty_T(H^{s-1})}.
\end{aligned}$$

The claim (3.15) then yields the desired bound (3.14).

We now have to prove the claim (3.15). We first note that

$$\begin{aligned}
\nu\,\|\widetilde{\gamma}\|_{L^2_T(H^{s+1})} &\leqslant \nu\sqrt{T}\,\|\widetilde{\gamma}\|_{L^\infty_T(H^{s+1})}\\
&\leqslant \nu\sqrt{T}\mathcal{C}(\|(\theta, \varepsilon p)\|_{L^\infty_T(L^\infty_x)})(1 + \|(\theta, \varepsilon p)\|_{L^\infty_T(H^{s+1})})\\
&\leqslant \sqrt{T}\mathcal{C}(\|(\theta, \varepsilon p)\|_{L^\infty_T(H^{s+1}_\nu)}) \leqslant \sqrt{T}C.
\end{aligned}$$

To prove the second half of (3.15), we verify that, directly from the definitions, $\widetilde{\gamma}$ satisfies an equation of the form $\partial_t\widetilde{\gamma}+v\cdot\nabla\widetilde{\gamma}=f$ with f bounded in $L^2(0,T;H^s(\mathbb{R}^d))$ by a constant depending only on $\Omega+\|Q\|_{L^\infty_T(H^{s+1}_\nu)}$. Then, we apply the above mentioned estimate for hyperbolic equations:

$$\|\widetilde{\gamma}\|_{L^\infty_T(H^s)}\lesssim e^{TV}\,\|\widetilde{\gamma}(0)\|_{H^s}+\int_0^T e^{(T-t)V}\,\|f\|_{H^s}\,dt, \tag{3.16}$$

where $V=K\int_0^T\|\nabla v\|_{H^{s-1}}\,dt$ with $K=K(s,d)$. Since $\widetilde{\gamma}(0)=0$, by applying the Cauchy–Schwarz inequality, it is found that the $L^\infty_T(H^s)$ norm of $\widetilde{\gamma}$ is estimated by $\sqrt{T}e^{TV}\,\|f\|_{L^2_T(H^s)}$, thereby obtaining the claim. □

Step 7: Closed set of estimates. To complete the proof of Proposition 3.6, it remains to check that we have proved a closed set of estimates.

The obvious estimate $\|u\|_{H^\sigma}\leqslant\|J_{\varepsilon\nu}u\|_{H^\sigma}+\|(I-J_{\varepsilon\nu})u\|_{H^\sigma}$ implies that

$$\|(\nabla p,\operatorname{div}v)\|_{L^\infty_T(H^{s-1})}+\sqrt{\mu+\kappa}\,\|(\nabla p,\operatorname{div}v)\|_{L^2_T(H^s)}\lesssim\Omega_{\mathrm{LF}}+\Omega_{\mathrm{HF}},$$

and, similarly, $\|\theta\|_{L^\infty_T(H^{s+1}_\nu)}+\sqrt{\kappa}\,\|\nabla\theta\|_{L^2_T(H^{s+1}_\nu)}$ is estimated by

$$\|J_{\varepsilon\nu}\theta\|_{L^\infty_T(H^{s+1}_\nu)}+\sqrt{\kappa}\,\|J_{\varepsilon\nu}\nabla\theta\|_{L^2_T(H^{s+1}_\nu)}+\Omega_{\mathrm{HF}}.$$

The estimate $\|\varepsilon u\|_{H^{\sigma+1}_\nu}\lesssim\|\varepsilon u\|_{L^2}+\|\nabla u\|_{H^{\sigma-1}}+\|(I-J_{\varepsilon\nu})u\|_{H^{\sigma+1}_\nu}$ yields

$$\begin{aligned}\|(\varepsilon p,\varepsilon v)\|_{L^\infty_T(H^{s+1}_\nu)}+\sqrt{\mu}\,\|\nabla v\|_{L^2_T(H^{s+1}_{\varepsilon\nu})}&\lesssim\|(\varepsilon p,\varepsilon v)\|_{L^\infty_T(L^2)}+\|(\nabla p,\nabla v)\|_{L^\infty_T(H^{s-1})}\\&\quad+\sqrt{\mu}\,\|\nabla v\|_{L^2_T(H^s)}+\Omega_{\mathrm{HF}}.\end{aligned} \tag{3.17}$$

On the other hand, Corollary 2.5 implies that, if $d\neq2$, there exists a constant C_0 depending only on M_0 such that

$$\begin{aligned}\|\nabla v\|_{L^\infty_T(H^{s-1})}+\sqrt{\mu}\,\|\nabla v\|_{L^2_T(H^s)}&\leqslant C_0\,\|(\operatorname{div}v,\operatorname{curl}(\gamma_0 v))\|_{L^\infty_T(H^{s-1})}\\&\quad+C_0\sqrt{\mu}\,\|(\operatorname{div}v,\operatorname{curl}(\gamma_0 v))\|_{L^2_T(H^s)}.\end{aligned}$$

By using the estimate (3.8), one can verify that the term $\|(\varepsilon p,\varepsilon v)\|_{L^\infty_T(L^2)}$ (in the left-hand side of (3.17)) can be estimated as in the proof of Lemma 3.11. Therefore, according to Propositions 3.12–3.13 and Lemmas 3.18–3.19, we have proved that, if $d\neq2$, then $\Omega\leqslant\widetilde{C}$ where $\widetilde{C}=C_0e^{(\sqrt{T}+\varepsilon)C}+\sqrt{T}C'$ for some constants C_0, C, and C' depending only on M_0, Ω and $\Omega+\Sigma$, respectively.

This concludes the proof of Proposition 3.6 and hence Theorem 3.3.

4. Uniform estimates in the Sobolev spaces. With regards to the low Mach number limit problem, we mention that the convergence results[1] proved in [1] apply to general systems (not only for perfect gases). To avoid repetition, we only mention that one can rigorously justify the low Mach number limit for general initial data provided that one can prove that the solutions are uniformly bounded in Sobolev spaces (see Proposition 8.2 in [1]). The problem presents itself: Theorem 1.2 only

[1] These results are strongly based on a theorem of Métivier and Schochet [20] about the decay to zero of the local energy for a class of wave operators with variable coefficients.

gives uniform estimates for the derivatives of p and v. In this section, we give uniform bounds in Sobolev norms.

THEOREM 4.1. *Let $d \geqslant 1$ and $\mathbb{N} \ni s > 1 + d/2$. Assume that $Q = 0$. Also, assume that either $\chi_1 = \chi_1(\vartheta, \wp)$ is independent of ϑ or that $d \geqslant 3$. Then, for all $M_0 > 0$, there exist $T > 0$ and $M > 0$ such that, for all $a = (\varepsilon, \mu, \kappa) \in A$ and all initial data $(p_0, v_0, \theta_0) \in H^{s+1}(\mathbb{R}^d)$ satisfying*

$$\|(p_0, v_0, \theta_0)\|_{H^{s+1}} \leqslant M_0,$$

the Cauchy problem for (1.4) has a unique solution (p, v, θ) in $C^0([0,T]; H^{s+1}(\mathbb{R}^d))$ such that

$$\sup_{t\in[0,T]} \|(p(t), v(t), \theta(t))\|_{H^s} \leqslant M.$$

The first half of this result is proved in [1]. Indeed, the assumption that $\chi_1(\vartheta, \wp)$ does not depend on ϑ is satisfied by perfect gases. Therefore, we concentrate on the second half ($d \geqslant 3$). In view of Theorem 3.3, it remains only to prove *a posteriori* uniform L^2 estimates. More precisely, the proof of Theorem 4.1 reduces to establishing the following result.

LEMMA 4.2. *Let $d \geqslant 3$. Consider a family of solutions (p^a, v^a, θ^a) of (1.4) (for some source terms Q^a) uniformly bounded in the sense of the conclusion of Theorem 3.3:*

$$\sup_{a\in A} \|(p^a, v^a, \theta^a)\|_{\mathcal{X}_a^s(T)} < +\infty, \tag{4.1}$$

for some $s > 1 + d/2$ and fixed $T > 0$. Assume further that the source terms Q^a are uniformly bounded in $C^1([0,T]; L^1 \cap L^2(\mathbb{R}^d))$ and that the initial data $(p^a(0), v^a(0))$ are uniformly bounded in $L^2(\mathbb{R}^d)$. Then the solutions (p^a, v^a, θ^a) are uniformly bounded in $C^0([0,T]; L^2(\mathbb{R}^d))$.

Remark 4.3. We allow $Q^a \neq 0$ for application to the combustion equations. To clarify matters, we note that one can replace (4.1) by

$$\sup_{a\in A} \sup_{t\in[0,T]} \|(\nabla p^a(t), \nabla v^a(t))\|_{H^s} + \|\theta^a(t)\|_{H^{s+1}} < +\infty,$$

for some $s > 2 + d/2$.

Proof. For this proof, we set

$$R := \sup_{a\in A} \Big\{ \|(p^a, v^a, \theta^a)\|_{\mathcal{X}_a^s(T)} + \|(p^a(0), v^a(0))\|_{L^2} + \|Q^a\|_{C^1([0,T];L^1\cap L^2)} \Big\},$$

and we denote by $C(R)$ various constants depending only on R.

The strategy of the proof consists of transforming the system (1.4) so as to obtain L^2 estimates uniform in ε by a simple integration by parts argument.

To do that we claim that there exist $U^a \in C^1([0,T]; L^2(\mathbb{R}^d))$ satisfying the following properties:

$$\sup_{a\in A} \|(p^a, v^a)\|_{L^\infty_T(L^2)} \leqslant \sup_{a\in A} \|U^a\|_{L^\infty_T(L^2)} + C(R), \tag{4.2}$$

$$\sup_{a\in A} \|U^a(0)\|_{L^2} \leqslant C(R), \tag{4.3}$$

and U^a solves a system having the form

$$E^a\partial_t U^a+\varepsilon^{-1}S(\partial_x)U^a=F^a, \tag{4.4}$$

where $S(\partial_x)$ is skew-symmetric, the symmetric matrices $E^a=E^a(t,x)$ are positive definite, and one has the uniform bounds

$$\sup_{a\in A}\|\partial_t E^a\|_{L^\infty([0,T]\times\mathbb{R}^d)}+\big\|(E^a)^{-1}\big\|^{-1}_{L^\infty([0,T]\times\mathbb{R}^d)}+\|F^a\|_{L^1_T(L^2)}\leqslant C(R). \tag{4.5}$$

Before we prove the claim, let us prove that it implies Lemma 4.2. To see this, we combine two basic ingredients:

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\langle\, E^aU^a\,,U^a\,\rangle&=-\varepsilon^{-1}\langle\, S(\partial_x)U^a\,,U^a\,\rangle+\langle\, F^a\,,U^a\,\rangle+\frac{1}{2}\langle(\partial_tE^a)U^a\,,U^a\,\rangle\\&\leqslant\|F^a\|_{L^2}^2+C(R)\,\|U^a\|_{L^2}^2\,,\end{aligned}$$

and $\big\|U^a\big\|_{L^2}^2\leqslant\big\|(E^a)^{-1}\big\|_{L^\infty}^{-1}\langle\, E^aU^a\,,U^a\,\rangle$. Hence, by (4.3) and (4.5), the Gronwall's lemma implies that $\big\|U^a\big\|_{L^\infty_T(L^2)}\leqslant C(R)$. The estimate (4.2) thus implies the desired result.

To prove the claim, we set $U^a:=(p^a,v^a-V^a)^T$, where

$$V^a:=\kappa\chi_1(\phi^a)k(\theta^a)\nabla\theta^a+\nabla\Delta^{-1}\big(-\kappa\nabla\chi_1(\phi^a)\cdot k(\theta^a)\nabla\theta^a+\chi_1(\phi^a)Q^a\big).$$

The fact that V^a is well defined follows from Proposition 2.1. We next verify that U^a satisfies (4.4) with

$$E^a=\begin{pmatrix}g_1(\phi^a)&0\\0&g_2(\phi^a)\end{pmatrix},\quad S(\partial_x)=\begin{pmatrix}0&\operatorname{div}\\\nabla&0\end{pmatrix},$$

$$F^a=\begin{pmatrix}-g_1(\phi^a)v^a\cdot\nabla p^a\\-g_1(\phi^a)v^a\cdot\nabla v^a+\mu B_2(\phi^a,\partial_x)v^a-g_2(\phi^a)\partial_tV^a\end{pmatrix}.$$

By (2.3), (2.4), and (2.6), to prove that the bounds (4.2) and (4.5) hold, it suffices to prove that $\|\partial_t\phi^a\|_{H^{s-1}}\leqslant C(R)$. Yet, this is nothing new. Indeed, we first observe that, directly from the equations,

$$\|\partial_t\phi^a+v^a\cdot\nabla\phi^a\|_{H^{s-1}}\leqslant C(R).$$

On the other hand, the product rule (2.7) implies that $\|v^a\cdot\nabla\phi^a\|_{H^{s-1}}$ is estimated by $\|\nabla v^a\|_{H^{s-1}}\,\|\phi^a\|_{H^s}\leqslant C(R)$. This completes the proof. □

Remark 4.4. For our purposes, one of the main differences between $\mathbb{R}^3$ and $\mathbb{R}$ is the following. For all $f\in C_0^\infty(\mathbb{R}^3)$, Proposition 2.5 implies that there exists a vector field $u\in H^\infty(\mathbb{R}^3)$ such that $\operatorname{div}u=f$. In sharp contrast, the mean value of the divergence of a smooth vector field $u\in H^\infty(\mathbb{R})$ is zero. This implies that Lemma 4.2 is false with $d=1$.

The following result contains an analysis of the easy case where initially $\theta_0=O(\varepsilon)$. This regime is interesting for the incompressible limit (see [3]).

PROPOSITION 4.5. *Let $d\geqslant1$ and $\mathbb{R}\ni s>1+d/2$. For all $M_0>0$, there exists $T>0$ and $M>0$ such that for all $a\in A$ and all initial data $(p_0,v_0,\theta_0)\in H^s(\mathbb{R}^d)$ satisfying*

$$\|(p_0,v_0)\|_{H^s}+\varepsilon^{-1}\,\|\theta_0\|_{H^s}\leqslant M_0, \tag{4.6}$$

the Cauchy problem for (1.4) *has a unique solution* (p, v, θ) *in* $C^0([0,T]; H^s(\mathbb{R}^d))$ *such that*

$$\sup_{t\in[0,T]} \|(p(t), v(t))\|_{H^s} + \varepsilon^{-1} \|\theta(t)\|_{H^s} \leqslant M. \tag{4.7}$$

Proof. The proof of this result is based on the change of unknown $(p, v, \theta) \mapsto (G(\theta, \varepsilon p), v, \theta)$, where G is as given by Assumption (H2) in section 1.2. By setting $\rho = G(\theta, \varepsilon p)$ it is found that (p, v, θ) satisfies (1.4) if and only if

$$\begin{cases} \chi_3(\partial_t \rho + v \cdot \nabla \rho) + (\chi_3 - \chi_1) \operatorname{div} v = 0, \\ g_2(\partial_t v + v \cdot \nabla v) + \varepsilon^{-2}\gamma_1 \nabla\theta + \varepsilon^{-2}\gamma_2 \nabla\rho - \mu B_2 v = 0, \\ g_3(\partial_t \theta + v \cdot \nabla\theta) + \operatorname{div} v - \kappa\chi_3 \operatorname{div}(k\nabla\theta) = 0, \end{cases} \tag{4.8}$$

where $\gamma_1 = (\chi_1 g_3)/(\chi_3 g_1)$ and $\gamma_2 = 1/g_1$. Notice that Assumption (H2) implies that the coefficients g_i, γ_i, χ_3 and $\chi_3 - \chi_1$ are positive.

The key point is that the assumption (4.6) allows us to symmetrize the equations by setting $u := (\tilde\rho, v, \tilde\theta)$, where

$$\tilde\rho := \varepsilon^{-1}\rho, \quad \tilde\theta := \varepsilon^{-1}\theta.$$

The fact that this change of unknowns is singular in ε causes no difficulty. Indeed, directly from the assumption (4.6), we have $\|\tilde\theta(0)\|_{H^s} \leqslant M_0$. On the other hand, the assumption $G(0,0) = 0$ implies that there is a function C_G such that $\|G(u)\|_{H^\sigma} \leqslant C_G(\|u\|_{L^\infty}) \|u\|_{H^\sigma}$ for all $u \in H^\sigma$ with $\sigma > d/2$. Therefore, we have

$$\begin{aligned} \|\tilde\rho\|_{H^s} = \varepsilon^{-1} \|G(\theta, \varepsilon p)\|_{H^s} &\leqslant \varepsilon^{-1} C_G(\|(\theta, \varepsilon p)\|_{L^\infty}) \|(\theta, \varepsilon p)\|_{H^s} \\ &\leqslant C_G(\|(\theta, \varepsilon p)\|_{L^\infty}) \|(\tilde\theta, p)\|_{H^s}, \end{aligned} \tag{4.9}$$

hence, $\|\tilde\rho(0)\|_{H^s} \leqslant C_0$ for some constant depending only on M_0.

Because $(\vartheta, \wp) \mapsto (\vartheta, G(\vartheta, \wp))$ is a C^∞ diffeomorphism with $G(0,0) = 0$, one can write $\varepsilon p = P\big(\theta, G(\theta, \varepsilon p)\big) = P(\theta, \rho)$, for some C^∞ function P vanishing at the origin. Therefore one can see the coefficients $(g_i, \chi_i, \gamma_i \dots)$ as functions of (θ, ρ). Hence, with $u = (\tilde\rho, v, \tilde\theta)$ as above, one can rewrite system (4.8) under the form

$$A_0(\varepsilon u)\partial_t u + \sum_{1\leqslant j\leqslant d} A_j(u, \varepsilon u)\partial_j u + \frac{1}{\varepsilon} \sum_{1\leqslant j\leqslant d} S_j(\varepsilon u)\partial_j u - B(\varepsilon u, \partial_x)u = 0, \tag{4.10}$$

where the matrices S_j, A_j are symmetric (with A_0 positive definite) and the viscous perturbation $B(\varepsilon u, \partial_x)$ is as in (4.8).

Note that one can always assume that the matrices S_j have constant coefficients. Furthermore, since the matrix A_0 multiplying the time derivative depends only on the unknown through εu, and since the initial data $u(0)$ are uniformly bounded in H^s, the proof of the uniform existence theorem of [15] applies. By that proof, we conclude that the solutions of (4.10) exist and are uniformly bounded for a time T independent of ε. Once this is granted, it remains to verify that the solutions (p, v, θ) of system (1.4) exist and are uniformly bounded in the sense of (4.7). To see this, as for $\tilde\rho$ in (4.9), we note that

$$\begin{aligned} \|p\|_{H^s} &= \|P(\theta, \rho)\|_{H^s} \\ &\leqslant \varepsilon^{-1} C_P(\|(\theta, \rho)\|_{L^\infty}) \|(\theta, \rho)\|_{H^s} = C_P(\|(\theta, \rho)\|_{L^\infty}) \|(\tilde\theta, \tilde\rho)\|_{H^s} \\ &\leqslant C(\|(\tilde\theta, \tilde\rho)\|_{H^s}), \end{aligned}$$

so that $\|(p, v)\|_{H^s} + \varepsilon^{-1} \|\theta\|_{H^s} \leqslant C(\|u\|_{H^s})$. This completes the proof. ∎

Remark 4.6. Consider the Euler equations ($\mu = 0 = \kappa$ and $\varepsilon = 1$). By a standard rescaling, Proposition 4.5 just says that the classical solutions with small initial data of size δ exist for a time of order of $1/\delta$. Following the approach initiated by Alinhac in [2], several much more precise results have been obtained. In particular, the interested reader is referred to the recent advance of Godin [14] (for the 3D non-isentropic Euler equations).

5. Spatially periodic solutions. In this section, we consider the case where x belongs to the torus $\mathbb{T}^d$.

THEOREM 5.1. *Let $d \geqslant 1$ and $\mathbb{N} \ni s > 1 + d/2$. For all source term $Q \in C^\infty(\mathbb{R} \times \mathbb{T}^d)$ and for all $M_0 > 0$, there exist $T > 0$ and $M > 0$ such that, for all $a \in A$ and all initial data $(p_0, v_0, \theta_0) \in H^{s+1}(\mathbb{T}^d)$ satisfying*

$$\|(p_0, v_0)\|_{H^s} + \|(\theta_0, \varepsilon p_0, \varepsilon v_0)\|_{H^{s+1}} \leqslant M_0,$$

the Cauchy problem for (1.4) *has a unique solution (p, v, θ) in $C^0([0,T]; H^{s+1}(\mathbb{T}^d))$ such that*

$$\sup_{t\in[0,T]} \|\nabla p(t)\|_{H^{s-1}} + \|v(t)\|_{H^s} + \|(\theta(t), \varepsilon p(t))\|_{H^s} \leqslant M.$$

The proof follows from two observations: first, the results proved in Steps 1–6 in section 3 apply *mutatis mutandis* in the periodic case; and second, as proved below, the periodic case is easier in that one can prove uniform L^2 estimates for the velocity. This in turn implies that (as in [1, 20]) one can directly prove a closed set of estimates by means of the estimate

$$\|v\|_{H^s(\mathbb{T}^d)} \leqslant C\,\|\operatorname{div} v\|_{H^{s-1}(\mathbb{T}^d)} + C\,\|\operatorname{curl}(\gamma v)\|_{H^{s-1}(\mathbb{T}^d)} + C\,\|v\|_{L^2(\mathbb{T}^d)},$$

for some constant C depending only on $\|\log\gamma\|_{H^s(\mathbb{T}^d)}$ (compare with (2.15)).

Let us concentrate on the main new qualitative property.

LEMMA 5.2. *Let $d \geqslant 1$. Consider a family of solutions (p^a, v^a, θ^a) of* (1.4) *(for some source terms Q^a) such that*

$$\sup_{a\in A} \|(p^a, v^a, \theta^a)\|_{\mathcal{X}^s_a(T)} < +\infty,$$

for some $s > 1+d/2$ and fixed $T > 0$. If Q^a is uniformly bounded in $C^1([0,T]; L^2(\mathbb{T}^d))$ and $(p^a(0), v^a(0))$ is uniformly bounded in $L^2(\mathbb{T}^d)$, then v^a is uniformly bounded in $C^0([0,T]; L^2(\mathbb{T}^d))$.

Proof. The main new technical ingredient is, as used by Schochet in [23], an appropriate ansatz for the pressure.

Again, the proof makes use of the Fourier multiplier $\nabla\Delta^{-1}$. Note that $\nabla\Delta^{-1}$ is bounded from $L^2_\sharp(\mathbb{T}^d)$ to $H^1(\mathbb{T}^d)$, where $L^2_\sharp(\mathbb{T}^d)$ consists of these functions $u \in L^2(\mathbb{T}^d)$ such that $\langle u\rangle := \int_{\mathbb{T}^d} u(x)\,dx = 0$.

Set

$$F^a := \kappa\chi_1(\phi^a)\operatorname{div}(k(\theta^a)\nabla\theta^a) + \chi_1(\phi^a)Q^a,$$

and introduce the functions $V^a = V^a(t,x)$ and $P^a = P^a(t)$ by

$$P^a := \frac{\langle F^a\rangle}{\langle g_1(\phi^a)\rangle} \quad \text{and} \quad V^a := \nabla\Delta^{-1}\big(F^a - g_1(\phi^a)P^a\big),$$

so that

$$F^a = g_1(\phi^a)P^a + \operatorname{div} V^a.$$

This allows us to rewrite the first equation in (1.4) as

$$g_1(\phi^a)(\partial_t p^a + v^a \cdot \nabla p^a) + \varepsilon^{-1} \operatorname{div}(v^a - V^a) = g_1(\phi^a)P^a.$$

Therefore, by introducing

$$U^a := (q^a, v^a - V^a)^T \quad \text{with} \quad q^a(t,x) = p^a(t,x) - P^a(t),$$

we are back in the situation of Lemma 4.2: U^a satisfies

$$E^a(\partial_t U^a + v^a \cdot \nabla U^a) + \varepsilon^{-1} S(\partial_x) U^a = F^a, \tag{5.1}$$

where $S(\partial_x)$ is skew-symmetric, the matrices E^a are positive definite, and

$$\|(E^a, \partial_t E^a + v^a \cdot \nabla E^a)\|_{L^\infty([0,T]\times\mathbb{R}^d)} + \big\|(E^a)^{-1}\big\|^{-1}_{L^\infty([0,T]\times\mathbb{R}^d)} + \|F^a\|_{L^1_T(L^2)}$$

is uniformly bounded.

As before, the proof proceeds by multiplying by U^a and integrating on $\mathbb{T}^d$. We find that $\partial_t \langle\, E^a U^a, U^a \,\rangle$ is given by

$$\langle ((\partial_t + v^a \cdot \nabla)E^a) U^a, U^a \rangle + \langle\, E^a (\operatorname{div} v^a) U^a, U^a \,\rangle + 2\langle\, F^a, U^a \,\rangle,$$

and hence conclude that U^a is uniformly bounded in $C^0([0,T]; L^2(\mathbb{T}^d))$. Since V^a is uniformly bounded in $C^0([0,T]; L^2(\mathbb{T}^d))$, this yields the desired result. □

Remark 5.3. In the periodic case, as shown by Métivier and Schochet [21, 22] as well as Bresch, Desjardins, Grenier, and Lin [6], the study of the behavior of the solutions when $\varepsilon \to 0$ involved many additional phenomena.

6. Low Mach number combustion. The system (1.1) is relevant whenever all nuclear or chemical reactions are frozen, which is the case in many treatments of fluid mechanics. By contrast, for the combustion, one has to replace the energy evolution equation by

$$\partial_t(\rho e) + \operatorname{div}(\rho v e) + P \operatorname{div} v = \kappa \operatorname{div}(k \nabla T) + F(Y),$$

with $Y := (Y_1, \dots, Y_L)$, where the Y_ℓ's denote the relative concentrations of nuclear or chemical species. The new unknown Y_ℓ satisfies

$$\partial_t(\rho Y_\ell) + \operatorname{div}(\rho v Y_\ell) = \lambda \operatorname{div}(D_\ell \nabla Y_\ell) + \rho \omega_\ell(t,x), \tag{6.1}$$

where ω_ℓ is a given source term, $D_\ell > 0$, and λ measures the importance of diffusion processes.

Many results have been obtained for the reactive gas equations (see [7] and the references therein). Yet, the previous studies do not include the dimensionless numbers. Here we consider the system

$$\begin{cases} \alpha(\partial_t P + v \cdot \nabla P) + \operatorname{div} v = \kappa\beta \operatorname{div}(k \nabla T) + F_1(Y,T,P), \\ \rho(\partial_t v + v \cdot \nabla v) + \dfrac{\nabla P}{\varepsilon^2} = \mu\big(2 \operatorname{div}(\zeta D v) + \nabla(\eta \operatorname{div} v)\big), \\ \gamma(\partial_t T + v \cdot \nabla T) + \operatorname{div} v = \kappa\delta \operatorname{div}(k \nabla T) + F_3(Y,T,P), \\ \rho(\partial_t Y + v \cdot \nabla Y) = \lambda \operatorname{div}(D \nabla Y), \end{cases} \tag{6.2}$$

where α, β, γ, and δ are given functions of (Y,T,P).

As explained in the introduction, it is convenient to introduce (p,θ,y) by $P=\underline{P}e^{\varepsilon p}$, $T=\underline{T}e^{\theta}$, $Y=\underline{Y}e^{y}$, where $(\underline{P},\underline{T},\underline{Y})\in[0,+\infty)^{2+L}$. For smooth solutions, (p,v,θ,y) satisfies a system of the form

$$\text{(6.3)}\qquad \begin{cases} g_1(\Phi)(\partial_t p+v\cdot\nabla p)+\dfrac{1}{\varepsilon}\operatorname{div}v=\dfrac{\kappa}{\varepsilon}\chi_1(\Phi)\operatorname{div}(k(\theta)\nabla\theta)+\dfrac{1}{\varepsilon}Q_1(\Phi),\\ g_2(\Phi)(\partial_t v+v\cdot\nabla v)+\dfrac{1}{\varepsilon}\nabla p=\mu\chi_2(\Phi)\big(\operatorname{div}(\zeta(\theta)Dv)+\nabla(\eta(\theta)\operatorname{div}v)\big),\\ g_3(\Phi)(\partial_t\theta+v\cdot\nabla\theta)+\operatorname{div}v=\kappa\chi_3(\Phi)\operatorname{div}(k(\theta)\nabla\theta)+Q_3(\Phi),\\ g_4(\Phi)(\partial_t y+v\cdot\nabla y)=\lambda\chi_4(\Phi)\operatorname{div}(D(\theta)\nabla y), \end{cases}$$

where $\Phi=(y,\theta,\varepsilon p)$.

ASSUMPTION 6.1. *Denote by $(\mathsf{y},\vartheta,\wp)\in\mathbb{R}^N$ the place holder of the unknown $(y,\theta,\varepsilon p)$. Parallel to Assumption $(H2)$ in section 1.2, we suppose that g_i and χ_i $(i=1,2,3)$ are C^∞ positive functions of $(\mathsf{y},\vartheta,\wp)\in\mathbb{R}^N$, $\chi_1<\chi_3$, and there exist two functions F and G such that $(\mathsf{y},\vartheta,\wp)\mapsto(\mathsf{y},F(\mathsf{y},\vartheta,\wp),\wp)$ and $(\mathsf{y},\vartheta,\wp)\mapsto(\mathsf{y},\vartheta,G(\vartheta,\wp))$ are C^∞ diffeomorphisms from $\mathbb{R}^N$ onto $\mathbb{R}^N$, F and G vanish at the origin, and*

$$g_1\frac{\partial F}{\partial\vartheta}=-g_3\frac{\partial F}{\partial\wp}>0,\quad g_1\chi_3\frac{\partial G}{\partial\vartheta}=-g_3\chi_1\frac{\partial G}{\partial\wp}<0.$$

Moreover, Q_1 and Q_3 are C^∞ functions of $(\mathsf{y},\vartheta,\wp)$ vanishing at the origin.

Introduce

$$B:=\big\{(\varepsilon,\mu,\kappa,\lambda)\in(0,1]\times[0,1]\times[0,1]\times[0,2]\,|\,\lambda\geqslant\sqrt{\mu+\kappa}\,\big\}.$$

DEFINITION 6.2. *Let $T>0$, $s\in\mathbb{R}$, $b=(\varepsilon,\mu,\kappa,\lambda)\in B$, and set $a:=(\varepsilon,\mu,\kappa)$. The space $\mathcal{Z}^s_b(T)$ consists of these $(p,v,\theta,y)\in C^0([0,T];H^s(\mathbb{R}^d))$ such that*

$$(p,v,\theta)\in\mathcal{X}^s_a(T),\quad \nu y\in C^0([0,T];H^{s+1}(\mathbb{R}^d)),\quad \lambda y\in L^2(0,T;H^{s+2}_\nu(\mathbb{R}^d)),$$

where $\nu:=\sqrt{\mu+\kappa}$ and $\mathcal{X}^s_a(T)$ is as defined in Definition 3.1. The space $\mathcal{Z}^s_b(T)$ is given the norm

$$\|(p,v,\theta,y)\|_{\mathcal{Z}^s_b(T)}:=\|(p,v,\theta)\|_{\mathcal{X}^s_a(T)}+\|y\|_{L^\infty_T(H^{s+1}_\nu)}+\sqrt{\lambda}\,\|y\|_{L^2_T(H^{s+2}_\nu)}.$$

Having proved estimates for the solutions of system (1.4) with precised estimates in terms of the norm Σ of the source term Q (see (3.3)), we are now in position to assert the following.

THEOREM 6.3. *Assume that $d\neq2$. Given $M_0>0$ and $\mathbb{N}\ni s>1+d/2$, there exist $T>0$ and $M>0$ such that, for all $b\in B$ and all initial data $(p_0,v_0,\theta_0,y_0)\in H^{s+1}(\mathbb{R}^d)$ satisfying*

$$\|(\nabla p_0,\nabla v_0)\|_{H^{s-1}}+\|(y_0,\theta_0,\varepsilon p_0,\varepsilon v_0)\|_{H^{s+1}}\leqslant M_0,$$

the Cauchy problem for (1.4) has a unique solution (p,v,θ,y) in the ball $B(\mathcal{Z}^s_b(T);M)$.

Remark 6.4 (low Mach number limit). For the case of greatest physical interest $(d=3)$, notice that Lemma 4.2 implies that, if, in addition, (p_0,v_0) belongs uniformly to $L^2(\mathbb{R}^3)$ and Q_1 satisfies $|Q_1(\mathsf{y},\vartheta,0)|\leqslant K(\mathsf{y},\vartheta)(|\mathsf{y}|^2+|\vartheta|^2)$ for some smooth function K, then the solutions (p,v,θ,y) are uniformly bounded in $C^0([0,T];H^s(\mathbb{R}^3))$. This result has two corollaries. As alluded to in section 4, it allows us to rigorously

justify, at least in the whole space case, the computations given by Majda in [18]. By the way, this proves the well posedness of the Cauchy problem for the zero Mach number combustion in the whole space (this was known only in the periodic case [11]). Moreover, note that the solutions given by Theorem 6.3 satisfy uniform estimates recovering in the limit $\varepsilon \to 0$ those obtained by Embid for the limit system. Finally, we mention that a close inspection of the high frequency regime indicates that the previous analysis applies with $Q_i(\Phi)$ $(i = 1, 3)$ replaced by $\chi_i(\Phi)Q(\Phi, \nabla y, \nabla^2 y)$ for some smooth function Q, yet we will not address this issue.

Appendix. General equations of state. Recall that in order to study the full Navier–Stokes equations (1.1), we chose to work with the unknown (P, v, T). In order to close this system, we must relate (ρ, e) to (P, T) by means of two equations of state: $\rho = \rho(P, T)$ and $e = e(P, T)$. The purpose of this section is to show that Assumption (H2) in section 1.2 is satisfied under general assumptions on the partial derivatives of ρ and e with respect to P and T.

A.1. Computation of the coefficients. We begin by expressing the coefficients g_i and χ_i, which appear in (1.4), in terms of the partial derivatives of ρ and e with respect to P and T. To do that it is convenient to introduce the entropy. Here is where the first identity in (1.5) enters.

ASSUMPTION A.1. *The functions ρ and e are C^∞ functions of $(P, T) \in (0, +\infty)^2$, satisfying*

$$P\frac{\partial \rho}{\partial P} + T\frac{\partial \rho}{\partial T} = \rho^2 \frac{\partial e}{\partial P}\cdot$$

Introduce the 1-form ω defined by $T\omega := \mathrm{d}\, e + P\, \mathrm{d}(1/\rho)$, where we started using the notation $\mathrm{d}\, f = (\partial f/\partial T)\, \mathrm{d}\, T + (\partial f/\partial P)\, \mathrm{d}\, P$. Assumption A.1 implies that $\mathrm{d}\, \omega = 0$. Hence, the Poincaré lemma implies that there exists a C^∞ function $S = S(P, T)$, defined on $(0, +\infty)^2$, satisfying the second principle of thermodynamics:

$$T\, \mathrm{d}\, S = \mathrm{d}\, e + P\, \mathrm{d}(1/\rho). \tag{A.1}$$

By combining the evolution equations for ρ and e with (A.1) written in the form $\rho T\, \mathrm{d}\, S = \rho\, \mathrm{d}\, e - (p/\rho)\, \mathrm{d}\, \rho$, we get an evolution equation for S, so that

$$(\partial_t + v \cdot \nabla) \begin{pmatrix} \rho \\ S \end{pmatrix} = \begin{pmatrix} -\rho & 0 \\ 0 & (\rho T)^{-1} \end{pmatrix} \begin{pmatrix} \operatorname{div} v \\ \kappa \operatorname{div}(k \nabla T) + Q \end{pmatrix}.$$

On the other hand, one has

$$(\partial_t + v \cdot \nabla) \begin{pmatrix} \rho \\ S \end{pmatrix} = J(\partial_t + v \cdot \nabla) \begin{pmatrix} P \\ T \end{pmatrix} \quad \text{with} \quad J = \begin{pmatrix} \partial\rho/\partial P & \partial\rho/\partial T \\ \partial S/\partial P & \partial S/\partial T \end{pmatrix}.$$

Equating both right-hand sides and inverting the matrix J, we obtain

$$\begin{cases} (\partial_t P + v \cdot \nabla P) + a \operatorname{div} v - \kappa b \operatorname{div}(k \nabla T) = bQ, \\ (\partial_t T + v \cdot \nabla T) + c \operatorname{div} v - \kappa d \operatorname{div}(k \nabla T) = dQ, \end{cases} \tag{A.2}$$

where

$$a = \frac{\rho(\partial S/\partial T)}{\det(J)},\ b = -\frac{\partial\rho/\partial T}{\rho T \det(J)},\ c = -\frac{\rho(\partial S/\partial P)}{\det(J)},\ d = \frac{\partial\rho/\partial P}{\rho T \det(J)}\cdot$$

To express the coefficients g_i and χ_i in terms of physically relevant quantities, we need some more notations. We introduce

$$\begin{aligned} &K_T := \frac{1}{\rho}\frac{\partial \rho}{\partial P}, \quad K_P := -\frac{1}{\rho}\frac{\partial \rho}{\partial T}, \quad \mathcal{R} := -\rho\frac{\partial S/\partial P}{\partial \rho/\partial P}, \\ &C_P := T\frac{\partial S}{\partial T}, \quad C_V := T\frac{(\partial S/\partial T)(\partial \rho/\partial P) - (\partial S/\partial P)(\partial S/\partial T)}{\partial \rho/\partial P}. \end{aligned} \tag{A.3}$$

The functions K_T, K_P, C_V, and C_P are known as the coefficient of isothermal compressibility, the coefficient of thermal expansion, and the specific heats at constant volume and pressure, respectively (see section 2 in [12]). The function $\mathcal{R}$ generalizes the usual gas constant: for perfect gases one can check that $\mathcal{R} = R$.

We now have to convert system (A.2) into equations for the fluctuations p and θ as defined by (1.2). Performing a little algebra we find that

$$\begin{cases} \dfrac{K_T C_V P}{C_P}(\partial_t p + v\cdot\nabla p) + \dfrac{1}{\varepsilon}\operatorname{div} v - \dfrac{\kappa}{\varepsilon}\dfrac{K_P}{\rho C_P}\operatorname{div}(kT\nabla\theta) = \dfrac{1}{\varepsilon}\dfrac{K_P}{\rho C_P}Q, \\ \rho(\partial_t v + v\cdot\nabla v) + \dfrac{1}{\varepsilon}P\nabla p = \mu\big(2\operatorname{div}(\zeta Dv) + \nabla(\eta\operatorname{div} v)\big), \\ \rho C_V T(\partial_t\theta + v\cdot\nabla\theta) + \mathcal{R}\rho T\operatorname{div} v - \kappa\operatorname{div}(kT\nabla\theta) = Q. \end{cases}$$

Hence, (p, v, θ) satisfies (1.4) with

$$g_1^* = \frac{K_T C_V P}{C_P},\ g_2^* = \frac{\rho}{P},\ g_3^* = \frac{C_V}{\mathcal{R}},\ \chi_1^* = \frac{K_P}{\rho C_P},\ \chi_2^* = \frac{1}{P},\ \chi_3^* = \frac{1}{\mathcal{R}\rho T}, \tag{A.4}$$

where we used the following notation: for all $f\colon (0,+\infty)^2 \to \mathbb{R}$,

$$f^*(\vartheta, \wp) := f(\underline{T}e^{\vartheta}, \underline{P}e^{\wp}). \tag{A.5}$$

A.2. Properties of the coefficients.

ASSUMPTION A.2. *The functions ρ and e are C^∞ functions of $(P,T) \in (0,+\infty)^2$ such that $\rho > 0$ and*

$$\frac{\partial \rho}{\partial P} > 0, \quad \frac{\partial \rho}{\partial T} < 0 \quad \textit{and} \quad \frac{\partial e}{\partial T}\frac{\partial \rho}{\partial P} > \frac{\partial e}{\partial P}\frac{\partial \rho}{\partial T}. \tag{A.6}$$

Remark A.3. This assumption is satisfied by general equations of state. Indeed, (A.6) just means that the coefficients K_T, K_P, and C_V are positive.

The following result proves that Assumptions A.1 and A.2 imply that our main structural assumption is satisfied.

PROPOSITION A.4. *If Assumptions* A.1 *and* A.2 *are satisfied, then $\chi_1 < \chi_3$ and g_i, χ_i $(i = 1, 2, 3)$ are C^∞ positive functions.*

Proof. In view of (A.4), the proof reduces to establishing that

$$0 < K_T, \quad 0 < K_P, \quad 0 < C_V < C_P \quad \text{and} \quad 0 < \mathcal{R} < \frac{C_P}{TK_P}.$$

The first two inequalities follow from the definitions of K_T and K_P. To prove the last two, we first establish the Maxwell identity $\partial S/\partial P = \rho^{-2}(\partial\rho/\partial T)$. To see this, by (A.1), we compute

$$\frac{\partial S}{\partial P}\,\mathrm{d}\,T \wedge \mathrm{d}\,P = \mathrm{d}\big(T\,\mathrm{d}\,S\big) = \mathrm{d}\left\{\mathrm{d}\,e + P\,\mathrm{d}\left(\frac{1}{\rho}\right)\right\} = -\frac{1}{\rho^2}\frac{\partial\rho}{\partial T}\,\mathrm{d}\,P \wedge \mathrm{d}\,T.$$

Since $\partial\rho/\partial T < 0$, the Maxwell identity implies that $\partial S/\partial P < 0$. By combining this inequality with $\partial\rho/\partial P > 0$, we find $\mathcal{R} > 0$. Also, the identity $\partial S/\partial P = \rho^{-2}(\partial\rho/\partial T)$ implies that

$$\frac{C_P}{C_V} = \frac{(\partial S/\partial T)(\partial\rho/\partial P)}{(\partial S/\partial T)(\partial\rho/\partial P) - \rho^{-2}\big(\partial\rho/\partial T\big)^2},$$

which proves $C_V < C_P$.

In view of (A.1), the assumption $\frac{\partial e}{\partial T}\frac{\partial\rho}{\partial P} > \frac{\partial e}{\partial P}\frac{\partial\rho}{\partial T}$ is equivalent to

$$\frac{\partial S}{\partial T}\frac{\partial\rho}{\partial P} > \frac{\partial S}{\partial P}\frac{\partial\rho}{\partial T}.$$

This inequality has two consequences. First, it implies that $C_V > 0$. Second, it yields

$$\frac{TK_P\mathcal{R}}{C_P} = \frac{(\partial S/\partial P)(\partial\rho/\partial T)}{(\partial S/\partial T)(\partial\rho/\partial P)} < 1.$$

This concludes the proof. ∎

We now discuss the physical meaning of the functions F and G introduced in section 1.2. These are compatibility conditions between the singular terms and the viscous terms. To see this, suppose (p, v, θ) is a smooth solution of (1.4) and let $\Psi = \Psi(\vartheta, \wp) \in C^\infty(\mathbb{R}^2)$. Then $\psi := \Psi(\theta, \varepsilon p)$ satisfies

$$\begin{aligned} &g_1 g_3\big(\partial_t\psi + v\cdot\nabla\psi\big) \\ &\quad + \underbrace{\left(g_1\frac{\partial\Psi}{\partial\vartheta} + g_3\frac{\partial\Psi}{\partial\wp}\right)}_{=:\Gamma_1(\Psi)}\operatorname{div} v = \kappa\underbrace{\left(g_1\chi_3\frac{\partial\Psi}{\partial\vartheta} + g_3\chi_1\frac{\partial\Psi}{\partial\wp}\right)}_{=:\Gamma_2(\Psi)}\big(\operatorname{div}(k(\theta)\nabla\theta) + Q\big), \end{aligned}$$

where the coefficients g_i, χ_i, $\partial\Psi/\partial\vartheta$ and $\partial\Psi/\partial\wp$ are evaluated at $(\theta, \varepsilon p)$. We next show that for appropriate function Ψ one can impose

$$[\Gamma_1(\Psi) = 0 \text{ and } \Gamma_2(\Psi) > 0] \quad \text{or} \quad [\Gamma_1(\Psi) > 0 \text{ and } \Gamma_2(\Psi) = 0]. \tag{A.7}$$

Proposition A.5. *Assume that Assumptions* A.1 *and* A.2 *are satisfied and use the notation* (A.5). *The functions S^* and ρ^* satisfy*

$$g_1\frac{\partial S^*}{\partial\vartheta} = -g_3\frac{\partial S^*}{\partial\wp} > 0, \qquad g_1\chi_3\frac{\partial\rho^*}{\partial\vartheta} = -g_3\chi_1\frac{\partial\rho^*}{\partial\wp} < 0. \tag{A.8}$$

Remark A.6. The fact that $\Psi = S^*$ (or $\Psi = \rho^*$) satisfies the first (respectively, second) set of conditions in (A.7) now follows from $\chi_1 < \chi_3$.

Proof. By (A.4) and the definitions given in (A.3), one has

$$\frac{g_1^*}{g_3^*} = -\frac{P(\partial S/\partial P)}{T(\partial S/\partial T)}. \tag{A.9}$$

By definition (A.5), $\partial f^*/\partial\vartheta = \big[T(\partial f/\partial T)\big]^*$ and $\partial f^*/\partial\wp = \big[P(\partial f/\partial P)\big]^*$. This proves that S^* satisfies the first identity in (A.8). Next, we compute

$$\frac{\chi_1^*}{\chi_3^*} = \frac{(\partial\rho/\partial T)(\partial S/\partial P)}{(\partial\rho/\partial P)(\partial S/\partial T)}.$$

By (A.9), this yields $\chi_1^* g_3^* P(\partial\rho/\partial P) = -\chi_3^* g_1^* T(\partial\rho/\partial T)$. Which proves that ρ^* satisfies the second identity in (A.8). □

Remark A.7. Assumption (H2) in section 1.2 requires, in addition, that $F = S^*$ and $G = \rho^*$ define bijections. This only means that the thermodynamic state is completely determined by (P, T), or (P, S) or (ρ, T).

The following result contains an example of an equation of state such that χ_1 depends on ϑ.

PROPOSITION A.8. *Assume that the gas obeys Mariotte's law: $P = R\rho T$, for some positive constant R, and $e = e(T)$ satisfies $C_V := \partial e/\partial T > 0$. Then, Assumptions* A.1 *and* A.2 *are satisfied. Moreover,*

$$\chi_1^* = R/((C_V(T) + R)P),$$

so that $\chi_1(\vartheta, \wp)$ is independent of ϑ if and only if C_V is constant.

Acknowledgments. I warmly thank Guy Métivier for helpful discussions. Thanks to Didier Bresch, Christophe Cheverry, Raphaël Danchin, and David Lannes for stimulating comments about this work. Thanks also to the referees who pointed to a number of ways in which the presentation of section 3 could be improved.

REFERENCES

[1] T. ALAZARD, *Low Mach number limit of the full Navier-Stokes equations*, Arch. Ration. Mech. Anal., 180 (2006), pp. 1–73.
[2] S. ALINHAC, *Temps de vie des solutions régulières des équations d'Euler compressibles axisymétriques en dimension deux*, Invent. Math., 111 (1993), pp. 627–670.
[3] C. BARDOS AND B. NICOLAENKO, *Navier-Stokes equations and dynamical systems*, in Handbook of Dynamical Systems, vol. 2, North-Holland, Amsterdam, 2002, pp. 503–597.
[4] B. J. BAYLY, C. D. LEVERMORE, AND T. PASSOT, *Density variations in weakly compressible flows*, Phys. Fluids A, 4 (1992), pp. 945–954.
[5] S. BENZONI, R. DANCHIN, AND S. DESCOMBES, *Multi-dimensional Korteweg models*, preprint.
[6] D. BRESCH, B. DESJARDINS, E. GRENIER, AND C.-K. LIN, *Low Mach number limit of viscous polytropic flows: Formal asymptotics in the periodic case*, Stud. Appl. Math., 109 (2002), pp. 125–149.
[7] G.-Q. CHEN, D. HOFF, AND K. TRIVISA, *Global solutions to a model for exothermically reacting compressible flows with large discontinuous initial data*, Arch. Ration. Mech. Anal., 166 (2003), pp. 321–358.
[8] R. DANCHIN, *Zero Mach number limit for compressible flows with periodic boundary conditions*, Amer. J. Math., 124 (2002), pp. 1153–1219.
[9] R. DANCHIN, *Low Mach number limit for viscous compressible flows*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 459–475.
[10] B. DESJARDINS AND C.-K. LIN, *A survey of the compressible Navier-Stokes equations*, Taiwanese J. Math., 3 (1999), pp. 123–137.
[11] P. EMBID, *Well-posedness of the nonlinear equations for zero Mach number combustion*, Comm. Partial Differential Equations, 12 (1987), pp. 1227–1283.
[12] L. C. EVANS, *A survey of entropy methods for partial differential equation*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 409–438.
[13] I. GALLAGHER, *Résultats récents sur la limite incompressible*, Astérisque, 299 (2005), pp. Exp. No. 926, vii, 29–57.
[14] P. GODIN, *The lifespan of a class of smooth spherically symmetric solutions of the compressible Euler equations with variable entropy in three space dimensions*, Arch. Ration. Mech. Anal., 177 (2005), pp. 479–511.
[15] S. KLAINERMAN AND A. MAJDA, *Compressible and incompressible fluids*, Comm. Pure Appl. Math., 35 (1982), pp. 629–651.
[16] R. KLEIN, *Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics. I. One-dimensional flow*, J. Comput. Phys., 121 (1995), pp. 213–237.
[17] D. LANNES, *Sharp estimates for pseudo-differential operators with symbols of limited smoothness and commutator estimates*, J. Funct. Anal., 232 (2006), pp. 495–539.

[18] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Appl. Math. Sci. 53, Springer-Verlag, New York, 1984.

[19] H. Meng and V. Yang, *A unified treatment of general fluid thermodynamics and its application to a preconditioning scheme*, J. Comput. Phys., 189 (2003), pp. 277–304.

[20] G. Métivier and S. Schochet, *The incompressible limit of the non-isentropic Euler equations*, Arch. Ration. Mech. Anal., 158 (2001), pp. 61–90.

[21] G. Métivier and S. Schochet, *Limite Incompressible des Équations d'Euler non Isentropiques*, in Sémin. Équ. Dériv. Partielles, École Polytech., Palaiseau, 2001.

[22] G. Métivier and S. Schochet, *Averaging theorems for conservative systems and the weakly compressible Euler equations*, J. Differential Equations, 187 (2003), pp. 106–183.

[23] S. Schochet, *Singular limits in bounded domains for quasilinear symmetric hyperbolic systems having a vorticity equation*, J. Differential Equations, 68 (1987), pp. 400–428.

[24] S. Schochet, *The mathematical theory of low Mach number flows*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 441–458.

[25] P. Secchi, *On slightly compressible ideal flow in the half-plane*, Arch. Ration. Mech. Anal., 161 (2002), pp. 231–255.

[26] C. Villani, *Limites hydrodynamiques de l'équation de Boltzmann (d'après C. Bardos, F. Golse, C. D. Levermore, P.-L. Lions, N. Masmoudi, L. Saint-Raymond)*, Astérisque, 282 (2002), pp. Exp. No. 893, ix, 365–405.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1214–1234

GRADIENT FLOWS AS A SELECTION PROCEDURE FOR EQUILIBRIA OF NONCONVEX ENERGIES*

CHRISTOPH ORTNER†

Abstract. For atomistic material models, global minimization gives the wrong qualitative behavior; a theory of equilibrium solutions needs to be defined in different terms. In this paper, a concept based on gradient flow evolutions, to describe local minimization for simple atomistic models based on the Lennard–Jones potential, is presented. As an application of this technique, it is shown that an atomistic gradient flow evolution converges to a gradient flow of a continuum energy as the spacing between the atoms tends to zero. In addition, the convergence of the resulting equilibria is investigated in the case of elastic deformation and a simple damaged state.

Key words. gradient flows, λ-convexity, atomistic models, continuum limit

AMS subject classifications. 35A15, 35B38, 35K55, 26B25

DOI. 10.1137/050643982

1. Introduction.

This article is concerned with a possible concept for analyzing elastic energy functionals which do not satisfy the classical coercivity and weak lower semicontinuity conditions of the calculus of variations. The subject of study is the one-dimensional atomistic energy

$$E_{\text{atom}}\big((y_j)_{j=1}^n\big) = \sum_{j=1}^{n} \big[J(y_j - y_{j-1}) + f_j u_j\big], \tag{1.1}$$

where $n \in \mathbb{N}$, and y_j are the positions of the atoms with $y_0 = 0$. The family (f_j) represents a linear applied force. We assume that the Lennard–Jones type potential $J = J(z)$ satisfies

$$\begin{aligned} &J \in \mathrm{C}^2(0,\infty),\\ &J(z) = +\infty \text{ if } z \le 0 \text{ and } J(z) \to +\infty \text{ as } z \to 0,\\ &J'(1) = 0, J''(z) > 0 \text{ in } (0, z_1), \text{ and}\\ &J \text{ is concave, increasing and bounded above in } (z_1, \infty), \end{aligned} \tag{1.2}$$

with $1 < z_1 < +\infty$. The typical shape is shown in Figure 1.1. Note, that the nonconvexity of J is of a much more fundamental type than the geometric nonconvexity of classical elasticity.

It has been noted previously (see, for example, [24]) that, due to the sublinear growth of J, the energy in (1.1) should not be analyzed in terms of global minimization, as this would give unrealistic material behavior. The most popular example given is that a material described by (1.1) would break for arbitrarily small loads if it were to attain its global minimum. We shall describe this in more detail in section 1.1.

In general, for applications in mechanics, it is advantageous to consider metastable states. The difficulty here is that the number of critical points of E_{atom} tends to infinity as $n \to \infty$. Thus, we require a selection criterion to pick the "correct" equilibrium

*Received by the editors November 1, 2005; accepted for publication (in revised form) June 6, 2006; published electronically December 15, 2006.
http://www.siam.org/journals/sima/38-4/64398.html
†Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK (christoph.ortner@comlab.ox.ac.uk).

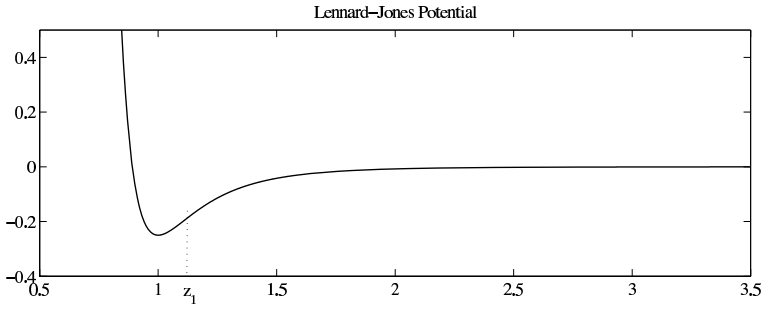


FIG. 1.1. *The shape of natural interaction potentials of Lennard–Jones type.*

points. Theoretically, we should consider the natural dynamics of the material and let time tend to infinity to find its equilibrium state. Here, we take a considerably easier route and use $|\cdot|_{\mathrm{H}^1}$-gradient flow dynamics. Our justification for the gradient flow is merely to accept it as a simple model for local minimization. Concerning the choice of the metric, there are also strong mathematical reasons for choosing an $|\cdot|_{\mathrm{H}^1}$-gradient flow evolution which are outlined in sections 2 and 3. Note that we do not try to analyze a physical evolution. Our aim is to simply demonstrate a concept which we believe gives better results than the traditional method of global minimization. The ideas in this paper have also important applications for the numerical analysis of coarse-graining techniques such as the QC method [17], as they give an indication how numerical optimization methods can be stabilized [20].

The main goal of the present work is to show that the $|\cdot|_{\mathrm{H}^1}$-gradient flow provides a selection criterion for critical points which results in good qualitative properties of the resulting equilibrium model. The simplicity of the one-dimensional model problem makes it possible to give complete results; however, many of the fundamental techniques applied here carry over to much more complicated settings. The additional challenges posed by higher dimensions will be discussed in section 5.

As an application of the idea to use gradient flows to analyze equilibrium points of nonconvex energies, we consider the continuum limit of a rescaled version of the atomistic functional E_{atom} as the number of atoms n tends to infinity. The novelty is that we primarily consider the convergence of the gradient flow evolutions (Theorem 3.1), and obtain the convergence of the equilibria almost as an afterthought (see Theorem 4.1 for elastic deformations and the discussion in section 4.2 for fracture). This procedure gives a different and, one might argue, more realistic continuum limit than previous work; see section 1.2 for a more extensive discussion. In addition, this shows that there is a strong relationship between the atomistic and continuum equilibria.

The local minimizers selected by the gradient flow are weak local minimizers, i.e., local minimizers with respect to the $\mathrm{W}^{1,\infty}$-norm. It is clear from the shape of the interaction potential (cf. Figure 1.1) and the comments at the end of section 4.1 that this is in fact the only possibility. In any weaker topology, even the elastic critical points are not local minimizers of the energy. The same is true for fractured states but the interpretation of $\mathrm{W}^{1,\infty}$ would be more subtle in this case.

If we replace the Lennard–Jones potential by a potential which is smooth at the origin and therefore J' Lipschitz-continuous, then the convergence analysis of the gradient flow requires only minor modifications of the classical convergence analysis of Galerkin discretizations. For the approach in this paper, however, convergence of

the energy is sufficient (cf. Theorem 2.4), which makes a result as general as Theorem 3.1 possible. To achieve this we use some generalizations of ideas in [1, Chapter 4].

For the analysis of equilibria, we use a lim inf condition for the slope of a family of functionals, whose proof is based on the notion of λ-convexity. This condition was also used in [23] to analyze the convergence of gradient flows. Using the techniques of this paper, which has a different aim than the present work, the convergence would have to be obtained by compactness principles (which are not available in our case) rather than λ-convexity.

1.1. The failure of global minimization. The Cauchy–Born hypothesis states that an atomistic body, subjected to a small affine boundary displacement, will follow this displacement in the bulk. Friesecke and Theil demonstrate in [15] a two-dimensional, mathematical version of this important foundation of continuum mechanics by considering global minima of an energy similar to (3.1) but with a quadratic interaction potential. When the potential has sublinear growth, global minimization will typically not reproduce this behavior.

Let us consider the atomistic energy $E_{\rm atom}$ in (1.1) with $f_i \equiv 0$, but apply a "Dirichlet" boundary condition at the right end of the domain as well. For each $\delta > 0$, we consider the minimization problem

$$\min_{y_n = n(1+\delta)} E_{\rm atom}\big((y_j)_{j=1}^n\big). \tag{1.3}$$

Concerning the formulation of the boundary displacement, note that the minimum of $J(z)$ is attained at $z = 1$. The choice of boundary displacement we have made here scales linearly with the number of atoms. An interesting different choice was made in [6] which we discuss briefly in section 1.2.

PROPOSITION 1.1. *There exist constants $\delta_0, C_0 > 0$, such that, for $\delta_0 > \delta > C_0 n^{-1/2}$, the affine state $y_j = (1+\delta)j$ is not the solution of* (1.3).

Proof. Consider the "fractured" deformation $y_j^f = j$ for $j = 0, 1, \ldots, n-1$ and $y_n^f = n(1+\delta)$. Then,

$$E_f(\delta) = E_{\rm atom}\big((y_j^f)_{j=1}^n\big) = (n-1)J(1) + J(1+n\delta) \le (n-1)J(1) + \sup_{z \ge 1} J(z).$$

The affine state $y_j^a = (1+\delta)j$ on the other hand has the energy

$$E_a(\delta) = E_{\rm atom}\big((y_j^a)_{j=1}^n\big) = nJ\big((1+\delta)\big).$$

The assumptions we have made in (1.2) allow us to estimate $J(z)$ from below by a quadratic

$$J(1) + c_0(z-1)^2 \le J(z) \text{ for } 1 \le z < \delta_0,$$

where $c_0 > 0$ and $\delta_0 > 0$ are appropriately chosen. Therefore, for $\delta < \delta_0$, we have $E_a(\delta) \ge n(J(1) + c_0\delta^2)$, and we obtain that $E_a(\delta) > E_f(\delta)$, if

$$\delta^2 \;>\; n^{-1}\Big(\frac{\sup_{z\ge 1} J(z) - J(1)}{c_0}\Big) =: C_0^2 n^{-1}. \qquad \square$$

The proof of Proposition 1.1, which is merely a review of well-known facts, actually suggests that not only is the Cauchy–Born hypothesis violated, but in fact any material with a sufficient number of atoms breaks for arbitrarily small boundary displacements or surface forces, if it were to attain its global energy minimum. This behavior is in clear contradiction to observations and, therefore, global minimization should be rejected for models of the type (1.1).

1.2. Continuum limits of atomistic energies. Continuum limits of atomistic models have been studied by many authors in the past. Because it is customary, we consider the case of Dirichlet boundary conditions in this section only. To be able to compute a continuum limit, we need to first rescale the energy (1.1) to a fixed, finite domain. The seemingly naive approach is to use a linear scaling of the energy as well as the boundary condition, which gives

$$E_n^{(1)}\big((y_j)_{j=0}^n\big) = \sum_{j=1}^n \frac{1}{n} J\big(n(y_j - y_{j-1})\big), \quad y_0 = 0, y_n = 1 + \delta. \tag{1.4}$$

If we assume that the body attains its global energy minimum, then for an arbitrarily small boundary displacement δ, the deformation will not be a continuum state (compare Proposition 1.1). This fact is reflected by the Γ-limit of $E_n^{(1)}$ as $n \to \infty$ (see, for example, [4, 5] and references therein) which gives the energy

$$E^{(1)}(y) = \int_0^1 J^{**}(y')\,\mathrm{d}x, y(0) = 0, y(1) = 1 + \delta,$$

where J^{**} is the convex envelope of J.

Motivated by an analysis quite similar to Proposition 1.1, it can be seen that if a different scaling is used, then the Γ-limit becomes more interesting [6]. If we define

$$E_n^{(2)}\big((u_j)_{j=0}^n\big) = \sum_{j=1}^n \Big[J\big(1 + \sqrt{n}(u_j - u_{j-1})\big) - J(1)\Big], \quad u_0 = 0, u_n = \delta,$$

then the Γ-limit turns out to be the Griffiths functional (compare [13])

$$G(u) = \alpha \int_0^1 |u'|^2\,\mathrm{d}x + \beta \sharp S_u, \quad u(0) = 0, u(1) = \delta,$$

where S_u is the set of jump-discontinuities of the displacement u, $\alpha = 1/2J''(1)$ and $\beta = \lim_{z\to\infty} J(z) - J(1)$. The boundary values of the possibly discontinuous functions u can be interpreted in a meaningful way. While it is interesting that the Griffiths functional can be obtained in this way, it should be noted that this model is typically used for crack propagation only, not crack initiation. In one dimension, however, only crack initiation can be analyzed.

The philosophy adopted in the present work is that the scaling of functional $E_n^{(1)}$ is actually the natural one; only the process of passing to the continuum limit is flawed. It will be shown that, if the continuum limit is analyzed in terms of an appropriate evolution, then the resulting model is in fact a very realistic candidate.

One of the problems addressed in this paper (see section 4) is to find the stable equilibrium that the material would "naturally" assume if we started in the reference configuration $y_i^n = x_i^n$, or a perturbation thereof, and then applied forces. In Theorem 4.1 we show that the resulting equilibria represent the correct elastic behavior. For this reason we prefer to work with surface forces rather than a prescribed displacement. This is, however, not a restriction. The entire convergence theory can also be repeated for Dirichlet conditions applied at both ends of the interval.

Closest in spirit to the approach advocated here is the work by Blanc, Le Bris, and Lions [3]. Except for the fact that they consider far more complicated atomistic interactions in three dimensions, their continuum limit is the same. In fact, the present

work may be seen as a small step towards a rigorous justification of the approach taken in [3].

From the point of view of numerical analysis, strong connections can be drawn to the local version of the quasicontinuum method [17]. In this respect, the results of E and Ming [11] have some similarities to our own.

For results on the continuum manifestation of some further interesting atomistic effects like finite-range interactions, the reader is referred to [25, 9].

1.3. Outline of the paper. We begin in section 2 by outlining the theoretical tools for the convergence analysis, a theory of gradient flows based on the notion of λ-convexity, and a corresponding approximation theory. We also review the notion of slope which is used to define the concept of critical points.

In section 3, we prove the convergence of an atomistic gradient flow evolution to the $|\cdot|_{\mathrm{H}^1}$-gradient flow of a nonconvex functional defined on H^1, giving a new type of continuum limit for atomistic functionals.

Finally, in section 4, we analyze the resulting equilibrium solutions which are obtained when $t \to \infty$ in the gradient flow. First, we consider the case of small loads and show that the equilibria obtained are the physically reasonable elastic deformations and not the "fractured" global energy minima. Then, we give a brief description of the behavior of the gradient flow evolution in the case when the loads are sufficiently large to create fracture. We demonstrate that the obtained equilibrium is reasonable given that we are always assuming perfect crystals and perfect equilibria. However, these critical points are highly unstable, as is demonstrated also in numerical computations. We may interpret this instability as the uncertainty of where fracture occurs in a material.

1.4. Connections to other models. In section 1.2, some connections to the works of Blanc, Le Bris, and Lions [3] and E and Ming [11] were briefly touched upon. In both of these works, the concept of global minimization of the energy is rejected and alternative means are sought to analyze equilibria of elastic energy functionals. A similar approach is taken by Rieger and Zimmer [22], who use a time-discrete gradient flow evolution of Young-measures to analyze material damage. In the slightly different setting of viscoelasticity [21, 2, 14], it is shown that dynamics can prevent the formation of finer and finer microstructure and therefore the attainment of a global energy minimum.

The model presented here is not to be confused, however, with quasistatic or rate independent evolutions (see, for example, [10, 13] for fracture, [8] for plasticity, or [16] for an abstract analysis). In their time-discrete form, at every timestep an equilibrium (typically a minimum) of a functional of the form

$$D(u_{j-1}, u) + E(u) \tag{1.5}$$

is sought, where D is a so-called dissipation metric. Rather, the gradient flow model we present here should be understood as a simple mechanism to find the equilibrium in the quasistatic evolution (1.5).

2. Approximation of gradient flows of nonconvex energies. Let $\mathscr{H}$ be a Hilbert space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$, let $\mathscr{A}$ be a closed convex subset of $\mathscr{H}$, and let $\phi\colon \mathscr{H} \to (-\infty, \infty]$. If ϕ is Fréchet differentiable at a point u, we denote the representation of its derivative, i.e., its gradient, by $\phi'(u)$. Second order derivatives are denoted by $\phi''(u; v_1, v_2)$. We denote the domain of definition of ϕ by $D(\phi) = \{u \in \mathscr{H}\colon \phi(u) < \infty\}$. By using the convention $+\infty \leq +\infty$, we do in fact not

need to make much explicit use of the domain of definition. For example, a functional ϕ would then be convex if and only if $D(\phi)$ is convex and ϕ is convex in $D(\phi)$.

Naively, we may call a curve $u \in \mathrm{C}^1(a,b;\mathscr{H})$ a gradient flow of ϕ, if

$$\dot{u}(t) = -\phi'(u(t)) \quad \forall t \in (a,b). \tag{2.1}$$

Equation (2.1) in infinite-dimensional spaces is usually restated only for convex functionals ϕ. The natural condition on ϕ, under which a considerable part of the theory of gradient flows for convex functionals can be recovered, is the condition of λ-convexity [1]. We say that ϕ is λ-convex in $\mathscr{A}$ if there exists $\lambda \in \mathbb{R}$ such that

$$\phi\big((1-t)v_0 + tv_1\big) \le (1-t)\phi(v_0) + t\phi(v_1) - \frac{\lambda}{2}t(1-t)\|v_0 - v_1\|^2 \quad \forall v_0, v_1 \in \mathscr{A} \quad \forall t \in (0,1). \tag{2.2}$$

To obtain a better feel for the meaning of λ-convexity, consider the following simple proposition (for a proof, see [19]).

Proposition 2.1.

(a) *The functional ϕ is λ-convex in $\mathscr{A}$ if and only if $u \mapsto \phi(u) - \frac{\lambda}{2}\|u\|^2$ is convex in $\mathscr{A}$.*

(b) One-sided Lipschitz continuity of the gradient: *If ϕ is differentiable at every point of $\mathscr{A}$ and satisfies*

$$\big(\phi'(v_1) - \phi'(v_0), v_1 - v_0\big) \ge \lambda\|v_1 - v_0\|^2 \quad \forall\, v_1, v_0 \in \mathscr{A}, \tag{2.3}$$

then ϕ is λ-convex in $\mathscr{A}$.

(c) Boundedness below of the Hessian: *If ϕ is twice differentiable at every nonextremal point of $\mathscr{A}$ and*

$$\phi''(u; v-u, v-u) \ge \lambda\|v-u\|^2 \quad \forall u, v \in \mathscr{A}, \tag{2.4}$$

then ϕ is λ-convex in $\mathscr{A}$.

(d) *If $\phi = \phi_1 + \phi_2$, where $\phi_i \colon \mathscr{A} \to (-\infty, +\infty]$, ϕ_1 is λ_1-convex and ϕ_2 is λ_2-convex, then ϕ is $(\lambda_1 + \lambda_2)$-convex.*

If a functional is λ-convex, then its gradient flows have an alternative characterization. Suppose that a curve $u \in \mathrm{C}^1(a,b;\mathscr{H})$ satisfies (2.1), where ϕ is λ-convex. By a relatively straightforward energy argument, one can show that u also satisfies the evolutionary variational inequality

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t) - v\|^2 + \frac{\lambda}{2}\|u(t) - v\|^2 + \phi(u(t)) \le \phi(v) \quad \forall\, v \in \mathscr{H}, \forall\, t \in (a,b).$$

This inequality is the basis for a powerful theory of gradient flows in metric spaces, then called curves of maximal slope, developed in Chapter 4 of [1]. Note, for example, that it makes sense to consider $u, v \in \mathscr{A}$ only, instead of all of $\mathscr{H}$. Theorem 2.2 is a collection of results in [1] translated to the Hilbert space setting which is sufficient for our purposes.

Theorem 2.2 (existence and uniqueness). *Let $\mathscr{A}$ be a closed, convex subset of a Hilbert space $\mathscr{H}$ and let $\phi \colon \mathscr{A} \to (-\infty, \infty]$ be (strongly) lower semicontinuous and λ-convex. For each $u_0 \in D(\phi)$, there exists a locally Lipschitz-continuous curve $u \colon [0,\infty) \to \mathscr{A}$ which is the unique solution of*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t) - v\|^2 + \frac{\lambda}{2}\|u(t) - v\|^2 + \phi(u(t)) \le \phi(v) \quad \forall\, v \in \mathscr{A} \text{ for a.e. } t > 0, \tag{2.5}$$

among all curves $v \in \mathrm{AC}_{\mathrm{loc}}(0,\infty;\mathscr{A})$*, satisfying* $v(0+) = u_0$*.*

For the remainder of the paper, we shall use the following definition for a gradient flow.

DEFINITION 2.3. *Let* $\mathscr{A}$ *be a closed, convex subset of a Hilbert space* $\mathscr{H}$ *and* $\phi\colon \mathscr{A} \to (-\infty,\infty]$ *a lower semicontinuous and* λ*-convex functional. We say that a locally Lipschitz-continuous curve* $u\colon [0,\infty) \to \mathscr{A}$ *is a gradient flow of* ϕ*, if it satisfies* (2.5)*.*

2.1. Approximation of gradient flows. Based on the evolutionary variational inequality stated above, an abstract convergence theory for gradient flows in a general metric setting for λ-convex functionals was developed in [19]. Theorem 2.4 below is one result therein which is relevant for the Hilbert space setting in the present work. For the sake of completeness, we give a sketch of the proof.

THEOREM 2.4. *Let* $\mathscr{A}$ *be a closed, convex subset of a Hilbert space* $\mathscr{H}$ *and, for* $n \in \mathbb{N}$*, let* $\phi, \phi_n\colon \mathscr{A} \to (-\infty,\infty]$ *be functionals defined on* $\mathscr{A}$*. Let* $u^0 \in D(\phi)$ *and* $u_n^0 \in D(\phi_n)$ *be given initial values, and assume that the following conditions are satisfied:*

(i) Lower semicontinuity: *The functionals* ϕ *and* ϕ_n *(*$n \in \mathbb{N}$*) are lower semicontinuous.*

(ii) Uniform λ-convexity: *There exists* $\lambda \in \mathbb{R}$*, such that the* ϕ_n *as well as* ϕ *are* λ*-convex.*

(iii) Equicoercivity: *There exists a point* $u^* \in \mathscr{A}$ *and* $\epsilon > 0$ *such that* $\inf_{n\in\mathbb{N}} \inf_{v\in\mathscr{A}, \|v-u^*\|\le\epsilon} \phi_n(v) > -\infty$*.*

(iv) Convergence of the initial data: $\sup_{n\in\mathbb{N}} \phi_n(u_n^0) < \infty$ *and* $\|u_n^0 - u^0\| \to 0$ *as* $n \to \infty$*.*

(v) Consistency: *If* $(w_n)_{n\in\mathbb{N}} \subset \mathscr{A}$ *is bounded in* $\mathscr{H}$*, then there exists a constant* $c_1 > 0$ *such that*

$$\limsup_{n\to\infty} \big(\phi(w_n) - \phi_n(w_n)\big) \le 0, \text{ and } \phi(w_n) \le c_1(1 + [\phi_n(w_n)]^+ + \|w_n\|^2).$$

(vi) Best approximation error: *For every* $n \in \mathbb{N}$*, there exists a Borel-measurable curve* $v_n\colon (0,\infty) \to \mathscr{A}$*, so that* $v_n \to u$ *in* $\mathrm{L}^2_{\mathrm{loc}}([0,\infty);\mathscr{H})$ *and*

$$\phi_n(v_n(t)) \to \phi(u(t)) \text{ and } \phi_n(v_n(t)) \le c_2(1 + [\phi(u(t))]^+ + \|u(t)\|^2),$$

where u *is the gradient flow of* ϕ *with initial data* u^0*.*

Then the gradient flows (in the sense of Definition 2.3*)* u_n *of* ϕ_n *with initial values* u_n^0 *converge in* $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathscr{H})$ *to the gradient flow* u *of* ϕ *with initial value* u^0*.*

Proof. Let u and u_n, respectively, satisfy

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)-v\|^2 + \frac{\lambda}{2}\|u(t)-v\|^2 + \phi(u(t)) \le \phi(v) \quad \forall\, v \in \mathscr{A}, \text{ and} \tag{2.6}$$

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_n(t)-v_n\|^2 + \frac{\lambda}{2}\|u_n(t)-v_n\|^2 + \phi(u_n(t)) \le \phi(v_n) \quad \forall\, v_n \in \mathscr{A}. \tag{2.7}$$

We test (2.6) with $v = u_n$ and choose a recovery sequence v_n satisfying (vi) to test (2.7). Adding (2.6) and (2.7) and some lengthy but relatively straightforward algebra gives the error estimate

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u-u_n\|^2 + \frac{\tilde\lambda}{2}\|u-u_n\|^2 \le \big(\phi_n(v_n) - \phi(u)\big) + \big(\phi(u_n) - \phi_n(u_n)\big) \\ + \frac{|\lambda|}{2}\|v_n-u\|^2 + \frac{1}{2}\|\dot u_n\|\|v_n-u\|,\end{aligned}$$

where $\tilde{\lambda} = \lambda - |\lambda|/2$. The λ-convexity can be used to derive an a priori estimate on the $\mathrm{L}^2(0,T)$-norm of $\|\dot{u}_n\|$ and $\phi(u_n)$. Using Gronwall's inequality, we obtain

$$\mathrm{e}^{2\tilde{\lambda}T}\|u(T) - u_n(T)\|^2 \leq \|u(0) - u_n(0)\|^2 + \int_0^T \mathrm{e}^{2\tilde{\lambda}t}(\text{error terms})\,\mathrm{d}t.$$

Using Fatou's lemma, the integral term on the right-hand side can be shown to tend to zero as $n \to \infty$, given the hypothesis of the theorem. □

Next, we state a result from [1], concerning the implicit Euler approximation of a gradient flow, which we will use frequently in section 4.

LEMMA 2.5. *Let $t_j = j\tau$, for $j = 0, 1, \ldots$, define a partition of $[0,\infty)$, with $0 < \tau < 1/\min(0, -\lambda)$. Let $u_0 \in \mathscr{H}$, and let the family $(u_i)_{i=1,2,\ldots}$ be defined by*

$$u_i = \operatorname{argmin}_{\mathscr{A}} \left[v \mapsto \frac{\|v - u_{i-1}\|^2}{2\tau} + \phi(v) \right].$$

Let $u(t)$ be the gradient flow of ϕ with $u(0) = u_0$ and let $\bar{u}_\tau(t)$ be the piecewise constant interpolant of (u_i), i.e.,

$$\bar{u}_\tau(0) = 0 \quad \text{and} \quad \bar{u}_\tau(t) = u_i \quad \text{if } t_{i-1} < t \leq t_i.$$

Then, $\bar{u}_\tau(t) \to u(t)$ in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty), \mathscr{H})$, as $\tau \to 0$.

2.2. The slope. So far we have only described gradient flow evolutions. However, we are also interested in analyzing the resulting equilibria, which can often be obtained by letting time tend to infinity. A natural concept of equilibrium, or critical point, is given by the concept of local slope,

$$|\partial\phi|(u) = \limsup_{v \to u} \frac{(\phi(u) - \phi(v))^+}{\|u - v\|}. \tag{2.8}$$

We say that $u^* \in \mathscr{H}$ is a critical point of the functional ϕ, if $|\partial\phi|(u^*) = 0$. The following lemma can be used in certain situations to show that an accumulation point of critical points of approximate functionals ϕ_n must again be a critical point.

LEMMA 2.6. *Let $\mathscr{H}$ be a Hilbert space, let $\phi, \phi_n \colon \mathscr{H} \to (-\infty, \infty]$ be λ-convex, with a uniform λ, and suppose that ϕ_n Γ-converges to ϕ in the strong topology of $\mathscr{H}$, i.e.,*

$$v_n \to v \Rightarrow \phi(v) \leq \liminf_{n\to\infty} \phi_n(v_n) \tag{2.9}$$

$$\forall v \in \mathscr{H}\ \exists (v_n)_{n\in\mathbb{N}} \subset \mathscr{H} \text{ s.t. } v_n \to v \text{ and } \phi(v) = \lim_{n\to\infty} \phi(v_n). \tag{2.10}$$

Then, the slopes satisfy the lim inf condition

$$u_n \to u \Rightarrow |\partial\phi|(u) \leq \liminf_{n\to\infty} |\partial\phi_n|(u_n). \tag{2.11}$$

Proof. The crucial observation [1, Theorem 2.4.9] is that for λ-convex functionals, the slope can be rewritten as

$$|\partial\phi|(u) = \sup_{v \neq u} \left[\frac{\phi(u) - \phi(v)}{\|u - v\|} + \frac{\lambda}{2}\|u - v\|^2 \right]^+.$$

Let $u_n \to u$, and for some fixed $v \neq u$ let $(v_n)_{n\in\mathbb{N}}$ be a recovery sequence for v, satisfying (2.10). Then, we have

$$\begin{aligned}
&\left[\frac{\phi(u)-\phi(v)}{\|u-v\|} + \frac{\lambda}{2}\|u-v\|^2\right]^+ \\
&\leq \left[\frac{\liminf_{n\to\infty}\phi_n(u_n) - \lim_{n\to\infty}\phi_n(v_n)}{\lim_{n\to\infty}\|u_n-v_n\|} + \frac{\lambda}{2}\lim_{n\to\infty}\|u_n-v_n\|^2\right]^+ \\
&\leq \liminf_{n\to\infty}\left[\frac{\phi_n(u_n)-\phi_n(v_n)}{\|u_n-v_n\|} + \frac{\lambda}{2}\|u_n-v_n\|^2\right]^+ \\
&\leq \liminf_{n\to\infty} |\partial\phi_n|(u_n).
\end{aligned}$$

Taking the supremum over $v \neq u$, we obtain (2.11). ∎

3. Convergence of an atomistic evolution. In section 1.2, it was outlined that different scalings of the atomistic energy $E_{\rm atom}$ give rise to different continuum limits. We have adopted the point of view that a linear scaling of all terms considered is the most natural choice. For the forces we assume that $f_n = O(1)$ and $f_j = O(1/n)$ for $1 \leq j \leq n-1$, i.e., f_n represents a boundary force. It is then natural to consider the rescaled energy

$$E_n\big((y_j^n)_{j=1}^n\big) = \sum_{j=1}^n \epsilon_n \left[J\left(\frac{y_j^n - y_{j-1}^n}{\epsilon_n}\right) - f_j^n(y_j^n + y_{j-1}^n)/2\right] - g y_n^n, \tag{3.1}$$

where $\epsilon_n = 1/n$. The family $(f_i^n)_{i=1,\dots,n}$ defines a linear body force, which we assume is obtained by averaging an L^1 function, i.e.,

$$f_i^n = \fint_{x_{i-1}^n}^{x_i^n} f(x)\,\mathrm{d}x,$$

where $x_i^n = i/n$, for each $i \in \mathbb{Z}$. The scalar g describes a linear surface force. For technical reasons, we may wish to impose an L^∞ bound on the deformations, i.e., we shall assume that $y_i^n \leq M$, where $M \in (z_1, \infty]$.

To rewrite E_n as an integral functional it is customary to identify the atomistic deformation with a piecewise affine function. To this end, we define the set of "admissible" atomistic deformations to be

$$\mathscr{A}_n := \big\{v \in \mathrm{H}^1(0,1)\colon v(0) = 0, v \leq M, \text{ and } v \text{ is piecewise affine w.r.t. } (x_i^n)\big\}.$$

Letting

$$\begin{aligned}
y_n'(x) &= \frac{y_i^n - y_{i-1}^n}{\epsilon_n} \quad \text{if } x \in (x_{i-1}^n, x_i^n), \text{ and} \\
y_n(x) &= \int_0^x y_n'(x)\,\mathrm{d}x,
\end{aligned}$$

y_n is the piecewise-affine interpolant of (y_i^n) and y_n' is its weak derivative, and we have in particular that $y_n \in \mathscr{A}_n$. Thus, we can rewrite E_n as

$$E_n(y_n) = \int_0^1 \big[J(y_n') - f_n y_n\big]\,\mathrm{d}x - g y_n(1) \qquad \text{for } y_n \in \mathscr{A}_n, \tag{3.2}$$

where f_n is the piecewise constant interpolant of f with

$$f_n(x) = f_n^i \quad \text{for } x \in (x_{i-1}, x_i). \tag{3.3}$$

In the formulation (3.2) it becomes obvious that the nonconvexity is with respect to the deformation gradient. In order to balance it out with the evolution, we need to consider the gradient flow with respect to the $|\cdot|_{\mathrm{H}^1}$-seminorm, which is in fact a norm in the spaces $\mathscr{A}_n$. We shall show below, though it is already quite obvious at this point, that the functionals E_n are uniformly λ-convex in the $|\cdot|_{\mathrm{H}^1}$-seminorm. Therefore, from Theorem 2.4, we expect the correct limit energy with respect to the $|\cdot|_{\mathrm{H}^1}$-gradient flow evolution to be

$$E(y) = \int_0^1 \big[J(y') - fy\big]\,\mathrm{d}x - gy(1), \tag{3.4}$$

defined for $y \in \mathscr{A} := \{v \in \mathrm{H}^1(0,1)\colon v(0) = 0, v \le M\}$.

While it is possible to consider gradient flows with respect to the full H^1-norm as well, the analysis of equilibria becomes significantly more technical. In addition, the $|\cdot|_{\mathrm{H}^1}$-seminorm seems to be the more natural metric for the gradient flow. All results can, however, be translated to the H^1-norm case [18].

Theorem 3.1 states that the (atomistic) $|\cdot|_{\mathrm{H}^1}$-gradient flow of E_n in $\mathscr{A}_n$ converges to the (continuum) $|\cdot|_{\mathrm{H}^1}$-gradient flow of E in $\mathscr{A}$. We embed $\mathscr{A}_n$ in $\mathscr{A}$ by setting $E_n(y) = +\infty$ if $y \in \mathscr{A} \setminus \mathscr{A}_n$.

THEOREM 3.1. *Let $y^0 \in D(E)$, and let $y_n^0 \in \mathscr{A}_n$ be the piecewise affine interpolant of y^0 with respect to the mesh (x_i^n). Then, the $|\cdot|_{\mathrm{H}^1}$-gradient flow y_n of E_n with initial data y_n^0 converges in $\mathrm{L}^\infty_{\mathrm{loc}}([0,\infty);\mathscr{A})$ to the $|\cdot|_{\mathrm{H}^1}$-gradient flow y of E with initial data y^0.*

The convergence proof consists of three steps: first, establishing the λ-convexity of the functionals; second, estimating the perturbations caused by the discrete forcing term; and third, constructing a recovery sequence for the solution which satisfies condition (vi) of Theorem 2.4.

LEMMA 3.2. *With respect to the norm $|\cdot|_{\mathrm{H}^1}$, the functionals E and E_n ($n \in \mathbb{N}$) are λ-convex in $\mathscr{A}$, with $\lambda = \min_{z>0} J''(z)$, and lower semicontinuous.*

Proof. For the λ-convexity as well as the lower semicontinuity, note that the linear, continuous terms need not be considered and we assume without loss of generality that $f, g \equiv 0$. In the spirit of Proposition 2.1, we define $F(z) = J(z) - (\lambda/2)z^2$. By the definition of λ, $F''(y) \ge 0$ whenever $y > 0$, hence F is convex in $(0,\infty)$. Since $F(z) = +\infty$ for $z \le 0$, F is convex on $\mathbb{R}$. Therefore, the functional

$$G(y) = \int_0^1 \Big(J(y') - \frac{\lambda}{2}|y'|^2\Big)\,\mathrm{d}x \;=\; \int_0^1 F(y')\,\mathrm{d}x$$

is convex as well which implies, by Proposition 2.1, that E is λ-convex. Since $E(y) = G(y) - \lambda/2|y|^2_{\mathrm{H}^1}$, a sum of a convex and a continuous functional, E is lower semicontinuous. To see that E_n is lower semicontinuous, simply note that under the assumption that $f, g \equiv 0$, $E_n = E|_{\mathscr{A}_n}$, where $\mathscr{A}_n$ is convex and closed and hence the proof carries over to E_n as well. ∎

LEMMA 3.3. *If $f \in \mathrm{L}^1(0,1)$, then, for every $v \in \mathscr{A}$, we have*

$$\Big|\int_0^1 (f_n - f)v\,\mathrm{d}x\Big| \le |v|_{\mathrm{H}^1}\|f - f_n\|_{\mathrm{L}^1(0,1)}, \text{ and}$$
$$\|f - f_n\|_{\mathrm{L}^1} \to 0 \quad \text{as } n \to \infty,$$

where f_n is defined as in (3.3).

Proof. Hölder's inequality gives

$$\Big| \int_0^1 (f_n - f) v \,\mathrm{d}x \Big| \leq \|v\|_{\mathrm{L}^\infty} \|f - f_n\|_{\mathrm{L}^1(0,1)}.$$

Using $v(0) = 0$, we also have $\|v\|_{\mathrm{L}^\infty} \leq \|v'\|_{\mathrm{L}^1} \leq |v|_{\mathrm{H}^1}$, which gives the first result. The convergence $\|f_n - f\|_{\mathrm{L}^1} \to 0$ follows from the fact that f_n is the L^2-projection of f onto the piecewise constant functions with respect to the mesh (x_i^n), using also the density of $\mathrm{L}^2(0,1)$ in $\mathrm{L}^1(0,1)$. ∎

LEMMA 3.4. *Let E and E_n be, respectively, given by* (3.4) *and* (3.2)*, where $f \in \mathrm{L}^1(0,1)$ and f_n satisfies* (3.3)*. For every $y \in \mathscr{A}$ with $E(y) < +\infty$, the piecewise affine, continuous interpolants v_n of y with respect to the mesh (x_i^n) satisfy*

$$|v_n - y|_{\mathrm{H}^1} \to 0, E_n(v_n) \to E(y) \ \text{ as } n \to \infty,$$
$$|v_n|_{\mathrm{H}^1} \leq |y|_{\mathrm{H}^1}, \text{ and } \ E_n(v_n) \leq \left[2\|f\|_{\mathrm{L}^1}^2 + \sup_{z \geq 1} J(z) \right] + E(y) + 2|y|_{\mathrm{H}^1}^2.$$

Proof. Let $y \in \mathscr{A}$, and let v_n be the piecewise affine interpolant with respect to the mesh (x_i^n). Applying Jensen's inequality to

$$\int_{x_{i-1}^n}^{x_i^n} v_n' \,\mathrm{d}x = \int_{x_{i-1}^n}^{x_i^n} y' \,\mathrm{d}x,$$

and summing over i, we get $\|v_n'\|_{\mathrm{L}^2(0,1)} \leq \|y'\|_{\mathrm{L}^2(0,1)}$. It follows from standard interpolation error estimates and a simple density argument that $|y - v_n|_{\mathrm{H}^1} \to 0$ as $n \to \infty$.

To compute the bounds on the energy as well and to show its convergence, we start with the lower-order terms. Jensen's inequality gives $\|f_n\|_{\mathrm{L}^1} \leq \|f\|_{\mathrm{L}^1}$ and as in the proof of Lemma 3.3, $\|v_{(n)}\|_{\mathrm{L}^\infty} \leq |v_{(n)}|_{\mathrm{H}^1} \leq |v|_{\mathrm{H}^1}$. Thus, we have

$$\begin{aligned} -\int_0^1 f_n v_n \,\mathrm{d}x &= -\int_0^1 fy \,\mathrm{d}x + \int_0^1 \big[f(y - v_n) + (f - f_n) v_n \big] \,\mathrm{d}x \\ &\leq -\int_0^1 fy \,\mathrm{d}x + \|f\|_{\mathrm{L}^1} \|y - v_n\|_{\mathrm{L}^\infty} + \|f - f_n\|_{\mathrm{L}^1} \|v_n\|_{\mathrm{L}^\infty} \\ &\leq -\int_0^1 fy \,\mathrm{d}x + \|f\|_{\mathrm{L}^1} |y - v_n|_{\mathrm{H}^1} + \|f - f_n\|_{\mathrm{L}^1} |v|_{\mathrm{H}^1} \qquad (3.5) \\ &\leq -\int_0^1 fy \,\mathrm{d}x + 2\|f\|_{\mathrm{L}^1}^2 + 2|y|_{\mathrm{H}^1}^2. \qquad (3.6) \end{aligned}$$

Using Lemma 3.3 and the fact that $v_n(1) = y(1)$ for all $n \in \mathbb{N}$, we obtain from (3.5) and (3.6),

$$\begin{aligned} &-\int_0^1 f_n v_n \,\mathrm{d}x - g v_n(1) \to -\int_0^1 fy \,\mathrm{d}x - gy(1) \ \text{ as } n \to \infty, \text{ and} \qquad (3.7) \\ &-\int_0^1 f_n v_n \,\mathrm{d}x - g v_n(1) \leq -\int_0^1 fy \,\mathrm{d}x - gy(1) + 2\|f\|_{\mathrm{L}^2(0,1)}^2 + 2|y|_{\mathrm{H}^1}^2. \end{aligned}$$

To deal with the higher-order terms, let $J(z) = J_0(z) + J_1(z)$, where $J_0(z) = J(z)\chi_{(-\infty,1]}(z)$. In the interval (x_{i-1}^n, x_i^n), we have $v_n' = n \int_{x_{i-1}^n}^{x_i^n} y' \,\mathrm{d}x$ and, using

Jensen's inequality $J_0(v_n') \le n \int_{x_{i-1}^n}^{x_i^n} J_0(y')\,\mathrm{d}x$ (note that $1/n$ is the length of the interval). If we define

$$a_n(x) = n \int_{x_{i-1}^n}^{x_i^n} J_0(y')\,\mathrm{d}x + \sup_{z\ge 1} J(z) \quad \text{for } x \in (x_{i-1}^n, x_i^n),$$

then $J(v_n') \le a_n(x)$ a.e. in $(0,1)$ and

$$\int_0^1 a_n(x)\,\mathrm{d}x = \int_0^1 J_0(y')\,\mathrm{d}x + \sup_{z\ge 1} J(z) =: A.$$

In particular, we also have

$$\int_0^1 J(v_n')\,\mathrm{d}x \le \int_0^1 J(y')\,\mathrm{d}x + \sup_{z\ge 1} J(z),$$

which, together with (3.7) gives

$$E_n(v_n) \le \left[2\|f\|_{\mathrm{L}^1}^2 + \sup_{z\ge 1} J(z)\right] + E(y) + 2|y|_{\mathrm{H}^1}^2. \tag{3.8}$$

Since $x \mapsto J_0(y'(x)) \in \mathrm{L}^1(0,1)$, we have, by a slightly stronger version of Lebesgue's differentiation theorem ([12], section 1.7, Corollary 2),

$$\lim_{n\to\infty} a_n(x) = J_0(x) + \sup_{z\ge 1} J(z)$$

for a.e. $x \in (0,1)$, and similarly, $v_n' \to y'$ a.e. in $(0,1)$.

Using Fatou's lemma, and the fact that J is continuous in $(0,\infty)$, we have

$$\begin{aligned} 2A - \limsup_{n\to\infty} \int_0^1 \left|J(v_n') - J(y')\right| \mathrm{d}x &= \liminf_{n\to\infty} \int_0^1 \left[2a_n - |J(v_n') - J(y')|\right] \mathrm{d}x \\ &\ge \int_0^1 \liminf_{n\to\infty} \left[2a_n - |J(v_n') - J(y')|\right] \mathrm{d}x \\ &= 2\int_0^1 \left[J_0(y') + \sup_{z\ge 1} J(z)\right] \mathrm{d}x \\ &= 2A, \end{aligned}$$

and hence, using also (3.7), we have $E(v_n) \to E(y)$ as $n \to \infty$. ∎

We have now assembled all results required to prove Theorem 3.1.

Proof of Theorem 3.1. The result is a straightforward application of Theorem 2.4, using the preparations of this section.

Conditions (i) and (ii) were shown in Lemma 3.2. Condition (iii), the equi-coercivity, follows from the fact that J is bounded below and the forcing term is Lipschitz continuous. Condition (iv), the convergence of the initial data, is guaranteed by standard interpolation error results as well as Lemma 3.4. Condition (v) is controlled by Lemma 3.3, since E_n and $E|_{\mathscr{A}_n}$ differ only in the forcing term.

Let $v_n(t)$ be the piecewise affine interpolant of $y(t)$. Using Lemma 3.4 to obtain (vi) we only need to show that $t \mapsto v_n(t)$ is Borel measurable. In fact, it is fairly easy to see that it is even continuous. Since in one dimension, $\mathrm{H}^1(0,1)$ is embedded

in $\mathrm{C}[0,1]$, the mapping $t \mapsto y(t)$ lies in $\mathrm{C}(0,\infty;\mathrm{C}[0,1])$ and hence $t \mapsto y(t,x)$ is continuous as well. Since

$$v_n(t,x) = \sum_{j=1}^{n} y(t,x_j^n)\varphi_j^n(x),$$

where the φ_j^n are Lipschitz functions, this shows that $v \in \mathrm{C}(0,\infty;\mathrm{H}^1)$. ∎

4. Convergence of equilibria.

4.1. Elastic deformation. In this section we show that the gradient flows are a selection criterion which can be used to recover correct elastic behavior even when the energies are highly nonconvex.

The convergence result of Theorem 3.1 suggests the following procedure: for sufficiently small forces, there should be a critical point y_n^*, in fact a strict local minimum, of the atomistic functional E_n, such that $y_n^{*\prime} < z_1$, i.e., the deformation gradient lies in the region where J is convex. Hence, the gradient flow for sufficiently close starting points should converge to y_n^* as $t \to \infty$ and the deformation gradient should remain within the region where J is convex. Since the atomistic gradient flow converges to the continuum gradient flow, the continuum deformation gradient should remain in this region as well and therefore converge to a critical point in that set which should be the limit of the y_n^*. By y^* being a critical point of ϕ, we mean that $|\partial\phi|(y^*) = 0$, where $|\partial\phi|(y)$ is the $|\cdot|_{\mathrm{H}^1}$-slope of ϕ at y (see section 2.2).

The main difficulty is to show that the critical points y_n^* are "uniform local minimizers" in the sense that we do not require perturbations to tend to zero as $n \to \infty$.

Before we start with the suggested program, let us note that it would be quite easy to show all results for the continuum problem directly. However, we wish to show here that the elastic critical point of the continuum functional (3.4) arises as the limit of the elastic critical points of the atomistic functionals (3.2). Furthermore, it is an interesting feature of the analysis that all information about the continuum functional can be obtained from the knowledge about the atomistic evolution.

THEOREM 4.1. *Let $(E_n)_{n\in\mathbb{N}}, E$ be defined, respectively, by* (3.2) *and* (3.4)*, and assume that $|g| + \|f\|_{\mathrm{L}^1(0,1)} < J'(z_1)$ (compare* (1.2)*).*

(a) *There exist critical points y_n^* of E_n in $\mathscr{A}_n$, such that $y_n^{*\prime} < z_1$. These equilibria are stable in the sense that any $|\cdot|_{\mathrm{H}^1}$-gradient flow y_n of E_n with $y_n'(0,x) < z_1$ satisfies $\lim_{t\to\infty} y_n(t) = y_n^*$ in $\mathrm{H}^1(0,1)$.*

(b) *There exists a critical point $y^* \in \mathscr{A}$ of E such that $\lim_{n\to\infty} y_n^* = y^*$ and $\lim_{t\to\infty} y(t) = y^*$ in H^1, for every $|\cdot|_{\mathrm{H}^1}$-gradient flow y of E with $y'(0,x) \le z_1 - \epsilon$ for some $\epsilon > 0$.*

(c) *If, in addition, $f \equiv 0$, then $y_n^* = y^*$ are affine.*

On the one hand, Theorem 4.1 shows that the derived continuum model has the correct qualitative and quantitative behavior for small loads. On the other hand, it shows that in this situation, the atomistic model behaves essentially like a continuum. In particular, note that point (c) is the Cauchy–Born hypothesis for the model presented.

Note also that not all proofs in this section are "optimal." Particularly, the final proof of Theorem 4.1 is more technical than it needs to be. The purpose of this discussion is to show that most of the techniques used here can be applied to far more general problems and are, in particular, dimension independent.

The proof of Theorem 4.1 requires some preparation in the form of several lemmas which assemble information about the atomistic gradient flow. Let $\mathscr{B}$ be the set of all

deformations whose gradient remains in the region where J is convex, i.e., we define

$$\mathscr{B}_\epsilon = \{v \in \mathscr{A} \colon v'(x) \le z_1 - \epsilon \text{ for a.e. } x \in (0,1)\}, \tag{4.1}$$

and $\mathscr{B} = \mathscr{B}_0$.

LEMMA 4.2. *Suppose that $|g| + \|f\|_{\mathrm{L}^1(0,1)} \le J'(z_1 - \epsilon)$ for some $\epsilon > 0$; then there exists a unique critical point y_n^* of E_n in the set $\mathscr{B}_\epsilon$. The point y_n^* satisfies*

$$y_n^{*\prime}(x) = (J')^{-1}(F_j^n) \le z_1 - \epsilon \quad \text{for } x_{j-1}^n < x < x_j^n, \tag{4.2}$$

where F_j^n is defined by (4.3).

Proof. We compute the critical point by a change of variables. For $y_n \in \mathscr{A}_n$, let $r_j^n = (y_j^n - y_{j-1}^n)/\epsilon_n$. Then, setting

$$\tilde{f}_i^n = \begin{cases} \frac{1}{2} f_1^n & \text{if } i = 0, \\ \frac{1}{2}(f_i^n + f_{i+1}^n) & \text{if } 1 \le i \le n-1, \\ \frac{1}{2} f_n^n & \text{if } i = n, \end{cases}$$

we have, using $y_0^n = 0$,

$$\begin{aligned} E_n(y_n) &= \sum_{j=1}^n \epsilon_n J(r_j^n) - \sum_{j=0}^n \epsilon_n \tilde{f}_j^n y_j^n - g y_n^n \\ &= \sum_{j=1}^n \epsilon_n J(r_j^n) - \sum_{j=1}^n \epsilon_n \tilde{f}_j^n \sum_{i=1}^j \epsilon_n r_i^n - g \sum_{i=1}^n \epsilon_n r_i^n \\ &= \sum_{j=1}^n \epsilon_n J(r_j^n) - \sum_{i=1}^n \epsilon_n r_i^n \left[g + \sum_{j=i}^n \epsilon_n \tilde{f}_j^n \right] \\ &= \sum_{j=1}^n \epsilon_n \big[J(r_j^n) - F_j^n r_j^n \big], \end{aligned}$$

where

$$F_i^n \;=\; g + \sum_{j=i}^n \epsilon_n \tilde{f}_j^n \;=\; g + \frac{\epsilon_n}{2}(f_i^n + f_n^n) + \sum_{j=i+1}^{n-1} \epsilon_n f_j^n. \tag{4.3}$$

To compute r_j^n, we differentiate E_n with respect to r_j^n, which gives the equation

$$\frac{\partial E_n(y_n)}{\partial r_j^n} = \epsilon_n \big[J'(r_j^n) - F_j^n \big] = 0 \quad \text{for } j = 1, \dots, n,$$

or, equivalently, $J'(r_j^n) = F_j^n$. We estimate F_j^n, using the assumption that $\|f\|_{\mathrm{L}^1} + |g| \le J'(z_1 - \epsilon)$, by

$$\begin{aligned} |F_j^n| &= \left| g + \frac{1}{2} \int_{x_{j-1}^n}^{x_j^n} f(x)\,\mathrm{d}x + \int_{x_j^n}^{x_{n-1}^n} f(x)\,\mathrm{d}x + \frac{1}{2} \int_{x_{n-1}^n}^1 f(x)\,\mathrm{d}x \right| \\ &\le |g| + \int_{x_{j-1}^n}^1 |f(x)|\,\mathrm{d}x \\ &\le |g| + \|f\|_{\mathrm{L}^1(0,1)} \\ &\le J'(z_1 - \epsilon). \end{aligned} \tag{4.4}$$

In the region $\{z < z_1\}$, $J'(z)$ is strictly increasing and hence invertible. Therefore,

$$r_j^n = (J')^{-1}(F_j^n) \leq z_1 - \epsilon$$

describes the unique critical point of E_n in $\mathscr{B}_\epsilon$. ∎

LEMMA 4.3. *Under the conditions of Lemma* 4.2*, if* $y_n\colon [0,\infty) \to \mathscr{A}_n$ *is an* $|\cdot|_{\mathrm{H}^1}$*-gradient flow of* E_n *with* $y_n(0) \in \mathscr{B}_\epsilon$*, then* $y_n(t) \in \mathscr{B}_\epsilon$ *for all* $t > 0$.

Proof. Consider the time-discrete approximation $(U_n(t_j))_{j=0,1,\ldots}$, as described in Lemma 2.5, for some fixed, sufficiently small time-step τ. Let $R_n^i(t_j)$ be as in the proof of Lemma 4.2. Then, $R_n(t_j)$ minimizes

$$\frac{1}{2\tau}\|R_n(t_j) - R_n(t_{j-1})\|_{L^2}^2 + E_n(R_n(t_j)). \tag{4.5}$$

As in the proof of Lemma 4.2, we compute the Euler–Lagrange equation in terms of $R_n^i(t_j)$. At the minimum, the equation

$$\frac{1}{\tau}\Big(R_n^i(t_j) - R_n^i(t_{j-1})\Big) = F_j^n - J'(R_n^i(t_j))$$

has to be satisfied. For sufficiently small τ, there is a unique solution. Now assume inductively that $R_n^i(t_{j-1}) \leq z_1 - \epsilon$. To show that $R_n^i(t_j) \leq z_1 - \epsilon$, assume this is not true. Then $F_j^n - J'(R_n^i(t_j)) < 0$, which gives a contradiction. Hence, we have that for all $i = 1, \ldots, n$ and $j \in \mathbb{N}$, $R_n^i(t_j) \leq z_1 - \epsilon$. As $\tau \to 0$, the discrete solution converges to the gradient flow y_n and hence $y_n' \leq z_1 - \epsilon$ a.e. in $(0,1)$. ∎

COROLLARY 4.4. *Under the conditions of Lemma* 4.2*, every* $|\cdot|_{\mathrm{H}^1}$*-gradient flow* y_n *with* $y_n(0) \in \mathscr{B}_\epsilon$ *satisfies the evolutionary variational inequality*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|y_n - v|_{\mathrm{H}^1}^2 + \frac{\alpha}{2}|y_n - v|_{\mathrm{H}^1}^2 + E_n(y_n) \leq E_n(v) \quad \forall v \in \mathscr{B}_\epsilon, \tag{4.6}$$

where $\alpha = \min_{z \leq z_1 - \epsilon} J''(z) > 0$*. In particular, we have*

$$|y_n(t) - y_n^*|_{\mathrm{H}^1} \leq e^{-\alpha t}|y_n(0) - y_n^*|_{\mathrm{H}^1}.$$

Proof. We set $\tilde{E}_n = E_n|_{\mathscr{B}_\epsilon}$ and show that y_n is also a gradient flow for $\tilde{E}_n$ by considering the minimization problem (4.5) again. (Note that this procedure is equivalent to replacing E_n outside of $\mathscr{B}_\epsilon$ by a uniformly convex functional.) Since the minimizer remains in $\mathscr{B}_\epsilon$, it is also the minimizer of

$$\frac{1}{2\tau}\|R_n(t_j) - R_n(t_{j-1})\|_{\mathrm{L}^2}^2 + \tilde{E}_n(R_n(t_j)),$$

and hence the limit of the time-discretizations must also be the gradient flow of $\tilde{E}_n$. By arguing as in the proof of Lemma 3.2, we find that $\tilde{E}_n$ is α-convex (i.e., λ-convex with $\lambda = \alpha$), and hence y_n satisfies (4.6) if we replace E_n with $\tilde{E}_n$. For $v \in \mathscr{B}_\epsilon$, however, the functionals are the same.

On testing (4.6) with $v = y_n^*$, and multiplying the resulting inequality by $e^{2\alpha t}$, we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big(e^{\alpha t}|y_n(t) - y_n^*|_{\mathrm{H}^1}\Big)^2 \leq e^{\alpha t}\big(E_n(y_n^*) - E_n(y_n(t))\big) \leq 0.$$

Integrating from 0 to T gives the result. ∎

Proof of Theorem 4.1. Lemmas 4.2 and 4.3 and Corollary 4.4 immediately imply item (a) and we only need to establish the facts about the continuum limit. Note that almost all of the following analysis is independent of the specific structure of the problem. The only crucial condition which we require is that $y_n(t) \to y(t)$ as $n \to \infty$, for every $t \geq 0$, and $y_n(t) \to y_n^*$ as $t \to \infty$, uniformly in n.

For item (b), we first need to show that, given an initial condition $y(0)$ for the "continuum" $|\cdot|_{\mathrm{H}^1}$-gradient flow satisfying the assumptions of the theorem, there exist "atomistic" initial conditions $y_n(0)$ which satisfy the assumptions of Lemma 4.3. Let $y'(0,x) \leq z_1 - \epsilon$ for a.e. $x \in (0,1)$. Letting $y_n(0,x)$ be the piecewise affine interpolant of $y(0,x)$, we have

$$y_n'(0,x) = \frac{1}{\epsilon_n} \int_{x_{i-1}^n}^{x_i^n} y'(0,x)\,\mathrm{d}x \leq z_1 - \epsilon, \quad x \in (x_{i-1}^n, x_i^n).$$

Therefore, the atomistic $|\cdot|_{\mathrm{H}^1}$-gradient flows with starting point $y_n'(0,\cdot)$ converge uniformly in n (compare to Corollary 4.4) to the equilibria y_n^*, computed in item (a) or Lemma 4.2. We use this fact to estimate

$$\begin{aligned} |y_n^* - y_{n'}^*|_{\mathscr{H}} &\leq |y_n^* - y_n(t)|_{\mathscr{H}} + |y_n(t) - y_{n'}(t)|_{\mathscr{H}} + |y_n(t) - y_{n'}^*|_{\mathscr{H}} \\ &\leq 2\mathrm{const.} e^{-\alpha t} + |y_n(t) - y_{n'}(t)|_{\mathscr{H}}, \end{aligned}$$

thus showing that $(y_n^*)_{n\in\mathbb{N}}$ is a Cauchy-sequence. We denote its limit in $\mathscr{A}$ by y^*. To see that $y(t) \to y^*$ as $t \to \infty$, consider

$$|y(t) - y^*|_{\mathscr{H}} \leq \inf_{n\in\mathbb{N}} \left(|y(t) - y_n(t)|_{\mathscr{H}} + |y_n(t) - y_n^*|_{\mathscr{H}} + |y_n^* - y^*|\right) \leq \mathrm{const.} e^{-\alpha t}.$$

We have shown that the "discrete" equilibria y_n^* converge to a "continuum" deformation y^* and that $y(t) \to y^*$.

The fact that y^* is a critical point of E is easily verified by hand, but in fact this follows from the general theory as well, using the concepts introduced in section 2.2. It is straightforward to show that the functionals E_n $\Gamma(\mathrm{H}^1)$-converge to E in the strong H^1 topology. We merely note the limsup condition (2.10) is given by Lemma 3.4 while for the lim inf condition (2.9) E and E_n can be decomposed into a convex, lower semicontinous part and a continuous, uniformly convergent part (compare to the proof of λ-convexity in Lemma 3.2).

Since the functionals E and $(E_n)_{n\in\mathbb{N}}$ are also uniformly λ-convex, Lemma 2.6, shows that

$$|\partial E|(y^*) \leq \liminf_{n\to\infty} |\partial E_n|(y_n^*) = 0,$$

where $|\partial E_{(n)}|$ denotes the $|\cdot|_{\mathrm{H}^1}$-local slope of the functionals $E_{(n)}$. □

We conclude the discussion of elastic behavior with a remark on the structure of the elastic critical points. It may not be surprising that the continuum "elastic" critical point computed in section 4.1 are actually not local minimizers with respect to the H^1-topology. Indeed, let us assume that $f \equiv 0$ and $0 < g < J'(z_1)$ and define the curve $s \mapsto v(s)$ by

$$v'(s) = y^{*\prime} + \frac{1}{s}\chi_{(1/2,1/2+s^k)}.$$

It is straightforward to establish that for $k \geq 2p$, $v \in \mathrm{C}^{0,1/p}(0, s_0; \mathrm{W}^{1,p})$ and $E(v(s)) < E(y^*)$, where $s_0 > 0$ and $\mathrm{C}^{0,1/p}$ denotes the usual space of Hölder continuous functions. Thus, the critical point y^* is not an H^1-local minimum of the energy $E(y)$. This is also reflected by the fact that we only allow $\mathrm{W}^{1,\infty}$ perturbations in Theorem 4.1.

Why, we should ask ourselves, is this not in contradiction with Theorem 4.1? If there exists a curve along which the energy decreases, should the gradient flow not find this curve? The explanation is that the curve $v(s)$, which we have constructed, is not absolutely continuous in $\mathrm{H}^1(0,1)$ and hence is not a candidate for the gradient flow evolution. An interesting question is whether there actually can exist an absolutely continuous curve starting in y^* along which the energy decreases strictly. Unfortunately, we are unable to answer this question at this point. A negative answer would lead to an interesting selection criterion for equilibria. It would in particular imply that the choice of evolution is not so crucial after all, as such equilibria would be stable under any "sufficiently smooth" evolution.

4.2. Instability and fracture. If the forces f and g are sufficiently strong, then they will cause the material to break, i.e., the atoms debond. Mathematically, this means that the deformation gradient of the atomistic or continuum deformation enters the region where J is concave. In dimensions higher than one, though, the model is unable to describe fracture. There, effects other than debonding of atoms, most notably dislocations, become highly important and cannot be neglected. The discussion in this section can therefore not be generalized directly to higher dimensions.

If we do not restrict the motion of the material, i.e., if we let $M = \infty$ (compare to section 3), then the gradient flows $y_n(t)$ and $y(t)$ will not converge to a stationary point as $t \to \infty$, but diverge. Hence, we restrict the possible deformations by setting M to be a real number, $z_1 < M < \infty$. We assume throughout this section that $f \equiv 0$ and $g > J'(z_1)$.

PROPOSITION 4.5. *There exists $t_1 > 0$ and $\alpha \in \mathrm{W}^{1,\infty}(0,\infty)$ satisfying $\dot\alpha(t) > 0$ if $t < t_1$ and $\alpha(t) = M$ if $t \geq t_1$, such that the solution of the $|\cdot|_{\mathrm{H}^1}$-gradient flow in $\mathscr{A}$ with $y(0,x) = x$ is*

$$y(t,x) = \alpha(t)x.$$

Proof. We change coordinates to $r(t,x) = y'(t,x)$ to obtain, formally for the moment, the equation

$$r_t(t,x) = g - J'(r(t,x)),$$

which is the same ordinary differential equation for every point $x \in (0,1)$. Furthermore, $g - J'(r(t,x)) > 0$ for all x and t, hence $\alpha(t)$ is strictly increasing. Since the solution we have obtained is Lipschitz continuous in time, it is the required gradient flow.

When we reach a time t_1 for which $y(t_1, 1) = M$, the deformation y will be fixed at $y(1,t) = M$ for $t \geq t_1$. To see this, we consider again the time-discretization with initial value $r_0 = M$. The next timestep is the minimizer of

$$\int_0^1 \left[\frac{1}{2\tau}|r - M|^2 + J(r) - gr \right] \mathrm{d}x,$$

subject to $(r) := \int_0^1 r \,\mathrm{d}x \leq M$. If $(r) < M$, then r must satisfy

$$\frac{1}{\tau}(r - M) + J'(r) = g. \tag{4.7}$$

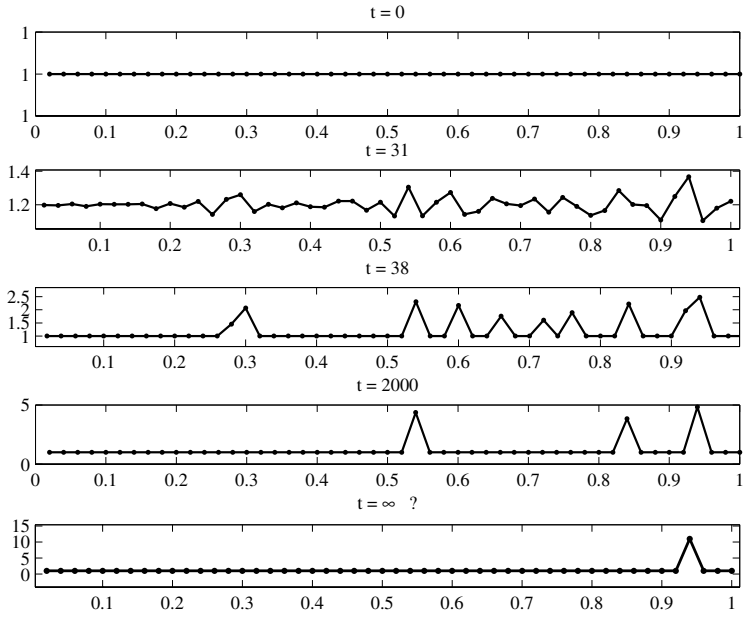


FIG. 4.1. *Snapshots of the deformation gradients of an* $|\cdot|_{\mathrm{H}^1}$*-gradient flow evolution, showing the instability of the final state, computed with* 51 *"atoms." The new final state* ($t = \infty$) *after instability sets in is not computed but guessed. This figure shows an* unstable *computation and should not be mistaken for the exact solution of the model! Note also the different scales in the respective plots.*

Since $(r) < m$, there must exist a set of positive measure where $r \leq M - \epsilon$ for some $\epsilon > 0$. However, since J' is bounded above, (4.7) cannot be satisfied in this set, if τ is sufficiently small.

By a uniqueness argument, we find that $y(t, x)$ satisfies the partial differential equation

$$-y_t'' = J'(y')' = J''(y')y'', \quad y(t,0) = 0, \quad y(t,1) = M, \quad y(t_1, x) = Mx,$$

which can be easily seen to be solved by $y(x,t) = Mx$. Therefore, the evolution remains in the affine state. □

Proposition 4.5 suggests that in our model fracture will never occur. However, the analytical solution obtained is highly unstable under perturbation as Figure 4.1, where a numerical computation is shown, demonstrates. In all computations, we chose $J(z)$ to be strictly increasing for $z > z_1$, i.e., there exists no threshold for the deformation gradient beyond which there are no internal forces.

In a second experiment we dominate the numerical round-off errors, and thus the instabilities in the $|\cdot|_{\mathrm{H}^1}$-gradient flow computation, by a controlled perturbation which could be interpreted, for example, as an impurity in the material. At time $t = 7.6$, we perturb the position of one node (or atom) by an amount of 10^{-8}. The effect of this is that the "fracture" occurs exactly at this position; see Figure 4.2 for the computational results. The instability of the evolution very much conforms with experimental observation that rupture in many types of materials is a highly unstable

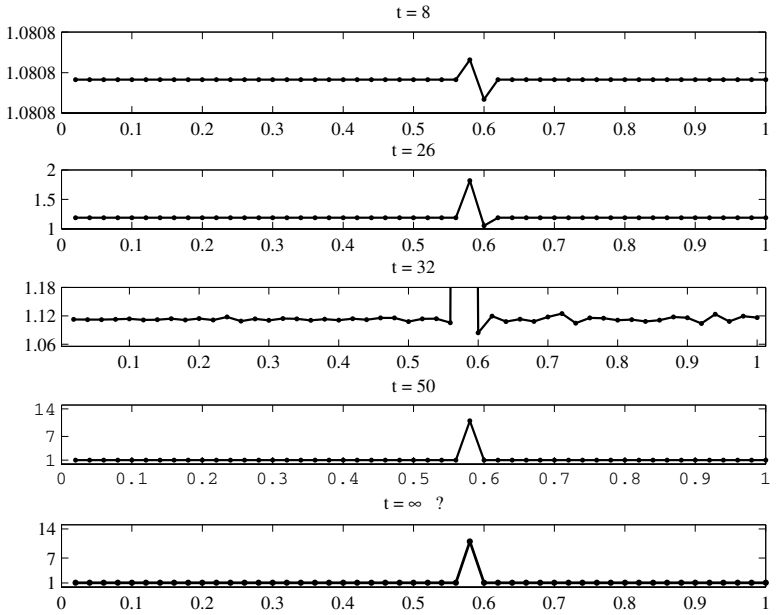


FIG. 4.2. *Snapshots of the deformation gradients of an* $|\cdot|_{\mathrm{H}^1}$*-gradient flow evolution, computed with* 51 *"atoms," with a controlled perturbation at time* $t = 7.6$ *by an amount of* 10^{-8}*. The final state* ($t = \infty$) *is not computed but guessed. Note also the different scales in the respective plots.*

process. No fracture experiment can be reproduced exactly. Thus, the instability of the evolution could be thought of as representing the uncertainty of where damage occurs.

5. Remarks on extensions to two and three dimensions. The simple problem we have investigated here has a fair amount of one-dimensional structure. Although most of the techniques developed here can be readily generalized, the extension to two and three dimensions, which is of great importance to the modeling of material behavior, is not entirely trivial.

The first difficulty to notice is that the passage to higher dimensions in a simple nearest-neighbor system based on the Lennard–Jones potential suffers from a loss of λ-convexity, since the atomistic deformations do not necessarily have to remain orientation preserving. By cutting off the Lennard–Jones potential at the origin, a process which is intuitively reasonable but difficult to justify rigorously, the convergence of the gradient flow can be recovered completely. A more interesting, and mathematically much more challenging, alternative would be to consider a gradient flow with respect to a different metric, which may allow the blow-up behavior of the Lennard–Jones potential, but such a metric seems to be presently unavailable.

To analyze elastic equilibria, it is necessary to obtain L^∞ bounds on the deformation gradient. This step poses the biggest challenge in higher dimensions as these bounds cannot be computed explicitly anymore. One possible avenue to obtain them would be to use the implicit function theorem, for which uniform bounds can be constructed with a slightly refined analysis. It would be necessary, however, that the solution of the linearized system lies in $\mathrm{W}^{1,\infty}(\Omega)$, which can only be obtained in

some very restrictive cases, e.g., with smooth domains and Dirichlet boundary conditions. At re-entrant corners or interfaces between Dirichlet and Neumann boundaries (for example, a crack tip), the nearest neighbor model is too simple to describe the material behavior accurately.

While the convergence theory for gradient flows can still be analyzed if finite-range interactions are added to the energy functional, the analysis of the equilibria seems to be far more difficult if we consider damaged states, but remains essentially unchanged for elastic deformation. The case of infinity-range interactions is completely unclear. For examples of atomistic models with finite-range interactions and their relation to continuum theories, see [25, 9].

Finally, it should be noted that different evolutions can be analyzed as well. For example, it is straightforward to extend the convergence result from the gradient flow evolution to linear viscoelasticity following, for example, the theory developed in [7]. It is more difficult in this setting, however, to analyze the resulting stationary points in similar detail.

Acknowledgments. The initial idea for this work arose in a discussion with Bernd Kirchheim and Georg Dolzmann about generalizations of the work of Friesecke and Theil [15] to realistic interaction potentials during a visit to the Max Planck Institute in Mathematical Sciences, Leipzig. Johannes Zimmer helped with useful comments throughout the work on this paper. I have also benefited from useful discussions with Endre Süli and Benson Muite.

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Math. ETH Zurich, Birkhäuser Verlag, Basel, Switzerland, 2005.

[2] J. M. Ball, P. J. Holmes, R. D. James, R. L. Pego, and P. J. Swart, *On the dynamics of fine structure*, J. Nonlinear Sci., 1 (1991), pp. 17–70.

[3] X. Blanc, C. Le Bris, and P.-L. Lions, *From molecular models to continuum mechanics*, Arch. Ration. Mech. Anal., 164 (2002), pp. 341–381.

[4] A. Braides, G. Dal Maso, and A. Garroni, *Variational formulation of softening phenomena in fracture mechanics: The one-dimensional case*, Arch. Ration. Mech. Anal., 146 (1999), pp. 23–58.

[5] A. Braides and M. S. Gelli, *Continuum limits of discrete systems without convexity hypotheses*, Math. Mech. Solids, 7 (2002), pp. 41–66.

[6] A. Braides, A. J. Lew, and M. Ortiz, *Effective cohesive behavior of layers of interatomic planes*, Arch. Ration. Mech. Anal., 180 (2006), pp. 151–182.

[7] C. Carstensen and G. Dolzmann, *Time-space discretization of the nonlinear hyperbolic system* $u_{tt} = \operatorname{div}(\sigma(Du) + Du_t)$, SIAM J. Numer. Anal., 42 (2004), pp. 75–89 (electronic).

[8] C. Carstensen, K. Hackl, and A. Mielke, *Nonconvex potentials and microstructures in finite-strain plasticity*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 299–317.

[9] M. Charlotte and L. Truskinovsky, *Linear elastic chain with a hyper-pre-stress*, J. Mech. Phys. Solids, 50 (2002), pp. 217–251.

[10] G. Dal Maso and R. Toader, *A model for the quasi-static growth of brittle fractures based on local minimization*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1773–1799.

[11] W. E and P. Ming, *Analysis of multiscale methods*, J. Comput. Math., 22 (2004), pp. 210–219. Special issue dedicated to the 70th birthday of Professor Zhong-Ci Shi.

[12] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[13] G. A. Francfort and J.-J. Marigo, *Revisiting brittle fracture as an energy minimization problem*, J. Mech. Phys. Solids, 46 (1998), pp. 1319–1342.

[14] G. Friesecke and J. B. McLeod, *Dynamics as a mechanism preventing the formation of finer and finer microstructure*, Arch. Ration. Mech. Anal., 133 (1996), pp. 199–247.

[15] G. Friesecke and F. Theil, *Validity and failure of the Cauchy–Born hypothesis in a two-dimensional mass-spring lattice*, J. Nonlinear Sci., 12 (2002), pp. 445–478.
[16] A. Mielke and F. Theil, *On rate-independent hysteresis models*, NoDEA: Nonlinear Differential Equations Appl., 11 (2004), pp. 151–189.
[17] R. E. Miller and E. B. Tadmor, *The quasicontinuum method: Overview, applications, and current directions*, J. Comput.-Aided Materials Design, 9 (2003), pp. 203–239.
[18] C. Ortner, *Continuum Limits of an Atomistic Energy Based on Local Energy Minimization*, Technical report NA05/11, Oxford University Computing Laboratory, Oxford, UK, 2005.
[19] C. Ortner, *Two Variational Techniques for the Approximation of Curves of Maximal Slope*, Technical report NA05/10, Oxford University Computing Laboratory, Oxford, UK, 2005.
[20] C. Ortner and E. Süli, *A Posteriori Analysis and Adaptive Algorithms for the Quasicontinuum Method in One Dimension*, Technical report NA06/13, Oxford University Computing Laboratory, Oxford, UK, 2006.
[21] R. L. Pego, *Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability*, Arch. Ration. Mech. Anal., 97 (1987), pp. 353–394.
[22] M. O. Rieger and J. Zimmer, *Young Measure Flow as a Model for Damage*, preprint, University of Bath, Bath, UK, 2005.
[23] E. Sandier and S. Serfaty, *Gamma-convergence of gradient flows with applications to Ginzburg-Landau*, Comm. Pure Appl. Math., 57 (2004), pp. 1627–1672.
[24] L. Truskinovsky, *Fracture as a phase transformation*, in Contemporary Research in Mechanics and Mathematics of Materials, R. C. Batra and M. F. Beatty, eds., CIMNE, 1996, pp. 322–332.
[25] L. Truskinovsky and A. Vainchtein, *The origin of nucleation peak in transformational plasticity*, J. Mech. Phys. Solids, 52 (2004), pp. 1421–1446.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1235–1248

UNIQUE SOLVABILITY OF A SYSTEM OF NONLINEAR ELLIPTIC PDES ARISING IN SOLID STATE PHYSICS*

X. BLANC†

Abstract. This paper presents a uniqueness theorem for a system of nonlinear elliptic PDEs. This system is related to a quantum chemistry model, namely the Thomas–Fermi–von-Weizsäcker model. The present result is concerned with the solution of a system modeling a semicrystal, i.e., a periodic structure filling a half-space. This kind of uniqueness theorem has been studied in a previous work [I. Catto, C. Le Bris, and P.-L. Lions, *Mathematical Theory of Thermodynamic Limits: Thomas–Fermi Type Models*, Clarendon Press, Oxford University Press, New York, 1998] for the case of a whole crystal. We give here a simpler version of their proof, and then an adaptation to the case of a semicrystal. The methods we use are the maximum principle, supersolution methods, and elliptic estimates.

Key words. nonlinear elliptic PDEs, uniqueness, maximum principle, supersolution methods, thermodynamic limit, energy minimization

AMS subject classifications. 35J45, 35J50, 35J60, 35Q40, 81V55

DOI. 10.1137/05064151X

1. Introduction.

In this article we study the following system:

$$\begin{cases} -\Delta u + u^{7/3} - \phi u = 0, \\ -\Delta \phi = m - u^2, \\ u \geq 0, \end{cases} \tag{1.1}$$

where m is a nonnegative measure satisfying the following hypotheses:

(H0) $\mathrm{Supp}(m) \subset \{x = (x_1, x_2, x_3) \in \mathbb{R}^3, \quad x_3 \leq 0\} := H$;

(H1) $\sup_{x \in H} m(B_1(x)) < +\infty$;

(H2) $\exists R > 0$ such that $\inf_{x \in H} m(B_R(x)) > 0$.

System (1.1) is sometimes called the Thomas–Fermi–von Weizsäcker (TFW) system, in link with the corresponding model. In such a setting, m is the measure defining the nuclei, for example,

$$m = \sum_{i=1}^{N} \delta_{X_i}$$

is the measure of N point nuclei of positions $\{X_i\}_{1 \leq i \leq N}$. These nuclei are surrounded by N electrons of density $\rho \geq 0$, and the corresponding energy reads

$$\mathcal{E}^{\mathrm{TFW}}(\rho, \{X_i\}) = \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2 + \frac{3}{5}\rho^{\frac{5}{3}} - \sum_{i=1}^{N} \int_{\mathbb{R}^3} \frac{\rho(x)}{4\pi|x - X_i|} dx + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{4\pi|x-y|} dxdy + \frac{1}{2} \sum_{i \neq j} \frac{1}{4\pi|X_i - X_j|}.$$

*Received by the editors September 29, 2005; accepted for publication (in revised form) June 12, 2006; published electronically December 15, 2006.
http://www.siam.org/journals/sima/38-4/64151.html

†Laboratoire Jacques-Louis Lions UMR 7598, Université Pierre et Marie Curie, Boîte courrier 187, 75252 Paris Cedex 05, France (blanc@ann.jussieu.fr).

The electrons are assumed to be in their ground state, that is, ρ is the solution of the following minimization problem:

$$\mathcal{E}^{\mathrm{TFW}}(\{X_i\}) = \inf\left\{\mathcal{E}^{\mathrm{TFW}}(\rho,\{X_i\}),\ \rho\geq 0,\ \sqrt{\rho}\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\rho = N\right\}.$$

Setting $u=\sqrt{\rho}$, the corresponding Euler–Lagrange equation reads (see [2]):

$$-\Delta u + u^{7/3} + \left[\left(u^2 - \sum_{i=1}^N \delta_{X_i}\right) * \frac{1}{4\pi|x|}\right] u = \theta u,$$

where $\theta\in\mathbb{R}$ is the Lagrange multiplier associated to the mass constraint. Therefore, setting $m=\sum_{i=1}^N \delta_{X_i}$ and $\phi = (m-u^2)*\frac{1}{4\pi|x|}+\theta$, one finds (1.1). The above derivation may be made rigorous (see [2]) if N is finite (basically because the energy $\mathcal{E}^{\mathrm{TFW}}(\rho,\{X_i\})$ is convex with respect to ρ), but it is not clear when N is infinite, or equivalently when m satisfies hypotheses of the type (H0), (H1), and (H2). This problem is usually referred to as the "thermodynamic limit problem." As is explained in section 4, the main difficulty in such a problem is to prove uniqueness for the limit system (1.1) with hypotheses (H0), (H1), and (H2).

Of course, one may also expect some difficulties for the existence of solutions to (1.1) due to the absence of boundary conditions. The natural strategy to deal with this problem is precisely to use a thermodynamic limit as described in section 4. Another possibility is to define the solution (u_R,ϕ_R) of the system (1.1) set on the ball B_R with Dirichlet boundary conditions, prove that this solution is bounded in suitable spaces, and pass to the limit. Both strategies are equivalent. The first one is described in detail in [7].

In the following, we use the notation (here, Ω is a domain of $\mathbb{R}^3$)

$$L^p_{\mathrm{unif}}(\Omega) = \left\{f\in L^1_{\mathrm{loc}}(\Omega),\quad \sup_{x\in\Omega}\|f\|_{L^p(B_1(x)\cap\Omega)}<\infty\right\}.$$

In [7] it is proved that (1.1) has a unique solution $(u,\phi)\in L^\infty(\mathbb{R}^3)\times L^1_{\mathrm{unif}}(\mathbb{R}^3)$ assuming (H1) and (H2), where H is replaced by $\mathbb{R}^3$ (see Theorem 2.1).

After giving (in section 2) a proof of Theorem 2.1 simpler than that in the original [7], we prove the following result in section 3.

THEOREM 1.1. *Let m be a nonnegative measure satisfying* (H0), (H1), *and* (H2). *Then the system* (1.1) *has a unique solution (u,ϕ) in $L^\infty(\mathbb{R}^3)\times L^1_{\mathrm{unif}}(\mathbb{R}^3)$.*

As already pointed out above, we give in section 4 an application to the TFW theory.

Remark 1.2. The power $\frac{7}{3}$ in (1.1) may be replaced by any $p>1$. Indeed, the important point is that the corresponding TFW energy is strictly convex in ρ. The term u^p in (1.1) corresponds to the term $\int\rho^{\frac{p+1}{2}}$ in the energy, which is strictly convex as far as $p>1$. Note, however, that technically, the case $1<p\leq 3$ may be treated with exactly the same methods, whereas in the case $p>3$, Lemmas 3.2 and 3.3 need to be adapted.

2. The case when m is bounded from below. As announced above, this section is devoted to a proof of Theorem 2.1, which, regarding the uniqueness, is simpler than that given in [7].

THEOREM 2.1 (uniqueness for a TFW crystal, [7]). *Let m be a measure on $\mathbb{R}^3$, which satisfies the following assumptions:*

(H1') $\sup_{x\in\mathbb{R}^3} m(B_1(x)) < +\infty$;

(H2') $\exists R > 0$ *such that* $\inf_{x\in\mathbb{R}^3} m(B_R(x)) > 0$.

Then, problem (1.1) *has a unique solution* $(u,\phi) \in L^\infty(\mathbb{R}^3) \times L^1_{\text{unif}}(\mathbb{R}^3)$.

Proof. We leave aside the question of existence, which is dealt with in [7]. In addition, we only recall, again from [7], that any solution $(u,\phi) \in L^\infty(\mathbb{R}^3)\times L^1_{\text{unif}}(\mathbb{R}^3)$ of (1.1) satisfies the following:

$$u \in C^0(\mathbb{R}^3) \quad \text{and} \quad \nabla u \in L^2_{\text{unif}}(\mathbb{R}^3); \tag{2.1}$$

$$\forall p < 3, \quad \phi \in L^p_{\text{unif}}(\mathbb{R}^3); \tag{2.2}$$

$$\exists \alpha > 0 \quad \text{such that} \quad \forall x \in \mathbb{R}^3, \quad u(x) \geq \alpha. \tag{2.3}$$

For the sake of simplicity, we assume that m is smooth (say, continuous), so that standard elliptic regularity results [8] imply that u and ϕ are locally smooth, and in particular that $u, \phi \in W^{1,\infty}(\mathbb{R}^3)$. The following proof is easily adapted to the case of m being a measure. Details of this generalization can be found in [7]. See also Remark 2.2.

Consider now two solutions, (u,ϕ) and (v,ψ) of (1.1). Setting $w = u - v$ and $\eta = \phi - \psi$, we have

$$\begin{cases} -\Delta w + u^{7/3} - v^{7/3} - \frac{1}{2}(\phi+\psi)w - \frac{1}{2}(u+v)\eta = 0, \\ -\Delta\eta = -(u+v)w. \end{cases} \tag{2.4}$$

Then multiplying the above equations by $w\xi^2$ and $\eta\xi^2$, respectively, where $\xi \in \mathcal{D}(\mathbb{R}^3)$, and integrating over $\mathbb{R}^3$ we get

$$\int_{\mathbb{R}^3} |\nabla(w\xi)|^2 + \left(u^{7/3} - v^{7/3}\right) w\xi^2 - \frac{1}{2}(\phi+\psi)w^2\xi^2 = \int_{\mathbb{R}^3} \frac{1}{2}(u+v)\eta w\xi^2 + w^2|\nabla\xi|^2, \tag{2.5}$$

$$\int_{\mathbb{R}^3} |\nabla(\eta\xi)|^2 = \int_{\mathbb{R}^3} -(u+v)w\eta\xi^2 + \eta^2|\nabla\xi|^2. \tag{2.6}$$

Since the operator $L_u = -\Delta + u^{4/3} - \phi$ satisfies $L_u u = 0$, with $u \geq 0$, we necessarily have $L_u \geq 0$. Similarly, $L_v = -\Delta + v^{4/3} - \psi \geq 0$, so that $\frac{1}{2}(L_u + L_v) \geq 0$:

$$\left\langle \frac{1}{2}\left(L_u + L_v\right)(w\xi), w\xi \right\rangle \geq 0.$$

In addition, we have

$$\left(u^{7/3} - v^{7/3}\right) w \geq \left(u^{4/3} + v^{4/3}\right) w^2. \tag{2.7}$$

Using (2.3), this implies $\left(u^{7/3} - v^{7/3}\right) w \geq \frac{1}{2}\left(u^{7/3} - v^{7/3}\right) w + \nu w^2$, for some constant $\nu > 0$. Inserting this into (2.5), we have

$$\nu \int_{\mathbb{R}^3} w^2\xi^2 \leq \int_{\mathbb{R}^3} \frac{1}{2}(u+v)\eta w\xi^2 + w^2|\nabla\xi|^2.$$

Hence, by using (2.6) we obtain

$$\nu \int_{\mathbb{R}^3} w^2\xi^2 + \frac{1}{4}\int_{\mathbb{R}^3} |\nabla\eta|^2\xi^2 \leq \int_{\mathbb{R}^3} (w^2+\eta^2)|\nabla\xi|^2. \tag{2.8}$$

We then take for ξ a sequence of functions in $\mathcal{D}(\mathbb{R}^3)$ converging to the function

$$\xi(x) = \frac{1}{(1+|x|^2)^{\frac{k}{2}}}, \quad \text{with} \quad \frac{1}{2} < k < 1. \tag{2.9}$$

Note that $\nabla\xi \in L^2(\mathbb{R}^3)$, $\Delta\xi \in L^2(\mathbb{R}^3)$, but $\xi \notin L^2(\mathbb{R}^3)$. Inequality (2.8) implies that

$$\int_{\mathbb{R}^3} w^2\xi^2 < +\infty \quad \text{and} \quad \int_{\mathbb{R}^3} |\nabla\eta|^2\xi^2 < +\infty.$$

So far, our proof reproduces that of [7]. As in [7], we want now to prove that

$$\int \eta^2\xi^2 < +\infty, \tag{2.10}$$

but we will now argue somewhat differently. For this purpose, we multiply the first equation of (2.4) by $\eta\xi^2$, and integrate over $\mathbb{R}^3$

$$\frac{1}{2}\int_{\mathbb{R}^3} (u+v)\eta^2\xi^2 = \int_{\mathbb{R}^3} (-\Delta w)\eta\xi^2 + \left(u^{7/3}-v^{7/3}\right)\eta\xi^2 - \frac{1}{2}\int_{\mathbb{R}^3} (\phi+\psi)w\eta\xi^2. \tag{2.11}$$

Noticing that u, v, ϕ, and ψ are bounded, we have, for some constant C independent of ξ:

$$\int_{\mathbb{R}^3} \left|u^{7/3}-v^{7/3}\right| |\eta|\xi^2 \leq C\int_{\mathbb{R}^3} |w|\,|\eta|\xi^2 \leq C\left(\int_{\mathbb{R}^3} w^2\xi^2\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^3} \eta^2\xi^2\right)^{\frac{1}{2}},$$

along with

$$\int_{\mathbb{R}^3} |\phi+\psi|\,|w|\,|\eta|\xi^2 \leq C\int_{\mathbb{R}^3} |w|\,|\eta|\xi^2 \leq C\left(\int_{\mathbb{R}^3} w^2\xi^2\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^3} \eta^2\xi^2\right)^{\frac{1}{2}},$$

and

$$\begin{aligned}
\left|\int_{\mathbb{R}^3} (-\Delta w)\eta\xi^2\right| &= \left|\int_{\mathbb{R}^3} w\left(-\Delta\left(\eta\xi^2\right)\right)\right| \\
&\leq \left|\int_{\mathbb{R}^3} w(-\Delta\eta)\xi^2\right| + \left|\int_{\mathbb{R}^3} w\eta\left(-\Delta\xi^2\right)\right| + \left|4\int_{\mathbb{R}^3} w\xi\nabla\eta\nabla\xi\right| \\
&\leq \int_{\mathbb{R}^3} (u+v)w^2\xi^2 + C + 4\left(\int_{\mathbb{R}^3} w^2\xi^2\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^3} |\nabla\eta|^2|\nabla\xi|^2\right)^{\frac{1}{2}} \\
&< +\infty.
\end{aligned}$$

Using these three inequalities to bound the right-hand side of (2.11), we find that, since u and v are bounded away from zero,

$$\int_{\mathbb{R}^3} \eta^2\xi^2 \leq C + C\left(\int_{\mathbb{R}^3} \eta^2\xi^2\right)^{\frac{1}{2}} \quad \text{hence} \quad \int_{\mathbb{R}^3} \eta^2\xi^2 < +\infty.$$

The end of the argument then mimics again that of [7]: Defining the function $\xi_\varepsilon(x) = \xi(\varepsilon x)$, we have $|\nabla \xi_\varepsilon|^2 \le \varepsilon^{2-2k}\xi^2$, so

$$\frac{\nu}{(1+\varepsilon^2 R^2)^k}\int_{B_R} w^2 \le \int_{\mathbb{R}^3} w^2\xi_\varepsilon^2 \le \int_{\mathbb{R}^3}(w^2+\eta^2)|\nabla\xi_\varepsilon|^2 \le \varepsilon^{2-2k}\int_{\mathbb{R}^3}\left(w^2+\eta^2\right)\xi^2.$$

Letting ε go to zero, we find that w is zero on B_R. Since this is valid for any $R>0$, we have $w=0$. This in turn implies that $\eta = 0$. □

Remark 2.2. Let us point out here that we have made additional regularity assumptions on the measure m in order to simplify the proof, but the general case may be treated in the same way. Roughly speaking, one needs in this case to use interpolation spaces, noting that $\phi, \psi \in L^{3,\infty}_{\text{unif}}(\mathbb{R}^3)$ (see [3] or [7]) instead of $L^\infty(\mathbb{R}^3)$. This is sufficient to carry out the estimates of the preceeding proof. For instance, (2.8) and (2.9) still imply $\int w^2\xi^2 + \int |\nabla\eta|^2\xi^2 < +\infty$ because

$$\begin{aligned}\int_{\mathbb{R}^3}\eta^2|\nabla\xi|^2 &\le C\sum_{j\in\mathbb{Z}^3}\frac{1}{(1+|j|^2)^{k+1}}\|\eta\|_{L^2_{\text{unif}}(\mathbb{R}^3)}\\ &\le C\sum_{j\in\mathbb{Z}^3}\frac{1}{(1+|j|^2)^{k+1}}\|\eta\|_{L^{3,\infty}_{\text{unif}}(\mathbb{R}^3)} < +\infty.\end{aligned}$$

The same kind of remark holds for the proof of (2.10).

The above proof may be easily adapted to prove the following result, which asserts that the kernel of the linearized operator of system (1.1) is reduced to $\{0\}$.

COROLLARY 2.3. *Let m be a measure satisfying* (H1') *and* (H2'), *and let $(u,\phi) \in L^\infty(\mathbb{R}^3)\times L^1_{\text{unif}}(\mathbb{R}^3)$ be the unique solution of* (1.1). *If $(w,\eta)\in L^\infty(\mathbb{R}^3)\times L^1_{\text{unif}}(\mathbb{R}^3)$ satisfies*

$$\begin{cases} -\Delta w + \frac{7}{3}u^{4/3}w - \phi w - u\eta = 0,\\ -\Delta\eta = -2uw, \end{cases} \tag{2.12}$$

then $w=0$ and $\eta = 0$.

3. Uniqueness for (1.1) with (H0). We now turn to the proof of Theorem 1.1. The problem here is that u is no longer bounded from below. However we will see that $u(x) = u(x_1,x_2,x_3)$ converges to 0 as x_3 goes to infinity, uniformly with respect to x_1 and x_2. A careful analysis of this decay then allows us to circumvent the difficulty.

Here again, we assume that m is smooth, so that by elliptic regularity [8], u and ϕ are smooth. The following proof may be adapted with exactly the technique of [7].

3.1. A priori bounds. We first prove the following lemma.

LEMMA 3.1. *Let m be a smooth function satisfying* (H0), (H1), *and* (H2). *Consider $(u,\phi)\in L^\infty(\mathbb{R}^3)\times L^1_{\text{unif}}(\mathbb{R}^3)$ and $(v,\psi)\in L^\infty(\mathbb{R}^3)\times L^1_{\text{unif}}(\mathbb{R}^3)$ two solutions of* (1.1). *Then, setting $w = u-v$ and $\eta = \phi-\psi$, there exists two constants, $A>0$ and $\alpha>0$, such that*

$$\forall x\in H,\quad |w(x)| + |\eta(x)| \le Ae^{-\alpha|x_3|}. \tag{3.1}$$

Proof. We first prove that w and η go to zero as $x_3 \longrightarrow -\infty$. We argue by contradiction and assume that there exists a sequence x^n such that $x_3^n \longrightarrow -\infty$ as n goes to infinity, and $|w(x^n)| + |\eta(x^n)| \ge \varepsilon$ for some constant $\varepsilon > 0$. Then defining

$$\begin{aligned}\tilde{u}_n(x) = u(x+x^n),\quad \tilde{v}_n(x) &= v(x+x^n),\quad \tilde{\phi}_n(x) = \phi(x+x^n),\\ \tilde{\psi}_n(x) &= \psi(x+x^n),\quad m_n(x) = m(x+x_n),\end{aligned} \tag{3.2}$$

since all these functions are bounded, we can extract a converging subsequence in $\mathcal{D}'(\mathbb{R}^3)$. Elliptic regularity implies that $\tilde{u}_n$, $\tilde{v}_n$, $\tilde{\phi}_n$, and $\tilde{\psi}_n$ also converge in $L^\infty_{\rm loc}(\mathbb{R}^3)$ (at least) to some functions, which we, respectively, denote by $\tilde{u}$, $\tilde{v}$, $\tilde{\phi}$, and $\tilde{\psi}$. The limit of m_n is denoted by $\tilde{m}$, which satisfies (H1') and (H2'). Hence, $(\tilde{u}, \tilde{\phi})$ and $(\tilde{v}, \tilde{\psi})$ are two different bounded solutions of (1.1), with m satisfying the hypotheses of Theorem 2.1. This is a contradiction.

We next prove that the decay is indeed exponential. We set, for any $R > 0$,

$$M(R) = \sup_{x_3 \leq -R} |w(x)| + |\eta(x)|.$$

We want to prove that

$$\exists R > 0, \quad \exists \gamma \in (0,1) \quad \text{s.t.} \quad \forall T \geq 1, \quad M(R+T) \leq \gamma M(T). \tag{3.3}$$

Indeed, this implies that for any $S \geq R+1$, $M(S) \leq \gamma^{S-R-2} M(R+1)$, which implies (3.1).

Arguing again by contradiction, we assume that (3.3) is not satisfied, i.e., that there exist sequences R_n, γ_n, T_n satisfying the following:

$$\left\{ \begin{array}{l} R_n \longrightarrow +\infty, \\ \gamma_n \longrightarrow 1, \\ T_n \geq 1 \end{array} \right. \qquad \text{and} \quad M(R_n + T_n) > \gamma_n M(T_n).$$

Thus, one can find a sequence $x^n \in \mathbb{R}^3$ such that $x^n_3 \leq -R_n - T_n$, $\frac{|w(x^n)|+|\eta(x^n)|}{M(R_n+T_n)} \longrightarrow 1$ and $|w(x^n)|+|\eta(x^n)| \geq \gamma_n M(R_n)$ as n goes to infinity. Using here again the notation (3.2), we find that, extracting a subsequence if necessary, $(\tilde{u}_n, \tilde{\phi}_n)$ converges to a solution of (1.1). Moreover, defining $\tilde{w}_n(x) = \frac{w(x+x_n)}{M(R_n)}$ and $\tilde{\eta}_n(x) = \frac{\phi(x+x_n)}{M(R_n)}$, we also have convergence (possibly after extracting a subsequence) for $(\tilde{w}_n, \tilde{\eta}_n)$. We pass to the limit in the corresponding equation, finding that the limit (w, η) of $(\tilde{w}_n, \tilde{\eta}_n)$ satisfies (2.12). According to Corollary 2.3, we thus have $w = 0$ and $\eta = 0$, which is a contradiction. ∎

Next, we study the decay of u and ϕ in H^c.

LEMMA 3.2. *Let m be a smooth function satisfying* (H0), (H1), *and* (H2). *Consider a solution (u, ϕ) of* (1.1). *Then, there exists a constant $A > 0$ such that*

$$\forall x \in H^c, \quad u(x) \leq \frac{A}{(1+x_3)^{\frac{3}{2}}}, \tag{3.4}$$

$$\forall x \in H^c, \quad \phi(x) \leq \frac{A}{(1+x_3)^2}. \tag{3.5}$$

Proof. These decay estimates are simple applications of a method of Benguria and Lieb [1]. It was also used by Solovej [11]. A complete proof is also given in [4]. We only indicate here the main ideas: Let $\omega \in \mathcal{D}(\mathbb{R}^3)$ be such that ω is radially symmetric, $\omega \geq 0$ and $\int \omega = 1$, and define, for any $R > 0$,

$$g_R(x) = \frac{1}{R^3} \omega\left(\frac{x}{R}\right) \quad \text{and} \quad \overline{\phi}_R = \phi * g_R.$$

We then have, using the fact that the operator $-\Delta + u^{4/3} - \phi$ is nonnegative,

$$\overline{\phi}_R \leq g_R * u^{4/3} + \frac{B}{R^2}, \tag{3.6}$$

where $B > 0$ is a constant depending only on ω. We then point out that ϕ is subharmonic in H^c, and use the mean-value inequality to prove that:

$$\text{if} \quad x_3 \geq R, \quad \overline{\phi}_R(x) \geq \phi(x). \tag{3.7}$$

We then define $\tilde{\phi}_R(x) = \overline{\phi}_R(x) - \frac{B}{R^2}$, and get, using (3.6) and Jensen's inequality,

$$\tilde{\phi}_R \leq \left(g_R * u^2\right)^{2/3}.$$

Hence, convoluting the second equation of (1.1) by g_R,

$$-\Delta\tilde{\phi}_R + \left(\tilde{\phi}_R\right)_+^{3/2} \leq m * g_R.$$

In particular, if $x_3 \geq R$, we have $-\Delta\tilde{\phi}_R(x) + \left(\tilde{\phi}_R(x)\right)_+^{3/2} \leq 0$. Hence, using supersolution methods (note that $U(x) = \frac{C}{(1+x_3)^4}$ is a supersolution of $-\Delta U + U^{3/2} = 0$ in H^c if C is large enough), one easily gets that for any $R > 0$,

$$\tilde{\phi}_R(x) \leq \frac{B'R^2}{(x_3^2 - R^2)^2} \quad \text{if} \quad x_3 > R,$$

where B' is a universal constant. This estimate and (3.7) imply that

$$\text{if} \quad x_3 \geq R, \quad \phi(x) \leq \frac{B'R^2}{(x_3^2 - R^2)^2} + \frac{B}{R^2}.$$

Choosing $R = \frac{x_3}{2}$, we thus find (3.5).

Next, we use the first equation of (1.1) and have $-\Delta u + u^{\frac{7}{3}} = \phi u \leq \frac{3}{7}u^{7/3} + \frac{4}{7}\phi^{7/4}$, so that $-\Delta u + \frac{4}{7}u^{7/3} \leq \frac{4A^{7/4}}{7(1+x_3)^{7/2}}$. Using supersolution methods once again (here, a supersolution is given by $U(x) = \frac{C}{(1+x_3)^{3/2}}$, with $C > 0$ large enough) we find (3.4). □

Estimate (3.5) only provides an upper bound on ϕ. We now improve this estimate and show that $\phi(x)$ converges to some nonpositive constant l as x_3 goes to infinity

LEMMA 3.3. *Let m be a smooth function satisfying* (H0), (H1), *and* (H2). *If $(u, \phi) \in \left(L^\infty(\mathbb{R}^3)\right)^2$ is a solution of* (1.1), *then there exists two constants $C > 0$ and $l \leq 0$ such that*

$$\forall x \in H^c, \quad |\nabla\phi(x)| \leq \frac{C}{(1+x_3)^{7/4}} \tag{3.8}$$

$$\forall x \in H^c, \quad |\phi(x) - l| \leq \frac{C}{(1+x_3)^{3/4}}. \tag{3.9}$$

Proof. We first use standard regularity estimates [8], and get, since $\phi \in C^{2,\alpha}(\mathbb{R}^3)$ for some $\alpha > 0$,

$$\|\nabla\phi\|_{C^0(B_R)} \leq C\left(\frac{1}{R}\|\phi\|_{C^0(B_{2R})} + R\|\Delta\phi\|_{C^0(B_{2R})}\right), \tag{3.10}$$

for any $R > 0$, where C is a constant independent of R. Fixing $x \in H^c$, using $R = \frac{x_3}{4}$ and for B_R the ball centered at x, we have

$$\forall x \in H^c, \quad |\nabla \phi(x)| \leq \frac{C}{1+x_3}. \tag{3.11}$$

Next, we fix again a ball $B_R \subset H^c$, and define in this set two functions ϕ_1 and ϕ_2 by

$$\left\{ \begin{array}{l} -\Delta \phi_1 = -u^2, \\ \phi_{1|\partial B_R} = 0, \end{array} \right. \qquad \left\{ \begin{array}{l} -\Delta \phi_2 = 0, \\ \phi_{2|\partial B_R} = \phi. \end{array} \right.$$

Hence, $\phi = \phi_1 + \phi_2$ in B_R. The maximum principle and (3.4) imply that $-\frac{1}{2x_3} \leq \phi_1(x) \leq 0$ in B_R. We then use the Gagliardo–Nirenberg inequality [10], and get

$$\|\nabla \phi_1\|_{L^\infty(B_R)} \leq C \|\phi_1\|^{\frac{1}{2}}_{L^\infty(B_R)} \|\Delta \phi_1\|^{\frac{1}{2}}_{L^\infty(B_R)},$$

for some constant C independent of R. Hence, fixing $x \in H^c$ and using for B_R the ball $B_R(x)$ with $R = \frac{x_3}{2}$, we have

$$\forall z \in B_R, \quad |\nabla \phi_1(z)| \leq \frac{C}{(1+z_3)^2}. \tag{3.12}$$

We then point out that ϕ_2 is explicitly given by the following formula:

$$\phi_2(z) = -\int_{\partial B_R} \frac{(z-y)\cdot n}{|z-y|^3} \phi(y) d\sigma(y), \tag{3.13}$$

where σ is the Lebesgue measure on ∂B_R, and n denotes the normal vector of ∂B_R at the point y. We then slightly modify B_R, in a way which does not change estimate (3.12), by setting $B_R = B_R(x + Re_3)$, where e_3 is the third vector of the canonical basis. Letting R go to infinity, we get (3.13) with $J_x := \{x + y,\, y \in \partial H\} = \{y \in \mathbb{R}^3,\, y_3 = x_3\}$ instead of ∂B_R, namely

$$\phi_2(z) = (z_3 - x_3) \int_{J_x} \frac{\phi(y)}{|z-y|^3} d\sigma(y).$$

We then differentiate this expression with respect to z_1 and get

$$\begin{aligned} \partial_1 \phi_2(z) &= -3(z_3 - x_3) \int_{J_x} \frac{z_1 - y_1}{|z-y|^5} \phi(y) d\sigma(y) \\ &= -3(z_3 - x_3) \int_{J_x} \frac{z_1 - y_1}{|z-y|^5} \left(\phi(y) - \phi\left[(z_1, z_2, x_3)\right]\right) d\sigma(y). \end{aligned}$$

Hence, setting $R' > 0$ (which will be chosen later on), we get, using (3.11),

$$\begin{aligned} |\partial_1 \phi_2(z)| &\leq C|z_3 - x_3| \int_{y_1^2 + y_2^2 \leq R'^2} \frac{R' dy_1 dy_2}{|x_3| \left((z_3 - x_3)^2 + y_1^2 + y_2^2\right)^{\frac{5}{2}}} \\ &\quad + C|z_3 - x_3| \int_{y_1^2 + y_2^2 \geq R'^2} \frac{dy_1 dy_2}{\left((z_3 - x_3)^2 + y_1^2 + y_2^2\right)^{\frac{5}{2}}} \\ &\leq C \frac{R'}{|x_3|(z_3 - x_3)^2} + C \frac{|z_3 - x_3|}{R'^3}. \end{aligned}$$

Hence, setting $R' = z_3 - x_3$, we get $|\partial_1\phi_2(x)| \leq \frac{C}{x_3(z_3-x_3)} + \frac{C}{(z_3-x_3)^2}$. The same argument holds for $\partial_2\phi_2$ so that, using (3.12), we have, for any $x \in H^c$,

$$\forall z \in \mathbb{R}^3 \quad / \quad z_3 \geq x_3, \quad |\partial_1\phi(z)| + |\partial_2\phi(z)| \leq \frac{C}{x_3(z_3-x_3)} + \frac{C}{(z_3-x_3)^2}. \tag{3.14}$$

We then go back to (3.13), and carry out exactly the same computation as for $\partial_1\phi_2$, but for $\partial_3\phi_2$, finding

$$\forall z \in \mathbb{R}^3 \quad / \quad z_3 \geq x_3, \quad |\partial_3\phi_2(z)| \leq \frac{C}{x_3(z_3-x_3)^{1/2}} + \frac{C}{(z_3-x_3)^{3/2}}.$$

Setting $z_3 = 2x_3$, we thus have

$$\forall x \in H^c, \quad |\nabla\phi(x)| \leq \frac{C}{(1+x_3)^{3/2}}. \tag{3.15}$$

This estimate implies that ϕ converges to some limit l as x_3 goes to $+\infty$, uniformly with respect to (x_1, x_2). We then define $\tilde{\phi} = \phi - l$, which satisfies (3.15) and

$$\forall x \in H^c, \quad \left|\tilde{\phi}(x)\right| \leq \frac{C}{(1+x_3)^{1/2}}.$$

We then repeat exactly the same argument as above, but with $\tilde{\phi}$ instead of ϕ, and we finally have (3.8), which implies (3.9). The fact that $l \leq 0$ is a direct consequence of (3.5). □

LEMMA 3.4. *Let m be a smooth function satisfying* (H0), (H1), *and* (H2). *If $(u, \phi) \in \left(L^\infty(\mathbb{R}^3)\right)^2$ is a solution of* (1.1), *then there exist two constants $A > 0$ and $a > 0$ such that*

$$\forall x \in H^c, \quad u(x) \geq Ae^{-a|x_3|} \tag{3.16}$$

$$\forall x \in H, \quad u(x) \geq A. \tag{3.17}$$

Proof. The estimate (3.17) may be proved with exactly the same method as for (2.3). We refer to [7] for the proof. Turning to (3.16), we use the fact that $\phi \in L^\infty(\mathbb{R}^3)$, so that $-\Delta u + u^{\frac{7}{3}} = \phi u \geq -bu$, for some constant $b > 0$. Since u is also bounded, this implies that there exists some $c > 0$ such that

$$-\Delta u + cu \geq 0.$$

Hence, convoluting the above inequality with the Yukawa potential $\frac{e^{-\sqrt{c}|x|}}{|x|}$, we get that

$$4\pi u(x) \geq \frac{c}{2} u * \frac{e^{-\sqrt{c}|x|}}{|x|} \geq \frac{Ac}{2} \int_H \frac{e^{-\sqrt{c}|x-y|}}{|x-y|} dy \geq Be^{-\sqrt{c}x_3},$$

for some constant $B > 0$. □

3.2. Proof of Theorem 1.1. In this section, we conclude the proof of Theorem 1.1. We assume that (u,ϕ) and (v,ψ) are two solutions of (1.1), and use the same notation as in section 2, namely $w = u - v$ and $\eta = \phi - \psi$. We recall that, for simplicity, we have assumed that m is smooth, so that (2.1) and (2.2) hold in the present case. We now carry out the same calculation as in section 2, multiplying the first line of (2.4) by $w\xi^2$ and the second line by $\eta\xi^2$, where $\xi \in \mathcal{D}(\mathbb{R}^3)$, finding (2.5) and (2.6). Here again, the operators $L_u = -\Delta + u^{\frac{4}{3}} - \phi$ and $L_v = -\Delta + v^{\frac{4}{3}} - \psi$ are nonnegative, and (2.7) is still valid, so that (2.8) translates into

$$\frac{1}{2}\int_{\mathbb{R}^3}\left(u^{\frac{4}{3}} + v^{\frac{4}{3}}\right)w^2\xi^2 + \frac{1}{4}\int_{\mathbb{R}^3}|\nabla\eta|^2\xi^2 \leq \int_{\mathbb{R}^3}(w^2+\eta^2)|\nabla\xi|^2. \tag{3.18}$$

We then use for ξ a sequence of functions converging to $\xi(x) = \left(1 - \frac{|x-x_0|}{2R}\right)_+$, for any $R > 0$ and any $x_0 \in \mathbb{R}^3$, and find that there exists some constant $C > 0$ independent of R and x_0 such that

$$\int_{B_R(x_0)}\left(u^{\frac{4}{3}} + v^{\frac{4}{3}}\right)w^2 \leq CR, \quad \int_{B_R(x_0)}|\nabla\eta|^2 \leq CR. \tag{3.19}$$

Moreover, (2.11) is still valid, and allows to prove, using the same computations as in section 2, that for some constant C independent of x_0 and R,

$$\int_{B_R(x_0)}(u+v)\eta^2 \leq CR. \tag{3.20}$$

Next, we apply Lemma 3.3 to ϕ and ψ, and infer that η satisfies (3.9) for some $l \in \mathbb{R}$. Hence, for $t > 0$ large enough, we have $\eta^2(x_1,x_2,t) \geq \frac{l^2}{2}$, for any $(x_1,x_2) \in \mathbb{R}^2$. Applying (3.20) and (3.16), we have, for all $t > 0$ large enough,

$$CR \geq \int_{D_R(t)}(u+v)\eta^2 \geq 2Ae^{-at}R^2\frac{l^2}{2},$$

where $D_R(t) = \left\{(x_1,x_2,x_3) \in \mathbb{R}^3,\ \sqrt{x_1^2+x_2^2} \leq R,\ t \leq x_3 \leq t+1\right\}$. We then let R go to infinity and find that $l = 0$, and thus

$$\forall x \in H^c, \quad |\eta(x)| \leq \frac{C}{(1+x_3)^{3/4}}, \tag{3.21}$$

for some constant C. We then go back to (3.18), and using here again the same kind of function ξ, we have, using estimates (3.1), (3.4), and (3.21),

$$\int_{\mathbb{R}^3} w^2|\nabla\xi|^2 \leq \left(\int_{-\infty}^0 A^2e^{-2\alpha|x_3|}dx_3 + \int_0^{+\infty}\frac{dx_3}{(1+x_3)^3}\right)\frac{1}{R^2}\int_{x_1^2+x_2^2\leq R^2}dx_1dx_2 \leq C,$$

and

$$\int_{\mathbb{R}^3} \eta^2|\nabla\xi|^2 \leq \left(\int_{-\infty}^0 A^2e^{-2\alpha|x_3|}dx_3 + \int_0^{+\infty}\frac{dx_3}{(1+x_3)^{3/2}}\right)\frac{1}{R^2}\int_{x_1^2+x_2^2\leq R^2}dx_1dx_2 \leq C,$$

for some constant $C > 0$. Hence, using here again the same argument as the one leading to (3.19) and (3.20), we find that

$$\int_{\mathbb{R}^3}\left(u^{\frac{4}{3}} + v^{\frac{4}{3}}\right)w^2 < +\infty, \quad \int_{\mathbb{R}^3}|\nabla\eta|^2 < +\infty, \quad \int_{\mathbb{R}^3}(u+v)\eta^2 < +\infty. \tag{3.22}$$

We then use $\xi(x_1,x_2,x_3) = \left(1-\frac{x_1}{R}\right)_+ \left(1-\frac{x_2}{R}\right)_+$, and compute, setting

$$E_R = \left\{(x_1,x_2,x_3)\in\mathbb{R}^3, \quad |x_1|\le R \quad \text{and} \quad |x_2|\le R\right\},$$

and using (3.18) again,

$$\begin{aligned}
\frac{1}{8}\int_{E_{\frac{R}{2}}} \left(u^{\frac{4}{3}}+v^{\frac{4}{3}}\right) w^2 &\le \frac{1}{R^2}\int_{E_R} \eta^2 + w^2 \\
&\le \frac{1}{R^2}\int_{E_R\cap\{|x_3|\le A\}} \eta^2+w^2 + \int_{E_R\cap\{|x_3|\ge A\}} \eta^2+w^2 \\
&\le \frac{Ce^{\gamma A}}{R^2}\left[\int_{\mathbb{R}^3}\left(u^{\frac{4}{3}}+v^{\frac{4}{3}}\right)w^2 + (u+v)\eta^2\right] + \frac{CR^2}{R^2A^{\frac{1}{2}}} \\
&\le C\left(\frac{e^{\gamma A}}{R^2}+\frac{1}{A^{\frac{1}{2}}}\right),
\end{aligned}$$

where A is any positive number, and γ and C are positive constants independent of R and A. We first let R go to infinity, and then A, which proves that $w=0$. This clearly implies that $\eta = 0$. □

4. Application: Thermodynamic limit for a semicrystal in the Thomas–Fermi–von Weizsäcker model. In this section we give a few remarks on the notion of thermodynamic limit, and on the link between this problem and the present work.

4.1. The Thomas–Fermi–von Weizsäcker model for molecules. In the Thomas–Fermi–von Weizsäcker (TFW) model, a molecule consisting of N point nuclei, which have positions $\{X_i\}_{1\le i\le N}$ and charges $\{Z_i\}_{1\le i\le N}$, together with M electrons, has its energy modelled by the functional

$$\begin{aligned}
\mathcal{E}^{\mathrm{TFW}}(\rho,\{X_i\}) = \int_{\mathbb{R}^3} a|\nabla\sqrt{\rho}|^2 + b\rho^{\frac{5}{3}} - \sum_{i=1}^N \int_{\mathbb{R}^3} \frac{Z_i\rho(x)}{4\pi|x-X_i|}dx \\
+\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho(x)\rho(y)}{4\pi|x-y|}dxdy + \frac{1}{2}\sum_{i\ne j}\frac{Z_iZ_j}{4\pi|X_i-X_j|},
\end{aligned} \tag{4.1}$$

where a and b are two positive constants. The function $\rho\in L^1(\mathbb{R}^3)$ denotes the electronic density and is constrained by $\int\rho = M$. The energy is defined only for densities ρ such that $\rho\ge 0$ and $\sqrt{\rho}\in H^1(\mathbb{R}^3)$, and the minimization problem corresponding to the determination of the ground state reads

$$\mathcal{E}^{\mathrm{TFW}}(\{X_i\}) = \inf\left\{\mathcal{E}^{\mathrm{TFW}}(\rho,\{X_i\}),\ \rho\ge 0,\ \sqrt{\rho}\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\rho = M\right\}. \tag{4.2}$$

In the special case of $a=1$ and $b=\frac{3}{5}$, and if we set

$$m = \sum_{i=1}^N Z_i\delta_{X_i},$$

$u=\sqrt{\rho}$, and $\phi = (\rho-m)*\frac{1}{|x|}-\theta$, then the Euler–Lagrange equation of (4.2) is (1.1). The constant θ in the definition of ϕ is the Lagrange multiplier associated to the mass constraint in (4.2).

In the special case $a=0$, the model reduces to the well-known Thomas–Fermi model [9].

4.2. Thermodynamic limit for a semicrystal. The thermodynamic limit is a way of deriving a solid-state model from a molecular one. Given an infinite number of atoms, one approximates it by a finite system, then lets the number of particles go to infinity in a suitable way and passes to the limit. We refer to [7] and [9] for a general presentation and the definition of the ground state energy of crystals (infinite periodic lattices) in the setting of Thomas–Fermi type theories.

The setting we consider for the finite size system is (4.1)–(4.2), and we will assume that $Z_i = 1$ for all $i \in \{1, 2, \dots, n\}$, although this can be generalized. The positions of the nuclei are assumed to be a subset of a periodic lattice intersected with H. For the sake of simplicity (this can be easily generalized to other periodic lattices, or to nonperiodic geometry, as in [5]) we assume that the set of nuclei is

$$\{X_i\}_{i\in\mathbb{N}} = \mathbb{Z}^3 \cap H,$$

and is approached by the truncated sets

$$\Lambda_R = \mathbb{Z}^3 \cap H \cap B_R,$$

for $R > 0$, where B_R is the ball of radius R centered at zero. Note that other types of truncation are also possible, as pointed out in [7]. We then define, according to (4.1)–(4.2), the number

$$I_R = \mathcal{E}^{\mathrm{TFW}}(\Lambda_R),$$

and denote by ρ_R the corresponding solution. We address the following questions:

(L1) *Does the energy per unit volume* $\frac{I_R}{|B_R|}$ *converge as* R *goes to infinity?*

(L2) *Does the minimizer* ρ_R *of* I_R *converge to a limit* ρ_∞ *as* R *goes to infinity?*

(L3) *Is the limit* ρ_∞ *the solution of some minimization problem of the form* (4.2)*?*

The method developed in [7] to deal with such a problem may be summarized as follows:

1. First, derive bounds on the density ρ_R,
2. next, using these bounds, pass to the limit in the Euler–Lagrange equation of I_R,
3. show a uniqueness theorem on the limit equation,
4. use this uniqueness property to prove a uniform convergence of the solution, which implies convergence of the energy.

Let us point out here that in the above strategy, step 3 is by far the most difficult.

Now, looking at the proofs given in [7] for the case of a full crystal, it is clear that the proofs of the first point carry through the present case. The next step is also clearly straightforward. Turning to step 3, we point out that the Euler–Lagrange equation of problem I_R can be rewritten as

$$\begin{cases} -\Delta u_R + u_R^{7/3} - \phi_R u_R = 0, \\ -\Delta \phi_R = m_R - u_R^2, \end{cases}$$

with $u_R = \sqrt{\rho}_R$ and $\phi_R = (\rho_R - m_R) * \frac{1}{|x|} - \theta_R$ (the number θ_R is the Lagrange multiplier), and

$$m_R = \sum_{k\in\Lambda_R} \delta_k.$$

Passing to the limit in this system, one gets (1.1) with

$$m = \sum_{k \in \mathbb{Z}^3 \cap H} \delta_k, \tag{4.3}$$

which clearly satisfies (H0), (H1), and (H2). Hence, one can apply Theorem 1.1, proving that the limit of the solution is unique. The thermodynamic limit problem can thus be dealt with in the present case.

Our final point is to write down a minimization problem for the limit density: since the solution of (1.1) with m defined by (4.3) is unique, it is periodic of period one in x_1 and x_2. Moreover, the exponential decay proved in Lemma 3.1 implies that $u - u_{\rm per}$ decays exponentially as x_3 goes to $-\infty$, where $u_{\rm per}$ is the solution of (1.1) with

$$m_{\rm per} = \sum_{k \in \mathbb{Z}^3} \delta_k.$$

The same decay is true for ∇u. In order to write down the limit minimization problem, we first define a periodic Coulomb potential (see [4, 6] for more details):

$$G(x) = -2\pi|x_3| + \sum_{k \in \mathbb{Z}^2 \times \{0\}} \left(\frac{1}{|x-k|} - \int_{K \times \{0\}} \frac{dy}{|x-y-k|} \right), \tag{4.4}$$

where K is the unit square of $\mathbb{R}^2$. We assume that χ is a cut-of function depending only on x_3 such that

$$\begin{aligned} &\chi(x) = 1 \quad \text{if} \quad x_3 < -1, \\ &\chi(x) = 0 \quad \text{if} \quad x_3 > 1, \\ &\int_Q \left[(\chi(x) - \mathbf{1}_{x_3<0}) u_{\rm per}^2(x) \right] dx = 0, \end{aligned}$$

where Q is the "unit cell"

$$Q = K \times \mathbb{R} = \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \quad / \quad x_1 \in \left[-\frac{1}{2}, \frac{1}{2} \right), \quad x_1 \in \left[-\frac{1}{2}, \frac{1}{2} \right) \right\},$$

and we define the following energy, with $\rho_0 = \chi u_{\rm per}^2$,

$$\begin{aligned} \mathcal{E}(\tilde\rho) = &\int_Q (\tilde\rho + \rho_0)^{5/3} - \rho_0^{5/3} + \int_Q |\nabla \sqrt{\tilde\rho + \rho_0}|^2 - |\nabla \sqrt{\rho_0}|^2 - \int_Q \tilde\rho G \\ &+ \frac{1}{2} \int_Q \int_Q \tilde\rho(x) G(x-y) \tilde\rho(y) dx dy \\ &+ \int_Q \int_Q \tilde\rho(x) G(x-y) \rho_0(y) dx dy, \end{aligned} \tag{4.5}$$

where we have set formally $\tilde\rho = \rho - \rho_0$, and have subtracted the infinite energy of ρ_0 in order to get a finite value of the energy. This allows one to define the variational problem as follows:

$$\inf \left\{ \mathcal{E}(\tilde\rho), \quad \tilde\rho \in H^1_{\rm per}(Q), \quad \tilde\rho + \rho_0 \geq 0, \quad \sqrt{\tilde\rho + \rho_0} \in H^1_{\rm loc}(\mathbb{R}^3), \quad \int_Q \tilde\rho = 0 \right\}, \tag{4.6}$$

where $H^1_{\rm per}(Q)$ is the set of functions in $H^1(Q)$, which satisfy periodic boundary conditions in x_1 and x_2.

One can then check that the minimization problem (4.6) is well posed, that its solution does not depend on χ, and that its Euler–Lagrange equation is (1.1) with m defined by (4.3), where $u = \sqrt{\tilde{\rho} + \rho_0}$ and ϕ is the effective potential.

REFERENCES

[1] R. Benguria and E. H. Lieb, *The most negative ion in the Thomas–Fermi–von-Weizsäcker theory of atoms and molecules.*, J. Phys. B, 18 (1985), pp. 1045–1059.

[2] R. Benguria, H. Brézis, and E. H. Lieb, *The Thomas–Fermi–von Weizsäcker theory of atoms and molecules,* Comm. Math. Phys., 79 (1981), pp. 167–180.

[3] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction,* Grundlehren der Math. Wiss., 223, Springer-Verlag, Berlin-New York, 1976.

[4] X. Blanc and C. Le Bris, *Thomas–Fermi type theories for polymers and thin films*, Adv. Differential Equations, 5 (2000), pp. 977–1032.

[5] X. Blanc, C. Le Bris, and P.-L. Lions, *A definition of the ground state energy for systems composed of infinitely many particles*, Comm. Partial Differential Equations, 28 (2003), pp. 439–475.

[6] X. Blanc and R. Monneau, *Screening of an applied electric field inside a metallic layer described by the Thomas–Fermi-von Weizsäcker model*, Adv. Differential Equations, 7 (2002), pp. 847–876.

[7] I. Catto, C. Le Bris, and P.-L. Lions, *Mathematical Theory of Thermodynamic Limits: Thomas–Fermi Type Models,* Clarendon Press, Oxford University Press, New York, 1998.

[8] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1998.

[9] E. Lieb and B. Simon, *The Thomas–Fermi theory of atoms, molecules and solids*, Adv. Math., 23 (1977), pp. 22–116.

[10] L. Nirenberg, *On elliptic partial differential equations*, Ann. Sc. Norm. Sup. Pisa (3), 13 (1959), pp. 115–162.

[11] J. P. Solovej, *Universality in the Thomas–Fermi–von Weizsäcker model of atoms and molecules*, Comm. Math. Phys., 129 (1990), pp. 561–598.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1249–1261

AN ASYMPTOTIC FORMULA FOR THE DISPLACEMENT FIELD IN THE PRESENCE OF THIN ELASTIC INHOMOGENEITIES*

ELENA BERETTA† AND ELISA FRANCINI‡

Abstract. We consider a plane isotropic homogeneous elastic body with thin elastic inhomogeneities in the form of small neighborhoods of simple smooth curves. We derive a rigorous asymptotic expansion of the boundary displacement field as the thickness of the neighborhoods goes to zero.

Key words. Lamé system, thin inhomogeneities, asymptotic formulas

AMS subject classifications. 35C20, 35R05

DOI. 10.1137/050648596

1. Introduction.

Let $\Omega \subset \mathbb{R}^2$ be a bounded smooth domain representing the region occupied by an elastic material.

Let $\sigma_0 \subset \Omega$ be a simple smooth curve and define, for a positive small ϵ, the set

$$\omega_\epsilon = \{x \in \Omega \ : \ d(x, \sigma_0) < \epsilon\},$$

which represents an inclusion of small size made of a different elastic material.

Let $\mathbb{C}_0$ and $\mathbb{C}_1$ be the elastic tensor fields in $\Omega \setminus \overline{\omega}_\epsilon$ and ω_ϵ, respectively.

Given a traction field g on $\partial\Omega$, the displacement field u_ϵ, generated by this traction in the body containing the inclusion ω_ϵ, solves the following system of linearized elasticity:

$$\begin{cases} \operatorname{div}(\mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon) &= 0 \quad \text{in} \quad \Omega, \\ (\mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon) \cdot \nu &= g \quad \text{on} \quad \partial\Omega, \end{cases} \tag{1}$$

where $\mathbb{C}_\epsilon = \mathbb{C}_0 \chi_{\Omega \setminus \omega_\epsilon} + \mathbb{C}_1 \chi_{\omega_\epsilon}$, $\widehat{\nabla} u_\epsilon = \frac{1}{2}\left(\nabla u_\epsilon + (\nabla u_\epsilon)^T\right)$ is the symmetric deformation tensor and ν denotes the outward unit normal to $\partial\Omega$.

Let us also introduce the background displacement u_0, namely the solution of

$$\begin{cases} \operatorname{div}(\mathbb{C}_0 \widehat{\nabla} u_0) &= 0 \quad \text{in} \quad \Omega, \\ (\mathbb{C}_0 \widehat{\nabla} u_0) \cdot \nu &= g \quad \text{on} \quad \partial\Omega. \end{cases} \tag{2}$$

The goal of this paper is to find an asymptotic expansion for $(u_\epsilon - u_0)_{|\partial\Omega}$ as $\epsilon \to 0$. An analogous expansion has been derived in [BFV] for the case of thin conductivity inclusions. These expansions represent a powerful tool to solve the inverse problem of identifying the inclusions, given boundary measurements (see, [ABF] and [ABF2] for the case of thin conductivity inclusions and [AK] for further references).

*Received by the editors December 28, 2005; accepted for publication (in revised form) June 15, 2006; published electronically December 15, 2006. This work is partly supported by MIUR under grant 2004011204.

http://www.siam.org/journals/sima/38-4/64859.html

†Dipartimento di Matematica "G. Castelnuovo," Università di Roma "La Sapienza," P.le Aldo Moro, 2 - 00185 Rome, Italy (beretta@mat.uniroma1.it)

‡Dipartimento di Matematica "U. Dini," Università degli Studi di Firenze, Viale Morgagni, 67A - 50134 Firenze, Italy (francini@math.unifi.it)

In [AKNT] the authors derive an asymptotic expansion for the boundary displacement field $(u_\epsilon - u_0)_{|\partial\Omega}$ in the case of diametrically small elastic inclusions, namely inclusions of the form $z + \epsilon B$, where z is a point in Ω and B is a bounded domain containing the origin. The approach they use, based on the method of layer potentials (see [AK]), allows them to find a very accurate expansion. Unfortunately, this method does not seem to work in the case of thin elastic inclusions. Hence, in order to derive the expansion in our context, we apply similar arguments as in [BFV] for conductivity inclusions; more precisely, we use a variational approach and fine regularity estimates for solutions of elliptic systems with discontinuous coefficients obtained by Li and Nirenberg in [LN]. The main difficulty arising in the framework of linear elasticity, compared to the conductivity case, consists of finding an explicit representation formula for the tensor appearing in the first order term of the asymptotic expansion.

The plan of the paper is as follows: in section 2 we introduce some notation and state the main result. Section 3 is devoted to the derivation of some estimates and properties of the displacement field, and in section 4 we prove our main result.

2. The main result. Let us introduce some notation and assumptions that will be useful in what follows.

(a) We will assume that σ_0 is of class C^3 and that there exists some $K > 0$ such that

$$\begin{aligned} d(\sigma_0, \partial\Omega) &\geq K^{-1}, \\ \|\sigma_0\|_{C^3} &\leq K, \\ K^{-1} \leq \text{length}(\sigma_0) &\leq K. \end{aligned} \tag{3}$$

Moreover, we assume that for every $x \in \sigma_0$ there are two discs, B_1 and B_2, of radius K^{-1} such that

$$\overline{B}_1 \cap \overline{B}_2 = \overline{B}_1 \cap \sigma_0 = \overline{B}_2 \cap \sigma_0 = \{x\}.$$

The latter assumption guarantees that different parts of σ_0 do not get too close, so that ω_ϵ does not self-intersect for small ϵ.

(b) Ω and ω_ϵ are both homogeneous and isotropic, i.e., the elastic tensor fields $\mathbb{C}_0$ and $\mathbb{C}_1$ are of the following form:

$$(\mathbb{C}_m)_{ijlk} = \lambda_m \delta_{ij}\delta_{kl} + \mu_m(\delta_{ki}\delta_{lj} + \delta_{kj}\delta_{li}), \text{ for } i,j,k,l = 1,2, \;\; m = 0,1, \tag{4}$$

where (λ_0, μ_0) and (λ_1, μ_1) are the Lamè coefficients corresponding to $\Omega \setminus \overline{\omega}_\epsilon$ and ω_ϵ, respectively, and $(\lambda_0 - \lambda_1)^2 + (\mu_0 - \mu_1)^2 \neq 0$.

(c) There are two positive constants α_0 and β_0 such that

$$\min(\mu_0, \mu_1) \geq \alpha_0, \quad \min(2\lambda_0 + 2\mu_0, 2\lambda_1 + 2\mu_1) \geq \beta_0. \tag{5}$$

We note that the last conditions ensure that $\mathbb{C}_\epsilon$ is strongly convex in Ω, i.e., if we set $\xi_0 = \min(2\alpha_0, \beta_0)$, then

$$\mathbb{C}_\epsilon A \cdot A \geq \xi_0 |A|^2,$$

for any symmetric 2×2 matrix A, where $A \cdot B = \sum_{ij} a_{ij} b_{ij}$ and $|A|^2 = A \cdot A$.

(d) We shall prescribe a traction field $g \in H^{-1/2}(\partial\Omega, \mathbb{R}^2)$ satisfying the compatibility condition

$$\int_{\partial\Omega} g \cdot R = 0, \tag{6}$$

for every infinitesimal rigid displacement R, that is $R(x) = c + Wx$, where c is any constant vector in $\mathbb{R}^2$ and W any constant 2×2 skew matrix.

Under our assumptions there exist weak solutions u_ϵ and $u_0 \in H^1(\Omega, \mathbb{R}^2)$ to the problems (1) and (2), respectively (see, for example, [V] or [F]). Concerning uniqueness we recall that solutions of the above problems are uniquely determined up to infinitesimal rigid displacements. Hence, in order to uniquely identify such solutions, we assume that u_ϵ and u_0 satisfy the normalization conditions

$$\int_{\partial\Omega} u = 0, \quad \int_\Omega \nabla u - (\nabla u)^T = 0. \tag{7}$$

It is easy to see that if u_ϵ is solution of (1), then it solves the Lamé system

$$\begin{cases} \mu_0 \Delta u_\epsilon + (\lambda_0 + \mu_0) \nabla(\operatorname{div} u_\epsilon) &= 0 \quad \text{in} \quad \Omega \setminus \overline{\omega}_\epsilon, \\ \mu_1 \Delta u_\epsilon + (\lambda_1 + \mu_1) \nabla(\operatorname{div} u_\epsilon) &= 0 \quad \text{in} \quad \omega_\epsilon, \\ u_\epsilon^i &= u_\epsilon^e \quad \text{on} \quad \partial\omega_\epsilon, \\ \left(\mathbb{C}_0 \widehat{\nabla} u_\epsilon^i\right) \nu &= \left(\mathbb{C}_1 \widehat{\nabla} u_\epsilon^e\right) \nu \quad \text{on} \quad \partial\omega_\epsilon, \end{cases} \tag{8}$$

where ν is the outward unit normal to $\partial\omega_\epsilon$, and, for $x \in \partial\omega_\epsilon$,

$$u_\epsilon^e(x) = \lim_{\substack{y \to x \\ y \in \Omega \setminus \overline{\omega}_\epsilon}} u_\epsilon(y), \quad u_\epsilon^i(x) = \lim_{\substack{y \to x \\ y \in \omega_\epsilon}} u_\epsilon(y),$$

and

$$\widehat{\nabla} u_\epsilon^e(x) = \lim_{\substack{y \to x \\ y \in \Omega \setminus \overline{\omega}_\epsilon}} \widehat{\nabla} u_\epsilon(y), \quad \widehat{\nabla} u_\epsilon^i(x) = \lim_{\substack{y \to x \\ y \in \omega_\epsilon}} \widehat{\nabla} u_\epsilon(y).$$

For $y \in \Omega$, we will denote by $N(\cdot, y)$ the Neumann function related to Ω, i.e., the weak solution to the problem

$$\begin{cases} \operatorname{div}\left(\mathbb{C}_0 \widehat{\nabla} N(\cdot, y)\right) &= -\delta_y \, \mathrm{I}_d \quad \text{in} \quad \Omega, \\ \left(\mathbb{C}_0 \widehat{\nabla} N(\cdot, y)\right) \cdot \nu &= -\frac{1}{|\partial\Omega|} \, \mathrm{I}_d \quad \text{on} \quad \partial\Omega, \end{cases} \tag{9}$$

with the normalization conditions (7) and where I_d is the identity matrix in $\mathbb{R}^2$.

Note that $N(x, y)$ is regular for $x \neq y$ and, at $x = y$, has the same singularities of $\Gamma(x - y)$, where $\Gamma = (\Gamma_{ij})_{i,j=1}^2$ is the fundamental solution in the free space of the system

$$\operatorname{div}\left(\mathbb{C}_0 \widehat{\nabla} \cdot\right) = 0 \quad \text{in} \quad \mathbb{R}^2,$$

and is given by

$$\Gamma_{ij}(x) = \frac{A}{2\pi} \, \delta_{ij} \ln |x| - \frac{B}{2\pi} \frac{x_i x_j}{|x|^2},$$

where $A = \frac{1}{2}\left(\frac{1}{\mu_0} + \frac{1}{\lambda_0 + 2\mu_0}\right)$ and $B = \frac{1}{2}\left(\frac{1}{\mu_0} - \frac{1}{\lambda_0 + 2\mu_0}\right)$.

Let us fix an orthonormal system (n, τ) on σ_0 such that n is a unit normal vector field to the curve and τ is a unit tangent vector field. If σ_0 is a closed curve, then we will take n to point in the outward direction of the domain it encloses. Let κ denote the curvature of σ_0 and, for $a, b \in \mathbb{R}^2$, let $a \otimes b$ denote the tensor product $a \otimes b = a_i b_j$.

We are now ready to state our main result.

THEOREM 2.1. *Let $\Omega \subset \mathbb{R}^2$ be a bounded smooth domain and let $\sigma_0 \subset\subset \Omega$ be a simple curve satisfying* (3). *Assume* (4), (5), *and* (6) *and let u_ϵ and u_0 be the solutions to* (1) *and* (2), *respectively, satisfying* (7). *For every $x \in \sigma_0$, there exists a fourth order symmetric tensor field $\mathcal{M}(x)$ such that, for $y \in \partial\Omega$ and $\epsilon \to 0$,*

$$(u_\epsilon - u_0)(y) = 2\epsilon \int_{\sigma_0} \mathcal{M}(x)\widehat{\nabla} u_0(x) \cdot \widehat{\nabla} N(x,y)\, d\sigma_0(x) + o(\epsilon). \tag{10}$$

The term $o(\epsilon)$ is bounded by $C\epsilon^{1+\theta}\|g\|_{H^{-1/2}(\partial\Omega)}$, with $0 < \theta < 1$ and C depending only on θ, Ω, α_0, β_0, and K.

Furthermore, on σ_0,

$$\begin{aligned}\mathcal{M}\widehat{\nabla} u_0 = {} & a \operatorname{div} u_0\, I_d + b\widehat{\nabla} u_0 \\ & + c\left(\frac{\partial(u_0 \cdot \tau)}{\partial \tau} + \kappa(u_0 \cdot n)\right)\tau \otimes \tau + d\frac{\partial(u_0 \cdot n)}{\partial n} n \otimes n,\end{aligned}$$

where

$$a = (\lambda_1 - \lambda_0)\frac{\lambda_0 + 2\mu_0}{\lambda_1 + 2\mu_1}, \quad b = 2(\mu_1 - \mu_0)\frac{\mu_0}{\mu_1}, \tag{11}$$

$$c = 2(\mu_1 - \mu_0)\left[\left(\frac{2\lambda_1 + 2\mu_1 - \lambda_0}{\lambda_1 + 2\mu_1} - \frac{\mu_0}{\mu_1}\right)\right], \tag{12}$$

and

$$d = 2(\mu_1 - \mu_0)\frac{\mu_1\lambda_0 - \mu_0\lambda_1}{\mu_1(\lambda_1 + 2\mu_1)}. \tag{13}$$

The proof of the theorem is contained in section 4.

3. Energy and a priori estimates. In this section we will show that, for $\epsilon \to 0$,

$$\|u_\epsilon - u_0\|_{H^1(\Omega)} = O(\epsilon^{1/2}). \tag{14}$$

In order to establish it we will need the following version of the Korn inequality.

LEMMA 3.1. *Let Ω be a Lipschitz connected open set in $\mathbb{R}^2$. Let $u \in H^1(\Omega, \mathbb{R}^2)$ and let $W_0 = \int_\Omega \frac{1}{2}\left(\nabla u - (\nabla u)^T\right)$.*

Then, there exists a constant C such that

$$\|\nabla u - W_0\|_{L^2(\Omega)} \le C\|\widehat{\nabla} u\|_{L^2(\Omega)}. \tag{15}$$

For the proof, see [T, section 3].

PROPOSITION 3.2. *Let u_ϵ and u_0 be solutions to* (1) *and* (2), *respectively. There exists a constant C depending on Ω, K, α_0, and β_0, such that*

$$\|u_\epsilon - u_0\|_{H^1(\Omega)} \le C\epsilon^{1/2}\|g\|_{H^{-1/2}(\partial\Omega)}. \tag{16}$$

Proof. Since $\int_{\partial\Omega}(u_\epsilon - u_0) = 0$, by the Poincaré inequality there exists a constant C, depending on Ω, such that

$$\int_\Omega |u_\epsilon - u_0|^2 \le C\int_\Omega |\nabla(u_\epsilon - u_0)|^2. \tag{17}$$

It thus suffices to estimate $\|\nabla(u_\epsilon - u_0)\|_{L^2(\Omega)}$.

By the strong convexity of $\mathbb{C}_\epsilon$ and the Korn inequality, in the form of Lemma 3.1, applied to $u_\epsilon - u_0$, and recalling that u_ϵ and u_0 satisfy (7), we get

$$\begin{aligned}\int_\Omega \mathbb{C}_\epsilon \widehat{\nabla}(u_\epsilon - u_0) \cdot \widehat{\nabla}(u_\epsilon - u_0) &\geq \xi_0 \int_\Omega |\widehat{\nabla}(u_\epsilon - u_0)|^2 \\ &\geq C \int_\Omega |\nabla(u_\epsilon - u_0)|^2, \end{aligned} \tag{18}$$

where C depends on α_0, β_0, and Ω.

Now, observe that

$$\int_\Omega \mathbb{C}_\epsilon \widehat{\nabla}(u_\epsilon - u_0) \cdot \widehat{\nabla}(u_\epsilon - u_0) = \int_{\omega_\epsilon} (\mathbb{C}_0 - \mathbb{C}_1)\widehat{\nabla} u_0 \cdot \widehat{\nabla}(u_\epsilon - u_0), \tag{19}$$

which follows by integration by parts and uses the fact that $(\mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon)\cdot \nu = (\mathbb{C}_0 \widehat{\nabla} u_0)\cdot \nu$ on $\partial\Omega$. Indeed

$$\begin{aligned} &\int_\Omega \mathbb{C}_\epsilon \widehat{\nabla}(u_\epsilon - u_0) \cdot \widehat{\nabla}(u_\epsilon - u_0) - \int_{\omega_\epsilon} (\mathbb{C}_0 - \mathbb{C}_1)\widehat{\nabla} u_0 \cdot \widehat{\nabla}(u_\epsilon - u_0) \\ &= \int_{\Omega\setminus\omega_\epsilon} \mathbb{C}_0 \widehat{\nabla}(u_\epsilon - u_0) \cdot \widehat{\nabla}(u_\epsilon - u_0) + \int_{\omega_\epsilon} (\mathbb{C}_1 \widehat{\nabla} u_\epsilon - \mathbb{C}_0 \widehat{\nabla} u_0) \cdot \widehat{\nabla}(u_\epsilon - u_0) \\ &= \int_\Omega \mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla}(u_\epsilon - u_0) - \int_\Omega \mathbb{C}_0 \widehat{\nabla} u_0 \cdot \widehat{\nabla}(u_\epsilon - u_0) \\ &= \int_{\partial\Omega} \Big(\Big(\mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon \Big) \cdot \nu - \Big(\mathbb{C}_0 \widehat{\nabla} u_0 \Big) \cdot \nu \Big) \cdot (u_\epsilon - u_0)\, d\sigma = 0, \end{aligned}$$

hence (19) holds.

On the other hand, by the Hölder inequality,

$$\begin{aligned} &\int_{\omega_\epsilon} (\mathbb{C}_0 - \mathbb{C}_1)\, \widehat{\nabla} u_0 \cdot \widehat{\nabla}(u_\epsilon - u_0)\, dx \\ &\leq \max\{2|\mu_0 - \mu_1|, |\lambda_0 - \lambda_1|\}\, \|\nabla u_0\|_{L^\infty(\omega_\epsilon)} |\omega_\epsilon|^{1/2} \|\nabla(u_\epsilon - u_0)\|_{L^2(\Omega)}. \end{aligned} \tag{20}$$

In order to bound $\|\nabla u_0\|_{L^\infty(\omega_\epsilon)}$ note that for small ϵ, say $\epsilon < K/2$, the distance between ω_ϵ and $\partial\Omega$ is bounded from below by $K/2$. Hence, by standard interior regularity estimates for elliptic systems (see [C]),

$$\|\nabla u_0\|_{L^\infty(\omega_\epsilon)} \leq C \|u_0\|_{H^1(\Omega)},$$

where C depends on α_0, β_0, and K.

By the divergence theorem, the trace theorem (see [LM]), and the Poincaré inequality,

$$\|u_0\|_{H^1(\Omega)} \leq C \|g\|_{H^{-1/2}(\partial\Omega)},$$

where C depends only on Ω. Finally,

$$\|\nabla u_0\|_{L^\infty(\omega_\epsilon)} \leq C \|g\|_{H^{-1/2}(\partial\Omega)}, \tag{21}$$

where C depends on Ω, α_0, β_0, and K.

So, by (18), (19), and (20) we obtain

$$\|\nabla(u_\epsilon - u_0)\|_{L^2(\Omega)} \leq C|\omega_\epsilon|^{1/2}\|g\|_{H^{-1/2}(\partial\Omega)},$$

where $C = C(\Omega, \alpha_0, \beta_0, K)$. By assumption (3), we can estimate

$$|\omega_\epsilon| \leq C\epsilon, \tag{22}$$

where C depends only on K. By putting together (21), (22), and the Poincarè inequality (17), we get (16). □

Besides the energy estimates (16), a key ingredient to establish the asymptotic expansion of Theorem 2.1 is a gradient estimate for elliptic systems modeling composite materials that has been established by Li and Nirenberg in [LN]. Here we state and use a simplified version of Proposition 5.1 in [LN].

Let D be the unit square $D = [-1, 1] \times [-1, 1]$, and let $f_0, \ldots, f_{l+1} \in C^2([-1, 1])$ such that

$$-1 = f_0(x_1) < f_1(x_1) < \cdots < f_{l+1}(x_1) = 1 \quad \text{for} \quad x_1 \in [-1, 1].$$

Let

$$D_m = \{x = (x_1, x_2) \in D : f_{m-1}(x_1) < x_2 < f_m(x_1)\} \quad \text{for} \quad 1 \leq m \leq l + 1.$$

We suppose that the origin does not belong to the graphs of the functions f_j, and we denote by m_0 the index for which

$$f_{m_0}(0) < 0 < f_{m_0+1}(0).$$

Let us also set $\frac{1}{2}D = [-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$.

Let $\mathbb{C}$ be a bounded symmetric Lamé tensor defined in D and such that $\mathbb{C}$ is constant in each D_m with corresponding Lamé coefficients λ_m and μ_m. Then the following estimate holds.

PROPOSITION 3.3. *Let $u \in H^1(D, \mathbb{R}^2)$ be a weak solution to*

$$\operatorname{div}\left(\mathbb{C}\widehat{\nabla}u\right) = 0 \quad \text{in} \quad D.$$

Then, for any $x \in \overline{D}_{m_0} \cap \frac{1}{2}D$,

$$|\nabla u(x) - \nabla u(0)| \leq C\|u\|_{L^2(D)}|x|^\alpha, \tag{23}$$

where $\alpha \in (0, 1/4)$ and C depends on α, l, λ_m, μ_m, and $\|f_m\|_{C^2([-1,1])}$, for $m = 1, \ldots, l + 1$.

For the proof of this result, see [LN, section 5].

4. Proof of Theorem 2.1. We divide the proof into several steps: in the first step we write $(u_\epsilon - u_0)_{|\partial\Omega}$ in terms of an integral over ω_ϵ of the product of $\widehat{\nabla}u_\epsilon^i$ and $\widehat{\nabla}N$.

In the second step, by using the estimate of Proposition 3.3, we reduce this integral to an integral over some part of $\partial\omega_\epsilon$. In the third part, we use the transmission conditions and introduce the tensor $\mathcal{M}$. The fourth part contains the conclusion of the proof.

First step.

We are going to show that, for $y \in \partial\Omega$,

$$(u_\epsilon - u_0)(y) = \int_{\omega_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y). \tag{24}$$

In order to get (24) we recall that, for $y \in \Omega$, the function $N(\cdot, y)$ satisfies

$$\int_\Omega \mathbb{C}_0 \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla} v = -v(y) + \frac{1}{|\partial\Omega|} \int_{\partial\Omega} v \quad \forall v \in H^1(\Omega).$$

By choosing $v = u_\epsilon - u_0$ and using the normalization (7) we get

$$\int_\Omega \mathbb{C}_0 \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla}(u_\epsilon - u_0) = -(u_\epsilon - u_0)(y). \tag{25}$$

Observe now that

$$\begin{aligned} \int_\Omega \mathbb{C}_0 \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla}(u_\epsilon - u_0) &= \int_{\omega_\epsilon} (\mathbb{C}_0 - \mathbb{C}_1)\, \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla} u_\epsilon \\ &+ \int_\Omega \mathbb{C}_\epsilon \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla} u_\epsilon - \int_\Omega \mathbb{C}_0 \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla} u_0. \end{aligned} \tag{26}$$

Since u_ϵ and u_0 are solutions to (1) and (2), respectively, we have

$$\begin{aligned} \int_\Omega \mathbb{C}_0 \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla} u_0 &= \int_{\partial\Omega} g \cdot N(\cdot, y) \\ &= \int_\Omega \mathbb{C}_\epsilon \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y), \end{aligned}$$

hence (26) becomes

$$\int_\Omega \mathbb{C}_0 \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla}(u_\epsilon - u_0) = \int_{\omega_\epsilon} (\mathbb{C}_0 - \mathbb{C}_1)\, \widehat{\nabla} N(\cdot, y) \cdot \widehat{\nabla} u_\epsilon,$$

and, by inserting this last relation into (25,) we get (24) for $y \in \Omega$.

Finally, since $u_\epsilon - u_0$ is continuous up to $\partial\Omega$, we get (24) for any $y \in \partial\Omega$.

Second step.

Let β be a constant $0 < \beta < 1$, and set

$$\omega'_\epsilon = \left\{ x + \mu n(x) \,:\, x \in \sigma_0,\, d(x, \partial\sigma_0) > \epsilon^\beta,\, \mu \in (-\epsilon, \epsilon) \right\}.$$

Notice that if σ_0 is a closed simple curve, then $\omega'_\epsilon = \omega_\epsilon$.

Let us write

$$\begin{aligned} \int_{\omega_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y) &= \int_{\omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y) \\ &+ \int_{\omega_\epsilon \setminus \omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y). \end{aligned} \tag{27}$$

Concerning the last term in (27)

$$\begin{aligned} \left| \int_{\omega_\epsilon \setminus \omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y) \right| &\le \left| \int_{\omega_\epsilon \setminus \omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla}(u_\epsilon - u_0) \cdot \widehat{\nabla} N(\cdot, y) \right| \\ &+ \left| \int_{\omega_\epsilon \setminus \omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_0 \cdot \widehat{\nabla} N(\cdot, y) \right|. \end{aligned} \tag{28}$$

In order to bound the first term on the right-hand side (RHS) of (28) we use the energy estimate (16) and the fact that, for $y \in \partial\Omega$, $\|\nabla N(\cdot, y)\|_{L^\infty(\omega_\epsilon)}$ is bounded uniformly in ϵ. Moreover, since

$$|\omega_\epsilon \setminus \omega'_\epsilon| \leq C\epsilon^{1+\beta},$$

we get

$$\left| \int_{\omega_\epsilon \setminus \omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla}(u_\epsilon - u_0) \cdot \widehat{\nabla} N(\cdot, y) \right| \leq C\epsilon^{1+\beta/2} \|g\|_{H^{-1/2}(\partial\Omega)},$$

where $C = C(\Omega, K, \alpha_0, \beta_0)$.

In the last term on the RHS of (28) we use the regularity estimates for u_0 so that

$$\begin{aligned} \left| \int_{\omega_\epsilon \setminus \omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_0 \cdot \widehat{\nabla} N(\cdot, y) \right| &\leq C \|\nabla u_0\|_{L^\infty(\omega_\epsilon)} \|\nabla N(\cdot, y)\|_{L^\infty(\omega_\epsilon)} \cdot |\omega_\epsilon \setminus \omega'_\epsilon| \\ &\leq C\epsilon^{1+\beta} \|g\|_{H^{-1/2}(\partial\Omega)}, \end{aligned}$$

where $C = C(K, \alpha_0, \beta_0)$ and so (27) becomes, for $\epsilon \to 0$,

$$\int_{\omega_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y) = \int_{\omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon \cdot \widehat{\nabla} N(\cdot, y) + O(\epsilon^{1+\beta/2}). \tag{29}$$

Now let us denote by σ'_μ, for $\mu \in [-\epsilon, \epsilon]$, the curve

$$\sigma'_\mu = \left\{ x + \mu n(x) \,:\, x \in \sigma_0,\ d(x, \partial\sigma_0) > \epsilon^\beta \right\}.$$

For every point $x + \mu n(x) \in \omega'_\epsilon$ (for $\mu \in (-\epsilon, \epsilon)$), let us consider the point $x + \epsilon n(x) \in \sigma'_\epsilon$ and let us compare $\nabla u_\epsilon(x + \mu n(x))$ with $\nabla u^i_\epsilon(x + \epsilon n(x))$.

More precisely, we will establish that, for $\alpha \in (0, 1/4)$,

$$|\nabla u_\epsilon(x + \mu n(x)) - \nabla u^i_\epsilon(x + \epsilon n(x))| \leq C\epsilon^{-\beta(2+\alpha)} \epsilon^\alpha \|g\|_{H^{-1/2}(\partial\Omega)}, \tag{30}$$

where $C = C(K, \alpha_0, \beta_0, \alpha)$.

Let ϵ be small enough to have

$$2\epsilon < \frac{\epsilon^\beta}{2\sqrt{2}}. \tag{31}$$

We note that the distance between $x + \mu n(x)$ and $x + \epsilon n(x)$ is smaller than 2ϵ and that, in a neighborhood of $x + \mu n(x)$ of radius ϵ^β, the boundary $\partial\omega_\epsilon$ is represented by graphs of smooth functions. In particular, if we set the origin to $x + \mu n(x)$, up to a rotation of the coordinate system (z_1, z_2), we have that there exist two functions g_1 and g_2 such that, for $-\frac{\epsilon^\beta}{\sqrt{2}} < z_1 < \frac{\epsilon^\beta}{\sqrt{2}}$,

$$\mathbb{C}_\epsilon(z_1, z_2) = \begin{cases} \mathbb{C}_0 & \text{if} \quad -\frac{\epsilon^\beta}{\sqrt{2}} < z_2 < g_1(z_1), \\ \mathbb{C}_1 & \text{if} \quad g_1(z_1) < z_2 < g_2(z_1), \\ \mathbb{C}_0 & \text{if} \quad g_2(z_1) < z_2 < \frac{\epsilon^\beta}{\sqrt{2}}. \end{cases}$$

By the a priori assumptions on σ_0 we know that

$$\|g_1\|_{C^2}, \quad \|g_2\|_{C^2} \leq K.$$

Consider now the function

$$v(y) = u_\epsilon\left(\frac{\epsilon^\beta}{\sqrt{2}}y\right),$$

which is defined in $[-1,1]\times[-1,1]$. Notice that v solves

$$\operatorname{div}\left(\tilde{\mathbb{C}}\widehat{\nabla}v\right) = 0 \quad \text{in} \quad (-1,1)\times(-1,1),$$

where

$$\tilde{\mathbb{C}}(y_1,y_2) = \begin{cases} \mathbb{C}_0 & \text{if} \quad -1 < y_2 < f_1(y_1) = \sqrt{2}\epsilon^{-\beta} g_1(\frac{\epsilon^\beta}{\sqrt{2}}y_1), \\ \mathbb{C}_1 & \text{if} \quad f_1(y_1) < y_2 < f_2(y_1) = \sqrt{2}\epsilon^{-\beta} g_2(\frac{\epsilon^\beta}{\sqrt{2}}y_1), \\ \mathbb{C}_0 & \text{if} \quad f_2(y_1) < y_2 < 1. \end{cases}$$

Let us check that we can apply Proposition 3.3, with $l = 2$ and $m_0 = 1$. Since $|g_i(y_1)| < \frac{\epsilon^\beta}{\sqrt{2}}$ and the derivative of g_i is bounded by K for $i = 1,2$, then

$$\|g_i\|_{L^\infty} \le (K+1)\frac{\epsilon^\beta}{\sqrt{2}}, \quad i = 1,2.$$

From this last estimate, and from the C^2 bounds on g_1 and g_2, we get

$$\|f_i\|_{C^2([-1,1])} \le 2K+1 \quad \text{for} \quad i = 1,2.$$

Since we set the origin at the point $x + \mu n(x)$, we have that

$$|x+\epsilon n(x)| = |x+\mu n(x) + (\epsilon-\mu)n(x)| = |(\epsilon-\mu)n(x)| \le 2\epsilon.$$

Hence, if we set $\overline{y} = \sqrt{2}\epsilon^{-\beta}(x+\epsilon n(x))$ we have, by (31),

$$|\overline{y}| = \sqrt{2}\epsilon^{-\beta}|x+\epsilon n(x)| \le \sqrt{2}\epsilon^{-\beta}\cdot 2\epsilon \le \frac{1}{2},$$

and, $\overline{y} \in \overline{D}_{m_0} \cap \frac{1}{2}D$. By Proposition 3.3,

$$|\nabla v(\overline{y}) - \nabla v(0)| \le C\|v\|_{L^2(D)}|\overline{y}|^\alpha,$$

where C depends only on K, α_0, β_0, and $\alpha \in (0,1/4)$. If we read this estimate for the function u_ϵ we get

$$\begin{aligned} \left|\nabla u_\epsilon(x+\mu n(x)) - \nabla u_\epsilon^i(x+\epsilon n(x))\right| &\le C\|u_\epsilon\|_{L^2(\Omega)}\epsilon^{-\beta(2+\alpha)}|(\mu-\epsilon)n(x)|^\alpha \\ &\le C\|u_\epsilon\|_{L^2(\Omega)}\epsilon^{-\beta(2+\alpha)}\epsilon^\alpha. \end{aligned}$$

Since

$$\|u_\epsilon\|_{L^2(\Omega)} \le C\|\nabla u_\epsilon\|_{L^2(\Omega)} \le C\|g\|_{H^{-1/2}(\partial\Omega)}$$

we finally have (30).

Due to (30), we can approximate the values of ∇u_ϵ in ${\omega_\epsilon}'$ with the values on σ_ϵ'.

Let us denote by ${\sigma_0}' = \{x \in \sigma_0 : d(x,\partial\sigma_0) > \epsilon^\beta\}$. Due to the regularity assumption on σ_0,

$$d\sigma_x^\mu = (1+O(\epsilon))d\sigma_x^0,$$

where $d\sigma_x^\mu$ and $d\sigma_x^0$ denote the infinitesimal arclengths on σ'_μ and σ'_0, respectively.

Hence,

$$\begin{aligned}
&\int_{\omega'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon(x) \cdot \widehat{\nabla} N(x,y)\, dx \\
&= \int_{-\epsilon}^{\epsilon} \int_{\sigma'_\mu} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon(x) \cdot \widehat{\nabla} N(x,y)\, d\sigma_x^\mu\, d\mu \\
&= \int_{-\epsilon}^{\epsilon} \int_{\sigma_0{}'} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon(x+\mu n(x)) \cdot \widehat{\nabla} N(x+\mu n(x), y)\, d\sigma_x^0\, d\mu + O(\epsilon^2) \\
&= \int_{-\epsilon}^{\epsilon} \int_{\sigma'_0} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon(x+\epsilon n(x)) \cdot \widehat{\nabla} N(x+\epsilon n(x), y)\, d\sigma_x^0\, d\mu \\
&\quad + O(\epsilon^{(1-\beta)(1+\alpha)}) \\
&= 2\epsilon \int_{\sigma'_\epsilon} (\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon^i \cdot \widehat{\nabla} N(\cdot, y) + O(\epsilon^{1+\alpha-\beta(2+\alpha)}), \qquad (32)
\end{aligned}$$

for $\beta < \alpha(2+\alpha)^{-1}$.

Third step.

Let us now extend the fields n and τ from σ_0 to ω'_ϵ. For $x \in \sigma'_0$ we set n and τ equal to $n(x)$ and $\tau(x)$ all along the line segment $x + \mu n(x)$, for $\mu \in [-\epsilon, \epsilon]$.

We will show, by using the transmission condition (8), that on σ'_ϵ,

$$(\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon^i = \mathcal{M}_\epsilon \widehat{\nabla} u_\epsilon^e, \qquad (33)$$

where

$$\begin{aligned}
\mathcal{M}_\epsilon \widehat{\nabla} u_\epsilon^e = a \operatorname{div} u_\epsilon^e\, \mathrm{I}_d + b \widehat{\nabla} u_\epsilon^e + c \left(\frac{\partial (u_\epsilon^e \cdot \tau)}{\partial \tau} + \kappa_\epsilon (u_\epsilon^e \cdot n) \right) \tau \otimes \tau \\
+ d \frac{\partial (u_\epsilon^e \cdot n)}{\partial n}\, n \otimes n,
\end{aligned}$$

with a, b, c, and d given by (11), (12), and (13), and κ_ϵ being the curvature of σ'_ϵ.

Let us express the transmission conditions (8) and $(\mathbb{C}_1 - \mathbb{C}_0)\, \widehat{\nabla} u_\epsilon^i$ in the n, τ coordinate system, namely

$$\begin{aligned}
&\frac{\partial (u_\epsilon^i \cdot \tau)}{\partial \tau} + \kappa_\epsilon (u_\epsilon^i \cdot n) = \frac{\partial (u_\epsilon^e \cdot \tau)}{\partial \tau} + \kappa_\epsilon (u_\epsilon^e \cdot n), \\
&\frac{\partial (u_\epsilon^i \cdot n)}{\partial \tau} - \kappa_\epsilon (u_\epsilon^i \cdot \tau) = \frac{\partial (u_\epsilon^e \cdot n)}{\partial \tau} - \kappa_\epsilon (u_\epsilon^e \cdot \tau), \qquad (34) \\
&\lambda_1 \left(\frac{\partial (u_\epsilon^i \cdot \tau)}{\partial \tau} + \kappa_\epsilon (u_\epsilon^i \cdot n) + \frac{\partial (u_\epsilon^i \cdot n)}{\partial n} \right) + 2\mu_1 \frac{\partial (u_\epsilon^i \cdot n)}{\partial n} \\
&\qquad = \lambda_0 \left(\frac{\partial (u_\epsilon^e \cdot \tau)}{\partial \tau} + \kappa_\epsilon (u_\epsilon^e \cdot n) + \frac{\partial (u_\epsilon^e \cdot n)}{\partial n} \right) + 2\mu_0 \frac{\partial (u_\epsilon^e \cdot n)}{\partial n}, \\
&\mu_1 \left(\frac{\partial (u_\epsilon^i \cdot \tau)}{\partial n} - \kappa_\epsilon (u_\epsilon^i \cdot \tau) + \frac{\partial (u_\epsilon^i \cdot n)}{\partial \tau} \right) = \mu_0 \left(\frac{\partial (u_\epsilon^e \cdot \tau)}{\partial n} - \kappa_\epsilon (u_\epsilon^e \cdot \tau) + \frac{\partial (u_\epsilon^e \cdot n)}{\partial \tau} \right),
\end{aligned}$$

and

$$
\begin{aligned}
(\mathbb{C}_1-\mathbb{C}_0)\,\widehat{\nabla}u_\epsilon^i &= (\lambda_1-\lambda_0)\left[\frac{\partial(u_\epsilon^i\cdot\tau)}{\partial\tau}+\frac{\partial(u_\epsilon^i\cdot n)}{\partial n}+\kappa_\epsilon(u_\epsilon^i\cdot n)\right](n\otimes n+\tau\otimes\tau)\\
&\quad+2(\mu_1-\mu_0)\left[\left(\frac{\partial(u_\epsilon^i\cdot\tau)}{\partial\tau}+\kappa_\epsilon(u_\epsilon^i\cdot n)\right)\tau\otimes\tau+\frac{\partial(u_\epsilon^i\cdot n)}{\partial n}\,n\otimes n\right.\\
&\quad\left.+\frac{1}{2}\left(\frac{\partial(u_\epsilon^i\cdot\tau)}{\partial n}+\frac{\partial(u_\epsilon^i\cdot n)}{\partial\tau}-\kappa_\epsilon(u_\epsilon^i\cdot\tau)\right)(\tau\otimes n+n\otimes\tau)\right]. \qquad (35)
\end{aligned}
$$

By solving the system (34) with respect to the derivatives of the components of u_ϵ^i, and inserting the result into (35), we derive (33).

Now, by inserting (33) into (32), we get

$$
\int_{\sigma_\epsilon'}(\mathbb{C}_1-\mathbb{C}_0)\,\widehat{\nabla}u_\epsilon^i\cdot\widehat{\nabla}N(\cdot,y)=2\epsilon\int_{\sigma_\epsilon'}\mathcal{M}_\epsilon\widehat{\nabla}u_\epsilon^e\cdot\widehat{\nabla}N(\cdot,y)+O(\epsilon^{1+\alpha-\beta(2+\alpha)}). \qquad (36)
$$

Fourth step.
We will show that

$$
\|\nabla u_\epsilon^e-\nabla u_0\|_{L^\infty(\sigma_\epsilon')}\le C\epsilon^\gamma\|g\|_{H^{-1/2}(\partial\Omega)} \qquad (37)
$$

for some positive γ.

In order to prove the above inequality, we need the following theorem.

THEOREM 4.1 (mean value property). *Let Ψ be a biharmonic scalar, vector, or tensor field in a open bounded domain D. Then, for any ball $B_\rho(y)\subset\subset D$,*

$$
\Psi(y)=\frac{1}{2\pi}\left[\frac{4}{\rho^2}\int_{B_\rho(y)}\Psi(x)\,dx-\frac{1}{\rho}\int_{\partial B_\rho(y)}\Psi(x)\,d\sigma_x\right]. \qquad (38)
$$

For the proof of Theorem 4.1, see [N].

Since $\nabla u_\epsilon-\nabla u_0$ is biharmonic in $\Omega\setminus\overline{\omega}_\epsilon$ we might use the mean value property (38) for points in the set $\Omega_K\setminus\overline{\omega}_d$, where $\Omega_K=\left\{x\in\Omega\,:\,d(x,\partial\Omega)>\frac{1}{2K}\right\}$ and d is such that $2\epsilon<d$.

Observe that, by (38), for every $y\in\Omega_K\setminus\overline{\omega}_d$ and for $0<\lambda\le\frac{d}{2}$,

$$
\nabla(u_\epsilon-u_0)(y)=\frac{1}{2\pi}\left[\frac{4}{\lambda^2}\int_{B_\lambda(y)}\nabla(u_\epsilon-u_0)-\frac{1}{\lambda}\int_{\partial B_\lambda}\nabla(u_\epsilon-u_0)\right]. \qquad (39)
$$

By using the divergence theorem we can rewrite (39) as follows:

$$
\nabla(u_\epsilon-u_0)(y)=\frac{1}{2\pi}\left[\frac{4}{\lambda^2}\int_{\partial B_\lambda(y)}(u_\epsilon-u_0)\otimes\tilde{\nu}\,d\sigma-\frac{1}{\lambda}\int_{\partial B_\lambda}\nabla(u_\epsilon-u_0)\right],
$$

where $\tilde{\nu}$ is the outward normal vector to ∂B_λ. If we multiply the last relation by λ^3 and integrate from 0 to $\rho=\frac{d}{2}$ we get

$$
\nabla(u_\epsilon-u_0)(y)=\frac{12}{\pi}\left[\frac{4}{d^4}\int_{B_{\frac{d}{2}}(y)}(u_\epsilon-u_0)\otimes\underline{r}\,dx-\frac{1}{d^4}\int_{B_{\frac{d}{2}}(y)}r^2\nabla(u_\epsilon-u_0)\,dx\right], \qquad (40)
$$

where $\underline{r}(x) = x - y$, $r = |\underline{r}|$.

From (40) and (16) we have that

$$\|\nabla(u_\epsilon - u_0)\|_{L^\infty(\Omega_K \setminus \omega_d)} \leq C d^{-2} \epsilon^{\frac{1}{2}}, \tag{41}$$

where $C = C(\Omega, K, \alpha_0, \beta_0)$. Let $x = z + \epsilon n(z)$ be any point on σ'_ϵ and let x_d be the point $x_d = z + dn(z)$. Since the entire line segment from x to x_d lies in $\Omega \setminus \overline{\omega}_\epsilon$ and has distance greater than $\frac{\epsilon^\beta}{2}$ from $\partial\sigma_0$, by Proposition 3.3 and arguing similarly as we did to prove (30), we have

$$|\nabla u^e_\epsilon(x) - \nabla u_\epsilon(x_d)| \leq C \epsilon^{-(2+\alpha)\beta} d^\alpha \|g\|_{H^{-1/2}(\partial\Omega)}, \tag{42}$$

where $C = C(K, \alpha_0, \beta_0)$.

By combining (41) and (42) we have that for any $x \in \sigma'_\epsilon$

$$\begin{aligned} |\nabla u^e_\epsilon(x) - \nabla u_0(x)| &\leq |\nabla u^e_\epsilon(x) - \nabla u_\epsilon(x_d)| \\ &\quad + |\nabla u_\epsilon(x_d) - \nabla u_0(x_d)| + |\nabla u_0(x_d) - \nabla u_0(x)| \\ &\leq C \left(d^\alpha \epsilon^{-(2+\alpha)\beta} + d^{-2} \epsilon^{\frac{1}{2}} + d \right) \|g\|_{H^{-1/2(\partial\Omega)}}. \end{aligned}$$

By choosing

$$d = \epsilon^{\frac{1}{2(\alpha+2)} + \beta}$$

we have (37) with $\gamma = \frac{\alpha}{2(\alpha+2)} - 2\beta$. Notice that $\gamma > 0$ if we choose $\beta < \frac{\alpha}{4(\alpha+2)}$.

By using (37), we have

$$2\epsilon \int_{\sigma'_\epsilon} \mathcal{M}_\epsilon \widehat{\nabla} u^e_\epsilon \cdot \widehat{\nabla} N(\cdot, y) = 2\epsilon \int_{\sigma'_\epsilon} \mathcal{M}_\epsilon \widehat{\nabla} u_0 \cdot \widehat{\nabla} N(\cdot, y) + O(\epsilon^{1+\gamma}).$$

Now, we recall that $d\sigma^\epsilon_x = (1 + O(\epsilon)) d\sigma_0$ and observe that, by assumption (3), $\mathcal{M}_\epsilon = (1 + O(\epsilon))\mathcal{M}$. Hence

$$2\epsilon \int_{\sigma'_\epsilon} \mathcal{M}_\epsilon \widehat{\nabla} u^e_\epsilon \cdot \widehat{\nabla} N(\cdot, y) = 2\epsilon \int_{\sigma_0} \mathcal{M} \widehat{\nabla} u_0 \cdot \widehat{\nabla} N(\cdot, y) + O(\epsilon^{1+\gamma}). \tag{43}$$

Finally, if we compare the remainders in the expansion in formulas (29), (32), (36), and (43), we have that (10) holds, if we choose $\alpha \in (0, 1/4)$ and $\beta \in (0, 1)$ such that $\beta < \frac{\alpha}{4(\alpha+2)}$. ∎

Remark 4.2. The asymptotic expansion also holds in the case where $\sigma_0 = \cup_{i=1}^M \sigma_i$ and $\sigma_1, \ldots, \sigma_M$ are disjoint and far from each other. In that case

$$(u_\epsilon - u_0)(y) = 2\epsilon \sum_{i=1}^M \int_{\sigma_i} \mathcal{M}_i \widehat{\nabla} u_0 \cdot \widehat{\nabla} N(\cdot, y)\, d\sigma_i + o(\epsilon),$$

where $\mathcal{M}_i$ is the restriction to σ_i of the tensor $\mathcal{M}$.

REFERENCES

[ABF] H. Ammari, E. Beretta, and E. Francini, *Reconstruction of thin conductivity imperfections*, Appl. Anal., 83 (2004), pp. 63–76.

[ABF2] H. Ammari, E. Beretta, and E. Francini, *Reconstruction of thin conductivity imperfections,* II. *The case of multiple segments*, Appl. Anal., 85 (2006), pp. 87–105.

[AK] H. Ammari and H. Kang, *Reconstruction of Small Inhomogeneities from Boundary Measurements*, Lecture Notes in Math. 1846, Springer-Verlag, Berlin, 2004.

[AKNT] H. Ammari, H. Kang, G. Nakamura, and K. Tanuma, *Complete asymptotic expansions of solutions of the system of elastostatics in the presence of an inclusion of small diameter and detection of an inclusion*, J. Elasticity, 67 (2002), pp. 97–129.

[BFV] E. Beretta, E. Francini, and M. S. Vogelius, *Asymptotic formulas for steady state voltage potentials in the presence of thin inhomogeneities. A rigorous error analysis*, J. Math. Pures Appl., 82 (2003), pp. 1277–1301.

[C] S. Campanato, *Sistemi Ellittici in Forma di Divergenza. Regolarità all'interno*, Quaderni, Scuola Normale Superiore di Pisa, Italy, 1980.

[F] G. Fichera, *Existence theorems in elasticity*, in Handbuch der Physik, Vol. VI, Springer-Verlag, Berlin, Heidelberg, New York, 1972, pp. 347–389.

[LM] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications,* Vol. I, Springer-Verlag, Berlin, 1972.

[LN] Y. Y. Li and L. Nirenberg, *Estimates for elliptic systems from composite material*, Commun. Pure Appl. Math., 56 (2003), pp. 892–925.

[N] M. Nicolesco, *Les Fonctions Polyharmoniques*, Hermann, Paris, 1936.

[T] A. Tiero, *On Korn's inequality in the second case*, J. Elasticity, 54 (1999), pp. 187–191.

[V] T. Valent, *Boundary Value Problems of Finite Elasticity*, Springer Tracts in Natural Philosophy 31, Springer-Verlag, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1262–1287

RIGOROUS UPSCALING OF THE REACTIVE FLOW THROUGH A PORE, UNDER DOMINANT PECLET AND DAMKOHLER NUMBERS*

ANDRO MIKELIĆ†, VINCENT DEVIGNE‡, AND C. J. VAN DUIJN§

Abstract. In this paper we present a rigorous derivation of the effective model for enhanced diffusion through a narrow and long 2D pore. The analysis uses a singular perturbation technique. The starting point is a local pore scale model describing the transport by convection and diffusion of a reactive solute. The solute particles undergo a first-order reaction at the pore surface. The transport and reaction parameters are such that we have large, dominant Peclet and Damkohler numbers with respect to the ratio of characteristic transversal and longitudinal lengths (the small parameter ε). We give a rigorous mathematical justification of the effective behavior for small ε. Error estimates are presented in the energy norm as well as in L^∞ and L^1 norms of the space variable. They guarantee the validity of the upscaled model. As a special case, we recover the well-known Taylor dispersion formula.

Key words. Taylor's dispersion, large Peclet number, singular perturbation, surface chemical reaction, large Damkohler number

AMS subject classifications. 35B25, 92E20, 76F25

DOI. 10.1137/050633573

1. Introduction.

We consider the transport of a reactive solute by diffusion and Poiseuille's convection in a semi-infinite 2D channel. The solute particles do not react among themselves. Instead they undergo a first-order chemical reaction at the wall of the channel. Following [12], we consider the following model for the solute concentration c^*:

(a) transport through channel $\Omega^* = \{(x^*, y^*) : 0 < x^* < +\infty, |y^*| < H\}$:

$$\frac{\partial c^*}{\partial t^*} + q(y^*)\frac{\partial c^*}{\partial x^*} - D^*\frac{\partial^2 c^*}{\partial (x^*)^2} - D^*\frac{\partial^2 c^*}{\partial (y^*)^2} = 0 \quad \text{in } \Omega^*, \tag{1}$$

where $q(z) = Q^*(1 - (z/H)^2)$ and where Q^* (velocity) and D^* (molecular diffusion) are positive constants.

(b) reaction at channel wall $\Gamma^* = \{(x^*, y^*) : 0 < x^* < +\infty, |y^*| = H\}$:

$$D^*\frac{\partial c^*}{\partial y^*} + k^* c^* = 0 \quad \text{on } \Gamma^*, \tag{2}$$

where k^* is the surface reaction coefficient.

*Received by the editors June 14, 2005; accepted for publication (in revised form) June 19, 2006; published electronically December 15, 2006.

http://www.siam.org/journals/sima/38-4/63357.html

†Institut Camille Jordan, UFR Mathématiques, Site de Gerland, Bât. A, Université Lyon 1, 50, avenue Tony Garnier, 69367 Lyon Cedex 07, France (Andro.Mikelic@univ-lyon1.fr). The research of this author was supported by the GDR MOMAS (Modélisation Mathématique et Simulations numériques liées aux problèmes de gestion des déchets nucléaires: 2439 - ANDRA, BRGM, CEA, EDF, CNRS) as a part of the project "Changements d'échelle dans la modélisation du transport multiphasique et réactif en milieux poreux: Application aux milieux fracturés et aux argiles."

‡Centre SITE, Ecole Nationale Supérieure des Mines de Saint-Etienne, 158, cours Fauriel, 42023 Saint-Etienne Cedex 2, France (VincentDevigne@emse.fr).

§Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, P. O. Box 513, 5600 MB Eindhoven, The Netherlands (c.j.v.duijn@tue.nl).

The natural way of analyzing this problem is to introduce appropriate scales. This requires characteristic or reference values for the parameters and variables involved. The obvious transversal length scale is H. For all other quantities we use reference values denoted by the subscript R. Setting

$$c = \frac{c^*}{\hat{c}_R},\ x = \frac{x^*}{L_R},\ y = \frac{y^*}{H},\ t = \frac{t^*}{T_R},\ Q = \frac{Q^*}{Q_R},\ D = \frac{D^*}{D_R},\ k = \frac{k^*}{k_R}, \tag{3}$$

where L_R is the "observation distance," we obtain the dimensionless equations

$$\frac{\partial c}{\partial t} + \frac{Q_R T_R}{L_R} Q(1-y^2)\frac{\partial c}{\partial x} - \frac{D_R T_R}{L_R^2} D\frac{\partial^2 c}{\partial x^2} - \frac{D_R T_R}{H^2} D\frac{\partial^2 c}{\partial y^2} = 0 \quad \text{in } \Omega \tag{4}$$

and

$$-\frac{D D_R}{H k_R}\frac{\partial c}{\partial y} = kc \quad \text{on } \Gamma, \tag{5}$$

where

$$\Omega = (0,+\infty)\times(-1,1) \quad \text{and} \quad \Gamma = (0,+\infty)\times\{-1,1\}. \tag{6}$$

The equations involve the following time scales:

$$T_L = \text{characteristic longitudinal time scale} = \frac{L_R}{Q_R},$$
$$T_T = \text{characteristic transversal time scale} = \frac{H^2}{D_R},$$
$$T_C = \text{superficial chemical reaction time scale} = \frac{H}{k_R},$$

and the nondimensional numbers

$$\mathbf{Pe} = \frac{L_R Q_R}{D_R} \quad \text{(Peclet number)},$$
$$\mathbf{Da} = \frac{L_R^2 k_R}{H D_R} \quad \text{(Damkohler number)}.$$

In this paper we fix the reference time by setting $T_R = T_L$. We are going to investigate the behavior of (4)–(5) with respect to the small parameter $\varepsilon = \frac{H}{L_R}$. Specifically, we will derive expressions for the effective values of the dispersion coefficient and velocity, and an effective 1D transport equation for small values of ε. To carry out the analysis, we need to compare the dimensionless numbers with respect to ε. For this purpose we set

$$\mathbf{Pe} = \varepsilon^{-\alpha} \quad \text{and} \quad \mathbf{Da} = \varepsilon^{\beta} \quad (\alpha, \beta \text{ to be chosen later}).$$

In the absence of chemical reactions, Taylor obtained in his well-known paper [19] an explicit expression for the enhanced diffusion originating from (1). It is known as Taylor's dispersion formula. We will recover this formula as a special case in our approach by setting $\alpha = 1$, $k = 0$ and by assuming $Q = \mathcal{O}(1)$. Note that

$$\frac{T_T}{T_L} = \frac{H Q_R}{D_R}\varepsilon = \varepsilon^2\ \mathbf{Pe} = \varepsilon \text{ (whenever } \alpha = 1).$$

Introducing the dimensionless numbers in (4)–(5) and considering constant initial/boundary conditions yields the problem

$$\frac{\partial c^\varepsilon}{\partial t} + Q(1-y^2)\frac{\partial c^\varepsilon}{\partial x} = D\varepsilon^\alpha \frac{\partial^2 c^\varepsilon}{\partial x^2} + D\varepsilon^{\alpha-2}\frac{\partial^2 c^\varepsilon}{\partial y^2} \quad \text{in } \Omega^+ \times (0,T), \tag{7}$$

$$-D\varepsilon^{\alpha-2}\frac{\partial c^\varepsilon}{\partial y} = -D\frac{1}{\varepsilon^2 \mathbf{Pe}}\frac{\partial c^\varepsilon}{\partial y} = k\frac{\mathbf{Da}}{\mathbf{Pe}}c^\varepsilon = k\varepsilon^{\alpha+\beta}c^\varepsilon \quad \text{on } \Gamma^+ \times (0,T), \tag{8}$$

$$c^\varepsilon(x,y,0) = 1 \quad \text{for } (x,y) \in \Omega^+, \tag{9}$$

$$c^\varepsilon(0,y,t) = 0 \quad \text{for } (y,t) \in (0,1)\times(0,T), \tag{10}$$

$$\frac{\partial c^\varepsilon}{\partial y}(x,0,t) = 0 \quad \text{for } (x,t) \in (0,+\infty)\times(0,T). \tag{11}$$

The latest condition results from the y-symmetry of the solution. Further

$$\Omega^+ = (0,+\infty)\times(0,1), \quad \Gamma^+ = (0,+\infty)\times\{1\},$$

and T is an arbitrarily chosen positive number.

We study the behavior of this problem as $\varepsilon \searrow 0$, while keeping the coefficients Q, D, and k all $\mathcal{O}(1)$. The most interesting case results when $\alpha+\beta=0$ and $0 \le \alpha < 2$, because then chemistry balances with flow in the limit as $\varepsilon \searrow 0$. Consequently, we shall restrict our attention to this situation.

In this paper we prove that the correct upscaling of (7)–(11) gives the 1D parabolic problem

$$\text{(EFF)} \qquad \begin{cases} \partial_t c + Q\left(\dfrac{2}{3} + \dfrac{4k}{45D}\varepsilon^{2-\alpha}\right)\partial_x c + k\left(1 - \dfrac{k}{3D}\varepsilon^{2-\alpha}\right)c \\ \\ = \left(D\varepsilon^\alpha + \dfrac{8}{945}\dfrac{Q^2}{D}\varepsilon^{2-\alpha}\right)\partial_{xx} c \ \text{ in } \ (0,+\infty)\times(0,T), \\ \\ c|_{x=0} = 0, \ c|_{t=0} = 1, \ \partial_x c \in L^2((0,+\infty)\times(0,T)). \end{cases}$$

We note that for $k=0$ and $\alpha=1$ this is exactly the effective model of Taylor [19].

What is known concerning the derivation of the effective problem (EFF), with or without chemical reactions? Below we give a short overview.

• In the absence of chemical reactions, Aris [1] presented a formal derivation using the method of moments.

• For the probabilistic justification of the Taylor dispersion, in the absence of chemical reactions, we refer to the lecture notes by Caflisch and Rubinstein [4]. This approach does not give an error estimate for the approximation.

• There have been numerous attempts to give a rigorous justification for the approximation in the absence of chemical reactions. The most convincing is the *near rigorous* derivation using the center manifold theory by Mercer and Roberts [13]. In this paper the initial value problem is studied and the Fourier transform with respect to x is applied. The center manifold theory is applied to obtain effective equations at various orders, however, without addressing the fact that one is dealing with an infinite dimensional case.

• Flow with chemistry, as described by (2), is considered by Paine, Carbonell, and Whitaker [15], who use the "single-point" closure schemes of turbulence modeling by Launder to obtain a closed model for the averaged concentration.

These studies do not provide a rigorous mathematical derivation of the Taylor dispersion formula, and in the presence of the chemical reactions it is even not clear how to average the problem.

It should be noted that the real interest is in deriving *dispersion equations* for reactive flows through porous media. If we consider a porous medium comprising a bundle of capillary tubes, then we arrive at our problem. The disadvantage is that a bundle of capillary tubes represents a geometrically oversimplified model of a porous medium. Nevertheless, there is considerable insight to be gained from the analysis of our model problem.

Our technique is strongly motivated by the paper by Rubinstein and Mauri [18], where effective dispersion and convection in porous media is studied using the homogenization approach. Their analysis is based on a hierarchy of time scales. In setting up the dimensionless equations, we followed their approach. To our knowledge, the only rigorous result concerning effective dispersion in porous media in the presence of high Peclet numbers (no chemistry) and with the characteristic transport time scale is given in the recent paper by Bourgeat, Jurak, and Piatnitski [3]. Their approach uses regular solutions with compatible data for the underlying linear transport equation. They assume a high order compatibility between the initial and boundary data, involving derivatives up to order five. They construct a smooth solution of the linear transport equation, add the appropriate boundary layer and initial layer, and add the correction due to the perturbation of the mean flow. The effective solution obtained in this way is an H^1-approximation of order ε and an L^2-approximation of order ε^2. However, in problems involving chemistry, one often encounters a jump between the initial value of the concentration and its value imposed at the injection boundary $x = 0$. This is also the case in the experiment described by Taylor [19].

Averaging the concentration in a tube with dissolution/precipitation occurring on the wall and with $\mathbf{Pe} = \mathcal{O}(1)$ is considered in [5].

For bounds on convection enhanced diffusion in porous media we refer to the work of Fannjiang, Papanicolaou, Zhikov, Kozlov, and Piatnitski. We do not give specific references because there is such an enormous number of papers on the subject. A detailed review of known results on the derivation of the effective equations of motion for the mean concentration, in the case of general heterogeneous media and transport velocities, is given in [11]. There one finds the rigorous homogenization theory for the spatio-temporal periodic velocity fields. However, in [11], the reference time is set to be the characteristic diffusion time, contrary to the choice made in [3] and the choice we have made. The case with chemical reactions, but in the absence of the transport, is considered in [2].

We note that our results also cover the case when the physical parameters result in large Peclet and small Damkohler numbers. This follows by setting $k = 0$, yielding Taylor's effective equation. In fact the effective equation (20) remains valid; only the effective terms representing the surface reaction are smaller and less important.

The plan of the paper is the following. In section 2 we study the homogenized problem. It turns out that it has an explicit solution having the form of a moving Gaussian, just as the 1D boundary layers of parabolic equations, when viscosity goes to zero (see [9]). Its behavior with respect to ε and t is singular.

In section 3 we give a justification of a lower order approximation, using a simple energy argument. In fact this approximation does not use Taylor's dispersion formula and gives an error of the same order in $L^\infty(L^2)$ as the solution to the linear transport equation.

In section 4 we give a formal derivation of the upscaled problem (EFF), using the approach proposed in [18].

The construction of the spatial boundary layer that takes care of the injection boundary is carried out in section 5.

Then in sections 6 and 7 we prove that the effective concentration satisfying the corresponding 1D parabolic problem, with Taylor's diffusion coefficient and the reactive correction, is an approximation in $L^\infty(L^2)$ and in $L^\infty(L^\infty)$ to the actual physical concentration.

To satisfy the curiosity of the reader not familiar with singular perturbation techniques, we give here the simplified version of the results stated in Theorems 5–7 from section 7. For simplicity, we compare only the physical concentration c^ε with c. Keeping the correction terms is necessary to have the same precision as stated in the theorems. Throughout the paper $H(x)$ denotes Heaviside's function

$$H(x) = 1,\ x > 0, \quad H(x) = 0,\ x \le 0. \tag{12}$$

Furthermore, using elementary parabolic theory one concludes that problem (7)–(11) has a unique bounded variational solution c^ε, with square integrable derivatives in x and y. Furthermore, c^ε belongs to C^∞ for $x > 0$ and stabilizes to 1 exponentially fast when $x \to \infty$.

THEOREM 1. *Let c be the unique solution of (EFF) and let $\Omega_K = (0, K) \times (0, 1)$, $K > 0$. Then we have*

$$\|t^3(c^\varepsilon - c)\|_{L^\infty(0,T;L^1(\Omega_K))} \le C\varepsilon^{2-\alpha}, \tag{13}$$

$$\|t^3(c^\varepsilon - c)\|_{L^\infty(0,T;L^2(\Omega_K))} \le C\big(\varepsilon^{2-5\alpha/4}H(1-\alpha) + \varepsilon^{3/2-3\alpha/4}H(\alpha-1)\big), \tag{14}$$

$$\|t^3\partial_y c^\varepsilon\|_{L^2(\Omega_K\times(0,T))} \le C\big(\varepsilon^{2-5\alpha/4}H(1-\alpha) + \varepsilon^{3/2-3\alpha/4}H(\alpha-1)\big), \tag{15}$$

$$\|t^3\partial_x(c^\varepsilon - c)\|_{L^2(\Omega_K\times(0,T))} \le C\big(\varepsilon^{2-7\alpha/4}H(1-\alpha) + \varepsilon^{3/2-5\alpha/4}H(\alpha-1)\big). \tag{16}$$

Note that c has disappeared in estimate (15) since it is only x and t dependent. This estimate is superior to estimate (16) because of the large $\mathcal{O}(\varepsilon^{\alpha-2})$ transversal diffusivity.

THEOREM 2. *Let c be the unique solution of (EFF). Then there exists a linear combination $\mathcal{C}_{cor}(x, y, t, \varepsilon)$ of products between polynomials in y and derivatives of c up to order 3, such that*

$$\|t^3(c^\varepsilon - c - \mathcal{C}_{cor})\|_{L^\infty((0,T)\times\Omega^+)} \le \begin{cases} C\varepsilon^{4-7\alpha/2-\delta} & \forall\delta > 0 \quad \text{if } \alpha < 1, \\ C\varepsilon^{3/2-\alpha-\delta} & \text{if } \alpha \ge 1. \end{cases} \tag{17}$$

COROLLARY 1. *Let $\alpha = 1$ and $k = 0$, as in Taylor's case. Then we have*

$$\|t^3(c^\varepsilon - c - \mathcal{C}_{cor})\|_{L^\infty((0,T)\times\Omega^+)} \le C\varepsilon^{1/2-\delta} \quad \forall\delta > 0, \tag{18}$$

$$\|t^3(c^\varepsilon - c)\|_{L^\infty(0,T;L^1(\Omega_K))} \le C\varepsilon. \tag{19}$$

The expression $\mathcal{C}_{cor}(x, y, t, \varepsilon)$ is given explicitly in Theorem 7 of section 7.

Our result could be stated in dimensional form as follows.

THEOREM 3. *Let us suppose that* $L_R > \max\{D_R/Q_R, Q_R H^2/D_R, H\}$. *Then the upscaled dimensional approximation to* (1), (2) *reads*

$$\frac{\partial c^{*,eff}}{\partial t^*} + \left(\frac{2}{3} + \frac{4}{45}\mathbf{Da}_T\right) Q^* \frac{\partial c^{*,eff}}{\partial x^*} + \frac{k^*}{H}\left(1 - \frac{1}{3}\mathbf{Da}_T\right) c^{*,eff} = D^*\left(1 + \frac{8}{945}\mathbf{Pe}_T^2\right)\frac{\partial^2 c^{*,eff}}{\partial (x^*)^2}, \tag{20}$$

where $\mathbf{Pe}_T = \frac{Q^* H}{D^*}$ *is the transversal Peclet number and* $\mathbf{Da}_T = \frac{k^* H}{D^*}$ *is the transversal Damkohler number.*

We conclude this section by noting that in the known literature on boundary layers for parabolic regularization, the transport velocity is assumed to be zero at the injection boundary (see [7]). Hence our result extends the existing framework.

One could try to get even higher order approximations. Unfortunately, our procedure then leads to higher order differential operators and it is not clear if they are easy to handle. In the absence of the boundaries, higher order terms were determined in [13] using the program REDUCE.

2. Study of the upscaled diffusion-convection equation on the half-line. For $\bar{Q}, \bar{D}, \varepsilon > 0$ and $\bar{k} \geq 0$, we consider the problem

$$\begin{cases} \partial_t u + \bar{Q}\partial_x u + \bar{k}u = \gamma \bar{D}\partial_{xx} u \ \text{ in } \ (0,+\infty)\times(0,T), \\ \partial_x u \in L^2((0,+\infty)\times(0,T)), \\ u(x,0) = 1 \ \text{ in } \ (0,+\infty), \quad u(0,t) = 0 \ \text{ at } \ x = 0. \end{cases} \tag{21}$$

The unique solution is given by the explicit formula

$$u(x,t) = e^{-\bar{k}t}\left(1 - \frac{1}{\sqrt{\pi}}\left[e^{\frac{\bar{Q}x}{\gamma \bar{D}}}\int_{\frac{x+t\bar{Q}}{2\sqrt{\gamma \bar{D}t}}}^{+\infty} e^{-\eta^2}\, d\eta + \int_{\frac{x-t\bar{Q}}{2\sqrt{\gamma \bar{D}t}}}^{+\infty} e^{-\eta^2}\, d\eta\right]\right). \tag{22}$$

This expression allows us to find the exact behavior of u with respect to γ. Note that for $\alpha \in [0,1]$, we will set $\gamma = \varepsilon^\alpha$; for $\alpha \in [1,2)$, we choose $\gamma = \varepsilon^{2-\alpha}$. The derivatives of u are found using Maple, and then their norms are estimated. Since the procedure is standard, we do not give the details. In more general situations, there are no explicit solutions and these estimates could be obtained using the technique and results from [9].

By the maximum principle we have

$$0 \leq u(x,t) \leq 1. \tag{23}$$

We first estimate the difference between $\chi_{\{x>\bar{Q}t\}}$ and u.

LEMMA 1. *Function* u *satisfies the estimates*

$$\int_0^\infty |e^{-\bar{k}t}\chi_{\{x>\bar{Q}t\}} - u(t,x)|\, dx = 3\sqrt{\gamma \bar{D}t}e^{-\bar{k}t} + C\gamma, \tag{24}$$

$$\|e^{-\bar{k}t}\chi_{\{x>\bar{Q}t\}} - u\|_{L^\infty(0,T;L^p((0,+\infty)))} \leq C\gamma^{1/(2p)} \quad \forall p \in (1,+\infty). \tag{25}$$

For the derivatives of u we have the following.

LEMMA 2. *Let ζ be defined by*

$$\zeta(t) = \begin{cases} \left(\dfrac{t}{\bar{D}\gamma}\right)^r & \text{for } 0 \le t \le \bar{D}\gamma, \\ 1 & \text{otherwise}, \end{cases} \tag{26}$$

with $r \ge q \ge 1$. Then

$$\|\zeta(t)(|\partial_t u| + |\partial_x u|)\|_{L^q((0,T)\times(0,+\infty))} \le C(\gamma\bar{D})^{\min\{1/(2q)-1/2,2/q-1\}},\ q \ne 3, \tag{27}$$

$$\|\zeta(t)(|\partial_t u| + |\partial_x u|)\|_{L^3((0,T)\times(0,+\infty))} \le C\left((\gamma\bar{D})^{-1}\log\left(\frac{1}{\gamma\bar{D}}\right)\right)^{1/3}. \tag{28}$$

Next we estimate the second-order derivatives.

LEMMA 3. *Let ζ be defined by* (26). *Then*

$$\begin{aligned} &\|\zeta(t)u_{tt}\|_{L^q((0,T)\times(0,+\infty))} + \|\zeta(t)u_{tx}\|_{L^q((0,T)\times(0,+\infty))} + \|\zeta(t)u_{xx}\|_{L^q((0,T)\times(0,+\infty))} \\ &\qquad \le C_q(\gamma\bar{D})^{\min\{1/(2q)-1,2/q-2\}},\ q \ne 3/2, \end{aligned} \tag{29}$$

$$\begin{aligned} &\|\zeta(t)u_{tt}\|_{L^{3/2}((0,T)\times(0,+\infty))} + \|\zeta(t)u_{tx}\|_{L^{3/2}((0,T)\times(0,+\infty))} + \|\zeta(t)u_{xx}\|_{L^{3/2}((0,T)\times(0,+\infty))} \\ &\qquad \le C\left((\gamma\bar{D})^{-1}\log\left(\frac{1}{\gamma\bar{D}}\right)\right)^{2/3}. \end{aligned} \tag{30}$$

For the third-order derivatives we have the following.

LEMMA 4. *Let ζ be defined by* (26). *Then*

$$\begin{aligned} &\|\partial_{xxx}(\zeta(t)u)\|_{L^q((0,T)\times(0,+\infty))} + \|\zeta(t)\partial_{xxt}u\|_{L^q((0,T)\times(0,+\infty))} \\ &\quad + \|\zeta(t)\partial_{xtt}u\|_{L^q((0,T)\times(0,+\infty))} \le C_q(\gamma\bar{D})^{2/q-3},\ q > 1, \end{aligned} \tag{31}$$

$$\begin{aligned} &\|\partial_{xxx}(\zeta(t)u)\|_{L^1((0,T)\times(0,+\infty))} + \|\zeta(t)\partial_{xxt}u\|_{L^1((0,T)\times(0,+\infty))} \\ &\quad + \|\zeta(t)\partial_{xtt}u\|_{L^1((0,T)\times(0,+\infty))} \le C_1(\gamma\bar{D})^{-1}\log\frac{1}{\gamma\bar{D}}. \end{aligned} \tag{32}$$

3. A simple L^2 error estimate. The simplest way to average problem (7)–(11) is to take the mean value with respect to y. Assuming that the mean of the product is the product of the means, which is in general wrong, we get the following problem for the "averaged" concentration $c_0^{eff}(x,t)$:

$$\begin{cases} \dfrac{\partial c_0^{eff}}{\partial t} + \dfrac{2Q}{3}\dfrac{\partial c_0^{eff}}{\partial x} + kc_0^{eff} = \varepsilon^\alpha D\dfrac{\partial c_0^{eff}}{\partial x^2} \ \text{ in } \ (0,+\infty)\times(0,T), \\ \partial_x c_0^{eff} \in L^2((0,+\infty)\times(0,T)), \quad c_0^{eff}|_{t=0} = 1, \quad c_0^{eff}|_{x=0} = 0. \end{cases} \tag{33}$$

This is problem (21) with $\tilde{Q} = \frac{2}{3}Q$, $\bar{k} = k$, and $\bar{D} = D$. The small parameter γ is equal to ε^α. For convenience we introduce the operator

$$\mathcal{L}^\varepsilon(w) := \frac{\partial w}{\partial t} + Q(1-y^2)\frac{\partial w}{\partial x} - D\varepsilon^\alpha\left(\frac{\partial^2 w}{\partial x^2} + \varepsilon^{-2}\frac{\partial^2 w}{\partial y^2}\right). \tag{34}$$

(7)–(11) are written as

$$\mathcal{L}^\varepsilon(c^\varepsilon) = 0 \quad \text{in } \Omega^+ \times (0,T), \tag{35}$$

$$c^\varepsilon(0,y,t) = 0 \quad \text{on } (0,1)\times(0,T), \tag{36}$$

$$\partial_y c^\varepsilon(x,0,t) = 0 \quad \text{on } (0,+\infty)\times(0,T), \tag{37}$$

$$-D\varepsilon^{\alpha-2}\partial_y c^\varepsilon(x,1,t) = kc^\varepsilon(x,1,t) \quad \text{on } (0,+\infty)\times(0,T), \tag{38}$$

$$c^\varepsilon(x,y,0) = 1 \quad \text{on } (0,+\infty)\times(0,1). \tag{39}$$

We want to approximate c^ε by c_0^{eff}. For this purpose we write

$$\mathcal{L}^\varepsilon(c_0^{eff}) = -kc_0^{eff} + Q\partial_x c_0^{eff}(1/3 - y^2) = R^\varepsilon,$$

$$\mathcal{L}^\varepsilon(c^\varepsilon - c_0^{eff}) = -R^\varepsilon \text{ in } \Omega^+\times(0,T), \quad \text{and} \tag{40}$$

$$-D\varepsilon^{\alpha-2}\partial_y(c^\varepsilon(x,1,t) - c_0^{eff}) = kc^\varepsilon(x,1,t) \quad \text{on } (0,+\infty)\times(0,T), \tag{41}$$

and we have the following useful estimate.

PROPOSITION 1. *Let $\Psi(x) = 1/(x+1)$ and let g^ε, ξ_0^ε, and R^ε be measurable functions satisfying*

$$\Psi g^\varepsilon \in H^1(\Omega^+\times(0,T)),\ \Psi\xi_0^\varepsilon \in L^2(\Omega^+),\ \text{and}\ \Psi R^\varepsilon \in L^2(\Omega^+\times(0,T)). \tag{42}$$

Furthermore, let $\xi \in L^\infty(\Omega^+\times(0,T))$, $\Psi\xi \in C([0,T];L^2(\Omega^+))$, $\Psi\nabla_{x,y}\xi \in L^2(\Omega^+\times(0,T))^2$ be a solution of the initial/boundary problem

$$\mathcal{L}^\varepsilon(\xi) = -R^\varepsilon \ \text{in}\ \Omega^+\times(0,T);\quad \partial_y\xi|_{y=0} = 0 \ \text{on}\ (0,+\infty)\times(0,T), \tag{43}$$

$$-D\varepsilon^{\alpha-2}\partial_y\xi|_{y=1} = k\xi|_{y=1} + g^\varepsilon|_{y=1} \quad \text{on}\ (0,+\infty)\times(0,T), \tag{44}$$

$$\xi|_{t=0} = \xi_0^\varepsilon \ \text{on}\ \Omega^+ \quad \text{and} \quad \xi|_{x=0} = 0 \ \text{on}\ (0,1)\times(0,T). \tag{45}$$

Then we have the energy estimate

$$\begin{aligned}\mathcal{E}(\xi,t) &= \frac{1}{2}\int_{\Omega^+}\Psi(x)^2\xi^2(t)\,dxdy + \frac{D}{2}\varepsilon^\alpha\int_0^t\int_{\Omega^+}\Psi(x)^2\\ &\times\left\{\varepsilon^{-2}|\partial_y\xi|^2 + |\partial_x\xi|^2\right\}dxdyd\tau + k\int_0^t\int_0^{+\infty}\xi^2|_{y=1}\Psi^2(x)\,dxd\tau\\ &\le -\int_0^t\int_{\Omega^+}\Psi(x)^2R^\varepsilon\xi\,dxdyd\tau - \int_0^t\int_0^{+\infty}g^\varepsilon|_{y=1}\xi|_{y=1}\Psi^2(x)\,dxd\tau\\ &+ 2D\varepsilon^\alpha\int_0^t\int_{\Omega^+}\Psi(x)^2\xi^2\,dxdyd\tau + \frac{1}{2}\int_{\Omega^+}\Psi(x)^2(\xi_0^\varepsilon)^2\,dxdy.\end{aligned} \tag{46}$$

Proof. We test (43)–(45) by $\Psi^2(x)\xi$ to obtain

$$\begin{aligned}&\frac{1}{2}\int_{\Omega^+}\xi^2(t)\Psi^2(x)\,dxdy + D\varepsilon^\alpha\int_0^t\int_{\Omega^+}\Psi^2(x)\left\{\varepsilon^{-2}|\partial_y\xi|^2 + |\partial_x\xi|^2\right\}dxdyd\tau\\ &\qquad + k\int_0^t\int_0^{+\infty}\xi^2|_{y=1}\Psi^2\,dxd\tau \le \frac{1}{2}\int_{\Omega^+}(\xi_0^\varepsilon)^2\Psi^2(x)\,dxdy\\ &\qquad - k\int_0^t\int_0^{+\infty}(g^\varepsilon\xi)|_{y=1}\Psi^2\,dxd\tau - D\varepsilon^\alpha\int_0^t\int_{\Omega^+}\partial_x\xi\xi\partial_x\Psi^2(x)\,dxdyd\tau.\end{aligned} \tag{47}$$

Next we use

$$D\varepsilon^\alpha \int_0^t \int_{\Omega^+} \partial_x \xi \xi \partial_x \Psi^2(x)\, dxd\tau \le \frac{D}{2}\varepsilon^\alpha \int_0^t \int_{\Omega^+} \Psi^2(x)|\partial_x \xi|^2\, dxdyd\tau$$
$$+ 2D\varepsilon^\alpha \int_0^t \int_{\Omega^+} \Psi^2(x)|\xi|^2\, dxdyd\tau, \tag{48}$$

giving directly (46). □

This simple result allows us to prove the following.

PROPOSITION 2. *There exist positive constants C_i^F, $i = 1, 2, 3$, such that*

$$\|\Psi(x)(c^\varepsilon - c_0^{eff})\|_{L^\infty(0,T;L^2(\Omega^+))} \le \varepsilon^{1-\alpha/2}\frac{F^0}{\sqrt{D}}, \tag{49}$$

$$\|\Psi(x)\partial_x(c^\varepsilon - c_0^{eff})\|_{L^2(0,T;L^2(\Omega^+))} \le \varepsilon^{1-\alpha}\frac{F^0}{D}, \tag{50}$$

$$\|\Psi(x)\partial_y(c^\varepsilon - c_0^{eff})\|_{L^2(0,T;L^2(\Omega^+))} \le \varepsilon^{2-\alpha}\frac{F^0}{D}, \tag{51}$$

where

$$F^0 = C_1^F \|\partial_x c_0^{eff}\|_{L^2((0,+\infty)\times(0,T))} + C_2^F k \le C_3^F \varepsilon^{-\alpha/4}. \tag{52}$$

Proof. We are in the situation of Proposition 1 with $\xi_0^\varepsilon = 0$ and $g^\varepsilon = kc_0^{eff}$. Consequently, for $\xi = c^\varepsilon - c_0^{eff}$ we have

$$\mathcal{E}(\xi, t) \le k\int_0^t \int_0^{+\infty} c_0^{eff}\left(\int_0^1 c^\varepsilon\, dy - c^\varepsilon|_{y=1}\right)\Psi^2\, dxd\tau + 2D\varepsilon^\alpha$$
$$\cdot \int_0^t \int_{\Omega^+} |\xi|^2\Psi^2(x)\, dxdyd\tau - \int_0^t \int_{\Omega^+} Q\left(\frac{1}{3} - y^2\right)\xi\partial_x c_0^{eff}\Psi^2\, dxdyd\tau. \tag{53}$$

It remains to estimate the first and third terms on the right. We have

$$\left|\int_0^t \int_{\Omega^+} Q\partial_x c_0^{eff}\left(\frac{1}{3} - y^2\right)\xi\Psi^2(x)\, dxdyd\tau\right|$$
$$= \left|\int_0^t \int_{\Omega^+} Q\partial_x c_0^{eff}\left(\frac{y}{3} - \frac{y^3}{3}\right)\partial_y\xi\Psi^2(x)\, dxdyd\tau\right| \tag{54}$$

and

$$k\left|\int_0^t \int_0^{+\infty} c_0^{eff}\left(\int_0^1 \xi\, dy - \xi|_{y=1}\right)\Psi^2\, dxd\tau\right|$$
$$\le \frac{D}{8}\varepsilon^\alpha \int_0^t \int_{\Omega^+} \Psi^2(x)|\partial_y\xi|^2\, dxdyd\tau + \frac{k^2}{D}\varepsilon^{2-\alpha}\int_0^t \int_0^{+\infty} (c_0^{eff})^2\Psi^2\, dxd\tau. \tag{55}$$

Inserting (54)–(55) into (53) gives

$$\mathcal{E}(c^\varepsilon - c_0^{eff}, t) \le \varepsilon^{2-\alpha}\int_0^t \int_0^{+\infty}\left\{\frac{2k^2}{D}(c_0^{eff})^2 + \frac{32}{315}\frac{Q^2}{D}(\partial_x c_0^{eff})^2\right\}\Psi^2\, dxd\tau$$
$$+ \int_0^t \int_0^1 \int_0^{+\infty} 2D\varepsilon^\alpha\Psi^2(x)(c^\varepsilon - c_0^{eff})^2\, dxdyd\tau, \tag{56}$$

and applying Gronwall's inequality results in (49)–(51). ∎

COROLLARY 2. *Let* $\Omega_K = (0,K)\times(0,1)$, $K>0$. *Then we have*

$$\|c^\varepsilon - c_0^{eff}\|_{L^\infty(0,T;L^2(\Omega_K))} \le C\varepsilon^{1-3\alpha/4}. \tag{57}$$

Remark 1. It is reasonable to expect L^1 estimates with better powers for ε. Unfortunately, testing (40) by the regularized sign $(c^\varepsilon - c_0^{eff})$ does not lead to a useful result. Hence at this stage claiming a $\sqrt{\varepsilon}$ estimate in L^1 is not justified.

Remark 2. There are results by Grenier [7] and Grenier and Guès [8] on singular perturbation problems. In [7] Grenier supposes that q is smooth and zero at $x=0$, together with its derivatives. Such hypotheses allow better estimates.

Remark 3. For $\alpha = 1$, the estimates (24) and (57) imply that the functions $\exp\{-kt\}\chi_{\{x>Qt\}}$ and c_0^{eff} approximate c^ε in $L^\infty(L^2)$ with the same order given by $C\varepsilon^{1/4}$.

Remark 4. Estimate (57) is not useful when $\alpha > 4/3$.

4. The formal 2-scale expansion leading to Taylor's dispersion. The estimate obtained in the previous section is not satisfactory. However, it is known that the Taylor dispersion model gives a very good 1D approximation. This motivates us to derive higher precision approximations. We give a formal 2-scale asymptotic expansion to obtain Taylor's (including the chemistry) dispersion formula.

We start with the problem (35)–(39) and search for c^ε in the form

$$c^\varepsilon = c^0(x,t;\varepsilon) + \varepsilon c^1(x,y,t) + \varepsilon^2 c^2(x,y,t) + \cdots. \tag{58}$$

After introducing (58) into (35) we get

$$\begin{aligned} &\varepsilon^0\Big\{\partial_t c^0 + Q(1-y^2)\partial_x c^0 - D\varepsilon^{\alpha-1}\partial_{yy}c^1\Big\} \\ &+\varepsilon\Big\{\partial_t c^1 + Q(1-y^2)\partial_x c^1 - D\varepsilon^{\alpha-1}\partial_{xx}c^0 - D\varepsilon^{\alpha-1}\partial_{yy}c^2\Big\} = \mathcal{O}(\varepsilon^2). \end{aligned} \tag{59}$$

To satisfy (59) for every $\varepsilon\in(0,\varepsilon_0)$, all coefficients in front of the powers of ε should be zero.

The problem corresponding to the ε^0 is

$$\begin{cases} -D\partial_{yy}c^1 = -\varepsilon^{1-\alpha}Q\left(\frac{1}{3}-y^2\right)\partial_x c^0 - \varepsilon^{1-\alpha}\left(\partial_t c^0 + 2Q\partial_x c^0/3\right) \ \text{ on } \ (0,1), \\ \partial_y c^1 = 0 \ \text{ on } \ y=0 \ \text{ and } \ -D\partial_y c^1 = k\varepsilon^{1-\alpha}c^0 \ \text{ on } \ y=1 \end{cases} \tag{60}$$

for every $(x,t)\in(0,+\infty)\times(0,T)$. By Fredholm's alternative, this problem has a solution if and only if

$$\partial_t c^0 + \frac{2Q\partial_x c^0}{3} + kc^0 = 0 \quad \text{in } \ (0,L)\times(0,T). \tag{61}$$

Unfortunately our initial and boundary data are incompatible, and therefore the solution to this hyperbolic equation is discontinuous. Since the asymptotic expansion for c^ε involves derivatives of c^0, (61) does not suit our needs. In [3] this difficulty was overcome by assuming compatible initial and boundary data. We proceed by following an idea from [18] and suppose that

$$\partial_t c^0 + \frac{2Q\partial_x c^0}{3} + kc^0 = \mathcal{O}(\varepsilon) \quad \text{in } \ (0,+\infty)\times(0,T). \tag{62}$$

This hypothesis will be justified a posteriori, after getting an equation for c^0.

Combining (60) and (61) and using hypothesis (62) leads us to consider

$$\left\{\begin{array}{l} -D\partial_{yy}c^1 = -\varepsilon^{1-\alpha}Q\left(\frac{1}{3}-y^2\right)\partial_x c^0 + \varepsilon^{1-\alpha}kc^0 \ \text{ on } \ (0,1), \\ \partial_y c^1 = 0 \ \text{ on } \ y=0 \ \text{ and } \ -D\partial_y c^1 = k\varepsilon^{1-\alpha}c^0 \ \text{ on } \ y=1 \end{array}\right. \tag{63}$$

for every $(x,t)\in(0,+\infty)\times(0,T)$. Consequently

$$c^1(x,y,t) = \varepsilon^{1-\alpha}\left(\frac{Q}{D}\left(\frac{y^2}{6}-\frac{y^4}{12}\right)\partial_x c^0 + \frac{k}{D}\left(\frac{1}{6}-\frac{y^2}{2}\right)c^0 + C_0(x,t)\right), \tag{64}$$

where C_0 is an arbitrary function.

The problem corresponding to ε^1 is

$$\left\{\begin{array}{l} -D\partial_{yy}c^2 = D\varepsilon\partial_{xx}c^1 - \varepsilon^{1-\alpha}Q(1-y^2)\partial_x c^1 + D\partial_{xx}c^0 - \varepsilon^{1-\alpha}\partial_t c^1 \\ -\varepsilon^{-\alpha}\left(\partial_t c^0 + \frac{2Q\partial_x c^0}{3} + kc^0\right) \quad \text{on } \ (0,1), \\ \partial_y c^2 = 0 \ \text{ on } \ y=0 \ \text{ and } \ -D\partial_y c^2 = k\varepsilon^{1-\alpha}c^1 \ \text{ on } \ y=1 \end{array}\right. \tag{65}$$

for every $(x,t)\in(0,+\infty)\times(0,T)$. This problem has a solution if and only if

$$\begin{aligned} &\partial_t c^0 + \frac{2Q\partial_x c^0}{3} + k(c^0+\varepsilon c^1|_{y=1}) + \varepsilon\partial_t\left(\int_0^1 c^1\ dy\right) - \varepsilon^\alpha D\partial_{xx}c^0 \\ &+ Q\varepsilon\partial_x\left(\int_0^1(1-y^2)c^1\ dy\right) - D\varepsilon^{1+\alpha}\partial_{xx}\left(\int_0^1 c^1 dy\right) = 0 \quad \text{in } \ (0,+\infty)\times(0,T). \end{aligned} \tag{66}$$

Note that this is the equation for c^0. In order to get the simplest possible equation we choose C_0 such that $\int_0^1 c^1\ dy = 0$. This implies

$$c^1(x,y,t) = \varepsilon^{1-\alpha}\left(\frac{Q}{D}\left(\frac{y^2}{6}-\frac{y^4}{12}-\frac{7}{180}\right)\partial_x c^0 + \frac{k}{D}\left(\frac{1}{6}-\frac{y^2}{2}\right)c^0\right), \tag{67}$$

and (66) becomes

$$\partial_t c^0 + Q\left(\frac{2}{3}+\frac{4k}{45D}\varepsilon^{2-\alpha}\right)\partial_x c^0 + k\left(1-\frac{k}{3D}\varepsilon^{2-\alpha}\right)c^0 = \varepsilon^\alpha\tilde{D}\partial_{xx}c^0 \quad \text{in } \ (0,+\infty)\times(0,T), \tag{68}$$

with

$$\tilde{D} = D + \frac{8}{945}\frac{Q^2}{D}\varepsilon^{2(1-\alpha)}. \tag{69}$$

As a result, problem (65) transforms into

$$\left\{\begin{array}{l} -D\partial_{yy}c^2 = \varepsilon^{2-2\alpha}\left\{-\frac{Q^2}{D}\partial_{xx}c^0\left\{\frac{8}{945}+(1-y^2)\left(\frac{y^2}{6}-\frac{y^4}{12}-\frac{7}{180}\right)\right\}\right. \\ +\partial_x c^0\frac{Qk}{D}\left\{\frac{2}{45}-(1-y^2)\left(\frac{1}{6}-\frac{y^2}{2}\right)\right\} + \frac{2kQ}{45D}\partial_x c^0 \\ -\frac{k^2}{3D}c^0 - \left(\frac{y^2}{6}-\frac{y^4}{12}-\frac{7}{180}\right)\left(\partial_{xt}c^0\frac{Q}{D} - \varepsilon^\alpha Q\partial_{xxx}c^0\right) \\ \left.-\left(\frac{1}{6}-\frac{y^2}{2}\right)\left(\partial_t c^0\frac{k}{D} - \varepsilon^\alpha k\partial_{xx}c^0\right)\right\} \ \text{ on } \ (0,1), \quad \partial_y c^2 = 0 \ \text{ on } \ y=0, \\ \text{and } \ -D\partial_y c^2 = \frac{Qk}{D}\varepsilon^{2-2\alpha}\partial_x c^0\frac{2}{45} - \frac{k^2}{3D}\varepsilon^{2-2\alpha}c^0 \ \text{ on } \ y=1. \end{array}\right. \tag{70}$$

If we choose c^2 such that $\int_0^1 c^2\ dy = 0$, then

$$
\begin{aligned}
c^2(x,y,t) = \varepsilon^{2-2\alpha}\Bigg\{ & -\frac{Q^2}{D^2}\partial_{xx}c^0\left(\frac{281}{453600} + \frac{23}{1512}y^2 - \frac{37}{2160}y^4 + \frac{1}{120}y^6 - \frac{1}{672}y^8\right) \\
& + \left(\frac{Q}{D^2}\partial_{xt}c^0 - \varepsilon^\alpha\frac{Q}{D}\partial_{xxx}c^0\right)\left(\frac{31}{7560} - \frac{7}{360}y^2 + \frac{y^4}{72} - \frac{y^6}{360}\right) \\
& + \frac{Qk}{D^2}\partial_x c^0\left(\frac{y^6}{60} - \frac{y^4}{18} + \frac{11y^2}{180} - \frac{11}{810}\right) \\
& + \left(\frac{k}{2D^2}\partial_t c^0 - \frac{k}{2D}\varepsilon^\alpha\partial_{xx}c^0\right)\left(-\frac{y^4}{12} + \frac{y^2}{6} - \frac{7}{180}\right) + \left(\frac{Qk}{45D^2}\partial_x c^0 - \frac{k^2}{6D^2}c^0\right)\left(\frac{1}{3} - y^2\right)\Bigg\}.
\end{aligned}
\tag{71}
$$

5. Boundary layer. The higher order approximations in the asymptotic expansion for c^ε do not satisfy the boundary conditions. Such incompatibility suggests that we should correct them using an appropriate boundary layer:

$$
\begin{cases}
-\Delta_{y,z}\beta = 0 \quad \text{for } (z,y) \in \Omega^+, \\
-\partial_y\beta = 0 \ \text{ for } y = 1, \quad \text{and for } y = 0, \\
\beta = \dfrac{y^2}{6} - \dfrac{y^4}{12} - \dfrac{7}{180} \ \text{ for } z = 0.
\end{cases}
\tag{72}
$$

Using the elementary variational theory for PDEs, we obtain the existence of a unique solution $\beta \in L^2_{loc}(\Omega^+)$ such that $\nabla\beta \in L^2_{loc}(\Omega^+)^2$. Since the average of the boundary value at $z = 0$ is zero, it follows that $\int_0^1 \beta(z,y)\ dy = 0$ for every $z \in (0,+\infty)$. This allows us to apply Poincaré's inequality in H^1:

$$
\int_0^1 \varphi^2\ dy \le \frac{1}{\pi^2}\int_0^1 |\partial_y\varphi|^2\ dy\ \ \forall\varphi \in H^1(0,1),\ \ \int_0^1 \varphi\ dy = 0,
\tag{73}
$$

and conclude that in fact $\beta \in H^1(\Omega^+)$. In order to prove that β represents a boundary layer, one should prove the exponential decay. We apply the theory from [14] and get the following result describing the decay of β as $z \to +\infty$.

PROPOSITION 3. *There exists a constant $\gamma_0 > 0$ such that the solution β of* (72) *satisfies the estimates*

$$
\int_z^{+\infty}\int_0^1 |\nabla_{y,z}\beta|^2\ dydz \le c_0 e^{-\gamma_0 z}, \quad z > 0,
\tag{74}
$$

$$
|\beta(y,z)| \le c_0 e^{-\gamma_0 z} \quad \forall (y,z) \in \Omega^+.
\tag{75}
$$

6. First correction. As explained, estimate (57) is not satisfactory. To get a better approximation we consider the correction, which was constructed using the formal 2-scale expansion in section 4.

Let $0 \leq \alpha < 2$. We start with the $\mathcal{O}(\varepsilon^{4-2\alpha})$ approximation and consider the function

(76)
$$c_1^{eff}(x,y,t;\varepsilon) = c(x,t;\varepsilon) + \varepsilon^{2-\alpha}\zeta(t)\left(\frac{Q}{D}\left(\frac{y^2}{6} - \frac{y^4}{12} - \frac{7}{180}\right)\frac{\partial c}{\partial x} + \frac{k}{D}\left(\frac{1}{6} - \frac{y^2}{2}\right)c\right),$$

where c is the solution to the effective problem with Taylor's dispersion coefficient including the reaction terms:

(77)
$$\begin{cases} \partial_t c + Q\left(\frac{2}{3} + \frac{4k}{45D}\varepsilon^{2-\alpha}\right)\partial_x c + k\left(1 - \frac{k}{3D}\varepsilon^{2-\alpha}\right)c \\ = \left(D\varepsilon^{\alpha} + \frac{8}{945}\frac{Q^2}{D}\varepsilon^{2-\alpha}\right)\partial_{xx} c \ \text{ in } \ (0,+\infty)\times(0,T), \\ c|_{x=0} = 0,\ c|_{t=0} = 1,\ \partial_x c \in L^2((0,+\infty)\times(0,T)). \end{cases}$$

The cut-off in time ζ is given by (26) and we use it to eliminate the time-like boundary layer appearing at $t = 0$. These effects are not visible in the formal expansion.

Let $\mathcal{L}^\varepsilon$ be the differential operator given by (34). Following the formal expansion from section 4, we find that $\mathcal{L}^\varepsilon$ applied to the correction without boundary layer functions and cut-offs would give $F_1^\varepsilon + F_2^\varepsilon + F_3^\varepsilon + F_4^\varepsilon + F_5^\varepsilon$, where

(78)
$$\begin{cases} F_1^\varepsilon = \partial_{xx} c\frac{Q^2}{D}\varepsilon^{2-\alpha}\left\{\frac{8}{945} + (1-y^2)\left(\frac{y^2}{6} - \frac{y^4}{12} - \frac{7}{180}\right)\right\}, \\ F_2^\varepsilon = \partial_x c\frac{Qk}{D}\varepsilon^{2-\alpha}\left\{-\frac{2}{45} + (1-y^2)\left(\frac{1}{6} - \frac{y^2}{2}\right)\right\}, \\ F_3^\varepsilon = \varepsilon^{2-\alpha}\left(\frac{y^2}{6} - \frac{y^4}{12} - \frac{7}{180}\right)\left\{\partial_{xt} c\frac{Q}{D} - \varepsilon^{\alpha}\partial_{xxx} cQ\right\}, \\ F_4^\varepsilon = \varepsilon^{2-\alpha}\left(\frac{1}{6} - \frac{y^2}{2}\right)\left\{\partial_t c\frac{k}{D} - \varepsilon^{\alpha}\partial_{xx} ck\right\}, \\ F_5^\varepsilon = \varepsilon^{2-\alpha}\left\{-\frac{2}{45}\partial_x c\frac{Qk}{D} + \frac{k^2}{3D}c\right\}. \end{cases}$$

These functions are not integrable up to $t = 0$, and for handling them we introduce the cut-off ζ.

Applying $\mathcal{L}^\varepsilon$ to c_1^{eff} gives

$$\begin{aligned} \mathcal{L}^\varepsilon(c_1^{eff}) &= \zeta(t)\sum_{j=1}^{5} F_j^\varepsilon + \left(\varepsilon^{2-\alpha}\partial_{xx} c\frac{Q^2}{D}\frac{8}{945} + Q\left(\frac{1}{3} - y^2\right)\partial_x c - kc\right) \\ &\times (1-\zeta(t)) + \zeta'(t)\varepsilon^{2-\alpha}\left(\partial_x c\frac{Q}{D}\left\{\frac{y^2}{6} - \frac{y^4}{12} - \frac{7}{180}\right\} + \frac{k}{2D}\left(\frac{1}{3} - y^2\right)c\right) \\ &=: \Phi_1^\varepsilon \ \text{ and } \ -\mathcal{L}^\varepsilon(c_1^{eff}) = \mathcal{L}^\varepsilon(c^\varepsilon - c_1^{eff}) = -\Phi_1^\varepsilon. \end{aligned} \tag{79}$$

At the lateral boundary $y = 1$ we have

$$-D\varepsilon^{\alpha-2}\partial_y c_1^{eff}|_{y=1} = \zeta(t)kc, \tag{80}$$

$$kc_1^{eff}|_{y=1} = k\left(c + \varepsilon^{2-\alpha}\frac{Q}{D}\zeta(t)\frac{2}{45}\partial_x c - \varepsilon^{2-\alpha}\frac{k}{3D}c\zeta(t)\right). \tag{81}$$

Now $c^\varepsilon - c_1^{eff}$ satisfies the system

$$\mathcal{L}^\varepsilon(c^\varepsilon - c_1^{eff}) = -\Phi_1^\varepsilon \text{ in } \Omega^+ \times (0,T), \tag{82}$$

$$\partial_y(c^\varepsilon - c_1^{eff})|_{y=0} = 0 \text{ on } (0,+\infty)\times(0,T), \tag{83}$$

$$-D\varepsilon^{\alpha-2}\partial_y(c^\varepsilon - c_1^{eff})|_{y=1} = k(c^\varepsilon - c_1^{eff})|_{y=1} + g^\varepsilon|_{y=1} \text{ on } (0,+\infty)\times(0,T), \tag{84}$$

$$(c^\varepsilon - c_1^{eff})|_{t=0} = 0 \text{ on } \Omega^+ \text{ and } (c^\varepsilon - c_1^{eff})|_{x=0} = \eta_0^\varepsilon \text{ on } (0,1)\times(0,T), \tag{85}$$

$$\text{with } g^\varepsilon = k\zeta(t)\varepsilon^{2-\alpha}\left(\partial_x c\frac{2Q}{45D} - c\frac{k}{3D}\right) + (1-\zeta)kc \tag{86}$$

$$\text{and } \eta_0^\varepsilon = -\varepsilon^{2-\alpha}\zeta(t)\partial_x c|_{x=0}\left(\frac{y^2}{6} - \frac{y^4}{12} - \frac{7}{180}\right)\frac{Q}{D}. \tag{87}$$

As a next step we estimate Φ_1^ε to find out if the right-hand side is smaller than in section 3.

PROPOSITION 4. *Let $O_t = \Omega^+ \times (0,t)$ and let $\varphi \in H^1(O_T)$ satisfy $\varphi = 0$ at $x = 0$. Then we have*

$$\left|\int_0^t\int_{\Omega^+} \zeta F_1^\varepsilon\varphi \, dxdyd\tau\right| \le C\varepsilon^{3(2-\alpha)/2}\|\zeta(\tau)\partial_{xx}c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}$$
$$\le C\big(\varepsilon^{3-5\alpha/2}H(1-\alpha) + \varepsilon^{1-\alpha/2}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}; \tag{88}$$

$$\left|\int_0^t\int_{\Omega^+} \zeta(\tau) F_3^\varepsilon\varphi \, dxdyd\tau\right| \le C\varepsilon^{3(2-\alpha)/2}$$
$$\cdot\left(\|\zeta(\tau)\partial_{xt}c\|_{L^2(O_t)} + \|\zeta(\tau)\partial_{xx}c\|_{L^2(O_t)}\right)\cdot\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}$$
$$\le C\big(\varepsilon^{3-5\alpha/2}H(1-\alpha) + \varepsilon^{1-\alpha/2}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}; \tag{89}$$

$$\left|\int_0^t\int_{\Omega^+} (1-\zeta)\partial_{xx}c\varepsilon^{2-\alpha}\frac{Q^2}{D}\varphi \, dxdyd\tau\right| \le C\varepsilon^{2-3\alpha/2}\|\varepsilon^{\alpha/2}\partial_x\varphi\|_{L^2(O_t)}$$
$$\cdot\|(1-\zeta)\partial_x c\|_{L^2(O_t)} \le C\varepsilon^{2-3\alpha/2}\|\varepsilon^{\alpha/2}\partial_x\varphi\|_{L^2(O_t)}; \tag{90}$$

$$\left|\int_0^t\int_{\Omega^+} (1-\zeta)Q\left(\frac{1}{3} - y^2\right)\partial_x c\varphi \, dxdyd\tau\right| \le C\varepsilon^{1-\alpha/2}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}$$
$$\cdot\|(1-\zeta)\partial_x c\|_{L^2(O_t)} \le C\varepsilon^{1-\alpha/2}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}; \tag{91}$$

$$\left|\int_0^t\int_{\Omega^+} \zeta'\left(\frac{t}{D\varepsilon}\right)\varepsilon^{2-\alpha}\left\{\partial_x c\frac{Q}{D}\left\{\frac{y^2}{6} - \frac{y^4}{12} - \frac{7}{180}\right\} - \frac{k}{2D}\left(\frac{1}{3} - y^2\right)c\right\}\cdot\varphi \, dxdyd\tau\right|$$
$$\le C\varepsilon^{3-3\alpha/2}\|\zeta'\partial_x c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}$$
$$\le C\big(\varepsilon^{3-5\alpha/2}H(1-\alpha) + \varepsilon^{1-\alpha/2}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}. \tag{92}$$

Proof. First note that in (88)–(89) and (91)–(92) the averages of the polynomials in y are zero. We write them in the form $P(y) = \partial_y P_1(y)$, where P_1 has zero traces at $y = 0, 1$. After partial integration with respect to y and applying the results from section 2, we obtain the estimates (88)–(89) and (91)–(92). Since $(1-\zeta)\partial_{xx}c$ is not square integrable, we cannot use the same approach to obtain (90). It is obtained by partial integration with respect to x. ∎

PROPOSITION 5. *With φ as in Proposition 4 we have*

$$\begin{aligned}\left|\int_0^t\int_{\Omega^+}\zeta F_2^\varepsilon\varphi\ dxdyd\tau\right| &\le C\varepsilon^{3(1-\alpha/2)}\|\zeta\partial_x c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\ &\le C\big(\varepsilon^{3-7\alpha/4}H(1-\alpha)+\varepsilon^{5/2-5\alpha/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)};\end{aligned}\tag{93}$$

$$\begin{aligned}\left|\int_0^t\int_{\Omega^+}\zeta F_4^\varepsilon\varphi\ dxdyd\tau\right| &\le C\varepsilon^{3-3\alpha/2}\Big(\|\zeta\partial_t c\|_{L^2(O_t)}+\varepsilon^\alpha\|\zeta\partial_{xx}c\|_{L^2(O_t)}\Big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\ &\le C\big(\varepsilon^{3-7\alpha/4)}H(1-\alpha)+\varepsilon^{5(2-\alpha)/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)};\end{aligned}\tag{94}$$

$$\begin{aligned}&\left|\int_0^t\int_0^{+\infty}\zeta\partial_x c\varepsilon^{2-\alpha}\Big(\int_0^1\varphi\ dy-\varphi|_{y=1}\Big)\ dxd\tau\right|\\ &\le C\varepsilon^{2-\alpha}\|\partial_x c\|_{L^2(0,t;L^2((0,+\infty)))}\left\|\int_0^1\varphi\ dy-\varphi|_{y=1}\right\|_{L^2(O_t)}\\ &\le C\big(\varepsilon^{3-7\alpha/4}H(1-\alpha)+\varepsilon^{5(2-\alpha)/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)};\end{aligned}\tag{95}$$

$$\left|\int_0^t\int_0^{+\infty}\zeta(t)c\varepsilon^{2-\alpha}\Big(\int_0^1\varphi\ dy-\varphi|_{y=1}\Big)\ dxd\tau\right|\le C\varepsilon^{3(1-\alpha/2)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)};\tag{96}$$

$$\begin{aligned}&\left|\int_0^t\int_0^{+\infty}(1-\zeta(t))c\Big(\int_0^1\varphi\ dy-\varphi|_{y=1}\Big)\ dxd\tau\right|\\ &\le C\big(\varepsilon H(1-\alpha)+\varepsilon^{2-\alpha}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}.\end{aligned}\tag{97}$$

COROLLARY 3. *With Φ_1^ε given by (79), g^ε by (86), and with φ as in Proposition 4 we have*

$$\begin{aligned}&\left|\int_0^t\int_{\Omega^+}\Phi_1^\varepsilon\varphi\ dxdyd\tau+\int_0^t\int_0^{+\infty}g^\varepsilon|_{y=1}\varphi|_{y=1}\ dxd\tau\right|\\ &\le C\big(\varepsilon^{1-\alpha/2}H(1-\alpha)+\varepsilon^{2-3\alpha/2}H(\alpha-1)\big)\Big\{\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}+\|\varepsilon^{\alpha/2}\partial_x\varphi\|_{L^2(O_t)}\Big\}.\end{aligned}\tag{98}$$

A natural next step would be to correct c_1^{eff} at $x=0$ and then apply Proposition 1. Due to the presence of the term containing the first-order derivative in x, the boundary layer corresponding to our problem does not enter into the theory from [14]. Consequently, one should generalize it to the second-order elliptic equations with first-order terms. The generalization in the case of periodic boundary conditions at the lateral boundary is done in [16]. To our knowledge, the generalization in the case of Neumann's boundary conditions at the lateral boundary was never published. It seems that the results from [16] apply also to this case (see [17]). In order to avoid developing a new boundary layer theory of second-order elliptic operators with important first-order terms, we simply use the boundary layer function corresponding to the Neumann problem for the Laplace operator (72). Then the transport term is ignored and a large error in the forcing term is created. The error is concentrated at small times, and by eliminating it we would obtain an appropriate estimate.

In order to use this particular point, we prove the following proposition.

PROPOSITION 6. *Let $\Psi(x) = 1/(x+1)$ and let g^ε and Φ^ε be measurable bounded functions satisfying the conditions*

$$\Psi g^\varepsilon \in H^1(\Omega^+ \times (0,T)) \ \text{ and } \ \Psi\Phi^\varepsilon \in L^2(\Omega^+ \times (0,T)). \tag{99}$$

Furthermore, let $\xi \in L^\infty(\Omega^+ \times (0,T))$, $\Psi\xi \in C^{0,\alpha_0}([0,T]; L^2(\Omega^+))$, $\Psi\nabla_{x,y}\xi \in L^2(\Omega^+ \times (0,T))^2$ be a solution of the initial/boundary problem

$$\mathcal{L}^\varepsilon(\xi) = -\Phi^\varepsilon \ \text{ in } \ \Omega^+ \times (0,T); \quad \partial_y\xi|_{y=0} = 0 \ \text{ on } \ (0,+\infty) \times (0,T), \tag{100}$$

$$-D\varepsilon^{\alpha-2}\partial_y\xi|_{y=1} = k\xi|_{y=1} + g^\varepsilon|_{y=1} \quad \text{on } \ (0,+\infty) \times (0,T), \tag{101}$$

$$\xi|_{t=0} = 0 \ \text{ on } \ \Omega^+ \quad \text{and} \quad \xi|_{x=0} = 0 \ \text{ on } \ (0,1) \times (0,T). \tag{102}$$

Then we have

$$\begin{aligned}
&\mathcal{E}(t^m\xi,t) = t^{2m}\int_{\Omega^+}\Psi(x)^2\xi^2(t)\,dxdy + D\varepsilon^\alpha\int_0^t\int_{\Omega^+}\Psi(x)^2\tau^{2m} \\
&\times\left\{\varepsilon^{-2}|\partial_y\xi|^2 + |\partial_x\xi|^2\right\}dxdyd\tau + k\int_0^t\int_0^{+\infty}\tau^{2m}\xi^2|_{y=1}\Psi^2(x)\,dxd\tau \\
&\le \left|\int_{\Omega^+}\tau^{2m}\Psi(x)^2\Phi^\varepsilon\xi\,dxdyd\tau\right| + \left|\int_0^t\int_0^{+\infty}\tau^{2m}g^\varepsilon\right|_{y=3D1}\xi|_{y=3D1}\Psi^2(x)\,dxd\tau| \\
&+ C_2D\varepsilon^\alpha\int_0^t\int_{\Omega^+}\tau^{2m}\Psi(x)^2\xi^2\,dxdyd\tau \quad \forall m \ge 1.
\end{aligned} \tag{103}$$

Remark 5. Clearly, we want to apply this abstract result to $\xi = c^\varepsilon - c_1^{eff}$. Then $\zeta(t)\partial_x c$ has the required regularity, since the cut-off eliminates the singularity. With c the analysis is more complicated. By direct calculation we have $\partial_t c \in L^q(0,T; L^2((0,+\infty)))$ for all $q \in [1,4/3)$ and we obtain the required Hölder regularity by the Sobolev imbedding. $\int_0^A\int_0^1 |\xi(x,y,t)|^2\,dxdy$ is Hölder-continuous with some exponent $\alpha_0 > 0$ for all $A < +\infty$, *independent of* ε. In complete analogy, c_0^{eff} defined by (33) also has the required regularity. Finally, the difference $c^\varepsilon - c_0^{eff}$ satisfies (40) and (41) and it is zero at $x = 0$ and $t = 0$. Then the classical parabolic regularity theory (see, e.g., [10]) implies the Hölder regularity in the time of the L^2 norm with respect to x, y. After combining all these results, we obtain the required regularity of ξ.

Proof. By the supposed Hölder continuity, there exists $t_M \in [0,T]$, $t_M > 0$, such that

$$\frac{1}{t_M^{\alpha_0}}\int_0^{+\infty}\int_0^1 |\xi(x,y,t_M)|^2\Psi^2(x)\,dxdy = \max_{t\in[0,T]}\frac{1}{t^{\alpha_0}}\int_0^{+\infty}\int_0^1 |\xi(x,y,t)|^2\Psi^2(x)\,dx. \tag{104}$$

Then we have

$$\begin{aligned}
\int_0^{t_M} m\tau^{2m-1}\int_{\Omega^+}|\xi|^2\Psi^2(x)\,dxdyd\tau &\le \int_{\Omega^+}\frac{|\xi|^2(t_M)}{t_M^{\alpha_0}}\Psi^2(x)\int_0^{t_M} m\tau^{2m-1+\alpha_0}\,d\tau \\
&= \frac{m}{2m+\alpha_0}t_M^{2m}\int_{\Omega^+}|\xi|^2(t_M)\Psi^2(x)\,dxdy
\end{aligned} \tag{105}$$

and

$$\begin{aligned}
&\frac{1}{2}t_M^{2m}\int_{\Omega^+}|\xi|^2(t_M)\Psi^2(x)\,dxdy + k\int_0^{t_M}\int_0^{+\infty}\tau^{2m}\xi^2|_{y=1}\Psi^2(x)\,dxd\tau\\
&+\int_0^{t_M} D\left(\varepsilon^\alpha\int_{\Omega^+}\tau^{2m}|\partial_x\xi|^2(\tau)\Psi^2(x)\,dxdy+\varepsilon^{\alpha-2}\int_{\Omega^+}\tau^{2m}|\partial_y\xi|^2(\tau)\Psi^2(x)\,dxdy\right)d\tau\\
&\qquad\le -\int_0^{t_M}\int_{\Omega^+}\tau^{2m}\Phi^\varepsilon\xi\,dxdyd\tau - k\int_0^t\int_0^{+\infty}\tau^{2m}\xi|_{y=1}g^\varepsilon|_{y=1}\Psi^2(x)\,dxd\tau\\
&\qquad + D\varepsilon^\alpha\int_0^{t_M}\int_{\Omega^+}\tau^{2m}\Psi^2(x)\xi^2\,dxdyd\tau + m\int_0^{t_M}\int_{\Omega^+}\tau^{2m-1}|\xi|^2\Psi^2\,dxdyd\tau. \qquad (106)
\end{aligned}$$

Using (105) we find (103) for $t = t_M$ and with $C_2 = 0$. Getting the estimate (103) for general $t \in (0,T)$ is now straightforward. ∎

To use this estimate we should refine the estimates in Propositions 4 and 5. First, we note that estimate (29) changes to

$$\begin{aligned}
&\|t^m\partial_{tt}c\|_{L^q((0,T)\times(0,+\infty))} + \|t^m\partial_{tx}c\|_{L^q((0,T)\times(0,+\infty))} + \|t^m\partial_{xx}c\|_{L^q((0,T)\times(0,+\infty))}\\
&\qquad\le C_q(m)(\gamma\bar{D})^{1/(2q)-1}. \qquad (107)
\end{aligned}$$

Hence one gains $\varepsilon^{\alpha/4}$ (respectively, $\varepsilon^{1/2-\alpha/4}$) for the L^2 norm. In analogy with Propositions 4 and 5 we have Proposition 7.

PROPOSITION 7. *Let φ be as in Proposition 4 and let $m > 1$. Then we have*

$$\begin{aligned}
&\left|\int_0^t\int_0^\infty\int_0^1\tau^m\zeta F_1^\varepsilon\varphi\,dxdyd\tau\right| \le C\varepsilon^{3(2-\alpha)/2}\|\tau^m\partial_{xx}c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\
&\qquad\le C\big(\varepsilon^{3-9\alpha/4}H(1-\alpha)+\varepsilon^{3/2-3\alpha/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}; \qquad (108)
\end{aligned}$$

$$\begin{aligned}
&\left|\int_0^t\int_0^\infty\int_0^1\tau^m\zeta F_3^\varepsilon\varphi\,dxdyd\tau\right| \le C\varepsilon^{3(2-\alpha)/2}\\
&\quad\cdot\left(\|\tau^m\partial_{xt}c\|_{L^2(O_t)}+\|\tau^m\partial_{xx}c\|_{L^2(O_t)}\right)\cdot\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\
&\qquad\le C\big(\varepsilon^{3-9\alpha/4}H(1-\alpha)+\varepsilon^{3/2-3\alpha/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}; \qquad (109)
\end{aligned}$$

$$\begin{aligned}
&\left|\int_0^t\int_{\Omega^+}\zeta\tau^m F_2^\varepsilon\varphi\,dxdyd\tau\right| \le C\varepsilon^{3(1-\alpha/2)}\|\tau^m\zeta\partial_x c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\
&\qquad\le C\big(\varepsilon^{3-7\alpha/4}H(1-\alpha)+\varepsilon^{5/2-5\alpha/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}; \qquad (110)
\end{aligned}$$

$$\begin{aligned}
&\left|\int_0^t\int_{\Omega^+}\zeta\tau^m F_4^\varepsilon\varphi\,dxdyd\tau\right| \le C\varepsilon^{3-3\alpha/2}\\
&\quad\cdot\left(\|\zeta\tau^m\partial_t c\|_{L^2(O_t)}+\varepsilon^\alpha\|\zeta\tau^m\partial_{xx}c\|_{L^2(O_t)}\right)\cdot\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\
&\qquad\le C\big(\varepsilon^{3-7\alpha/4)}H(1-\alpha)+\varepsilon^{5(2-\alpha)/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}; \qquad (111)
\end{aligned}$$

$$\begin{aligned}&\left|\int_0^t\int_0^{+\infty}\zeta\tau^m\partial_x c\varepsilon^{2-\alpha}\left(\int_0^1\varphi\ dy-\varphi|_{y=1}\right)\ dxd\tau\right|\\&\le C\varepsilon^{2-\alpha}\|\tau^m\partial_x c\|_{L^2(O_t)}\left\|\int_0^1\varphi\ dy-\varphi|_{y=1}\right\|_{L^2(O_t)}\\&\le C\big(\varepsilon^{3-7\alpha/4}H(1-\alpha)+\varepsilon^{5(2-\alpha)/4}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)};\end{aligned}\tag{112}$$

$$\begin{aligned}&\left|\int_0^t\int_0^{+\infty}\zeta(t)\tau^m c\varepsilon^{2-\alpha}\left(\int_0^1\varphi\ dy-\varphi|_{y=1}\right)\ dxd\tau\right|\\&\le C\varepsilon^{3(1-\alpha/2)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}.\end{aligned}\tag{113}$$

Proof. These estimates are straightforward consequences of Propositions 4 and 5. □

We improve these results with respect to the other terms.

Without proof we state the following.

PROPOSITION 8. *Let φ be as in Proposition 4 and let $m>1$. Then we have*

$$\begin{aligned}&\left|\int_0^t\int_0^{\infty}\int_0^1(1-\zeta)\tau^m\partial_{xx}c\frac{Q^2}{D}\varepsilon^{2-\alpha}\varphi\ dxdyd\tau\right|\\&\le C\varepsilon^{2-3\alpha/2}\|(1-\zeta)\tau^m\partial_x c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2}\partial_x\varphi\|_{L^2(O_t)}\\&\le C\big(\varepsilon^{m\alpha+2-3\alpha/2}H(1-\alpha)+\varepsilon^{m(2-\alpha)+2-3\alpha/2}H(\alpha-1)\big)\|\varepsilon^{\alpha/2}\partial_x\varphi\|_{L^2(O_t)};\end{aligned}\tag{114}$$

$$\begin{aligned}&\left|\int_0^t\int_0^{\infty}\int_0^1(1-\zeta)\tau^m Q\left(\frac13-y^2\right)\partial_x c\varphi\ dxdyd\tau\right|\\&\le C\varepsilon^{1-\alpha/2}\|(1-\zeta)\tau^m\partial_x c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\&\le C\big(\varepsilon^{m\alpha+1-\alpha/2}H(1-\alpha)+\varepsilon^{m(2-\alpha)+1-\alpha/2}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)};\end{aligned}\tag{115}$$

$$\begin{aligned}&\left|\int_0^t\int_{\Omega^+}\zeta'\left(\frac{t}{D\varepsilon}\right)\tau^m\varepsilon^{2-\alpha}\left\{\partial_x c\frac{Q}{D}\left\{\frac{y^2}{6}-\frac{y^4}{12}-\frac{7}{180}\right\}-\frac{k}{2D}\left(\frac13-y^2\right)c\right\}\cdot\varphi\ dxdyd\tau\right|\\&\le C\varepsilon^{3-3\alpha/2}\|\zeta'\tau^m\partial_x c\|_{L^2(O_t)}\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}\\&\le C\big(\varepsilon^{3-3\alpha/2+\alpha(m-1)}H(1-\alpha)+\varepsilon^{3-3\alpha/2+(2-\alpha)(m-1)}H(\alpha-1)\big)\|\varepsilon^{\alpha/2-1}\partial_y\varphi\|_{L^2(O_t)}.\end{aligned}\tag{116}$$

Before applying Proposition 6 and getting the final estimate, we should correct the trace at $x=0$. This is done by adding

$$\bar c_1^{eff}=-\varepsilon^{2-\alpha}\zeta(t)\beta^\varepsilon\partial_x c\frac{Q}{D},\tag{117}$$

where $\beta^\varepsilon(x,y)=\beta(x/\varepsilon,y)$ is the boundary layer function given by (72). Let $g_1^\varepsilon=g^\varepsilon-\varepsilon^{2-\alpha}\frac{kQ\zeta\partial_x c\beta^\varepsilon}{D}$. Then for $\xi^\varepsilon=c^\varepsilon-c_1^{eff}-\bar c_1^{eff}$ we have the problem

$$\mathcal{L}^\varepsilon(\xi^\varepsilon) = -\Phi^\varepsilon = -\Phi_1^\varepsilon + \partial_t\zeta\varepsilon^{2-\alpha}\partial_x c\frac{Q}{D}\beta^\varepsilon + \varepsilon^{2-\alpha}\beta^\varepsilon\zeta(t)$$
$$\times\left\{\partial_{xt}c\frac{Q}{D} - \varepsilon^\alpha\partial_{xxx}cQ\right\} + \partial_x\beta^\varepsilon\frac{Q^2}{D}(1-y^2)\zeta\varepsilon^{2-\alpha}\partial_x c - \varepsilon^{2-\alpha}Q\partial_{xx}c\zeta(t)$$
$$\times\left(2\varepsilon^\alpha\partial_x\beta^\varepsilon - \beta^\varepsilon(1-y^2)\frac{Q}{D}\right) \text{ in } \Omega^+\times(0,T), \tag{118}$$

$$-D\varepsilon^{\alpha-2}\partial_y\xi^\varepsilon = k\xi^\varepsilon + g_1^\varepsilon \text{ on } (0,+\infty)\times\{1\}\times(0,T), \tag{119}$$

$$\partial_y\xi^\varepsilon|_{y=0} = 0 \text{ on } (0,+\infty)\times(0,T), \tag{120}$$

$$\xi^\varepsilon|_{t=0} = 0 \text{ on } \Omega^+ \quad\text{and}\quad \xi^\varepsilon|_{x=0} = 0 \text{ on } (0,1)\times(0,T). \tag{121}$$

We need an estimate for the new terms. It will be obtained from the following auxiliary result.

LEMMA 5. *With β and c defined by* (72) *and* (77)*, respectively, and with $m \geq 1$, we have*

$$\|\tau^m\zeta'\beta^\varepsilon\partial_x c\|_{L^2(O_t)} \leq C\varepsilon^{m-(\alpha+3)/4}\big\{H(1-\alpha) + \varepsilon^{(\alpha-1)/2}H(\alpha-1)\big\} \leq C\varepsilon^{m-1}; \tag{122}$$

$$\|\tau^m\zeta\beta^\varepsilon|_{y=1}\partial_x c\|_{L^2(O_t)} \leq C\varepsilon^{m+1/4}\big\{\varepsilon^{-\alpha/4}H(1-\alpha) + \varepsilon^{\alpha/4-1/2}H(\alpha-1)\big\} \leq C\varepsilon^m; \tag{123}$$

$$\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_x c\|_{L^2(O_t)} \leq C\varepsilon^{m-3/4}\big\{\varepsilon^{-\alpha/4}H(1-\alpha) + \varepsilon^{\alpha/4-1/2}H(\alpha-1)\big\} \leq C\varepsilon^{m-1}; \tag{124}$$

$$\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_t c\|_{L^2(O_t)} \leq C\varepsilon^{m-5/4}\big\{\varepsilon^{\alpha/2}H(1-\alpha) + \varepsilon^{1-\alpha/2}H(\alpha-1)\big\} \leq C\varepsilon^{m-5/4}; \tag{125}$$

$$\begin{aligned}&\|\tau^m\zeta\beta^\varepsilon\partial_{xx}c\|_{L^2(O_t)}\\ &\leq C\varepsilon^m\big\{(\varepsilon^{-1/4-\alpha/2} + \varepsilon^{1/4-3\alpha/4})H(1-\alpha) + (\varepsilon^{\alpha/2-5/4} + \varepsilon^{-5/2+3\alpha/4})H(\alpha-1)\big\}\\ &\leq C\varepsilon^{m-7/4};\end{aligned} \tag{126}$$

$$\begin{aligned}&\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_{xx}c\|_{L^2(O_t)}\\ &\leq C\varepsilon^{m-1}\big\{(\varepsilon^{-1/4-\alpha/2} + \varepsilon^{1/4-3\alpha/4})H(1-\alpha) + (\varepsilon^{\alpha/2-5/4} + \varepsilon^{-5/2+3\alpha/4})H(\alpha-1)\big\}\\ &\leq C\varepsilon^{m-7/4}.\end{aligned} \tag{127}$$

Proof. Since

$$\begin{aligned}\int_0^{+\infty}|\partial_x c\beta^\varepsilon|^2\,dx &\leq C\int_0^{+\infty}\exp\left\{-\frac{2\gamma_0 x}{\varepsilon}\right\}\exp\left\{-\frac{(x-\tau\bar{Q})^2}{2\gamma\bar{D}\tau}\right\}\frac{dx}{\gamma\tau\bar{D}}\\ &\leq C(\varepsilon D\tau)^{-1/2}\exp\{-C_0\tau/\varepsilon\}\,dxd\tau,\end{aligned} \tag{128}$$

inequalities (122), (123), and (124) follow by integration with respect to τ. Furthermore, since

$$\begin{aligned}\int_0^{+\infty}|\partial_t c\beta^\varepsilon|^2\,dx &\leq C\int_0^{+\infty}x^2\exp\left\{-\frac{2\gamma_0 x}{\varepsilon}\right\}\exp\left\{-\frac{(x-\tau\bar{Q})^2}{2\gamma\bar{D}\tau^3}\right\}\frac{dx}{\gamma\tau\bar{D}}\\ &\leq C(\varepsilon D\tau^3)^{-1/2}\exp\{-C_0\tau/\varepsilon\}\,dxd\tau,\end{aligned} \tag{129}$$

inequality (125) follows. Finally, since

$$\begin{aligned}&\|\tau^m\zeta\beta^\varepsilon\partial_{xx}c\|_{L^2(O_t)}\\ &\le C(\|\tau^m\zeta\beta^\varepsilon\partial_x c\|_{L^2(O_t)}+\|\tau^m\zeta\beta^\varepsilon\partial_t c\|_{L^2(O_t)})(\varepsilon^{-\alpha}H(1-\alpha)+\varepsilon^{\alpha-2}H(\alpha-1)),\end{aligned}\tag{130}$$

we obtain (126) and (127). □

PROPOSITION 9. *With φ as in Proposition 4, we have*

$$\begin{aligned}&\left|\int_0^t\int_{\Omega^+}\varepsilon^{2-\alpha}\tau^m\zeta(\tau)\beta^\varepsilon\left\{\partial_{xt}c\frac{Q}{D}-\varepsilon^\alpha\partial_{xxx}cQ\right\}\varphi\ dxdyd\tau\right|\\ &\le C\varepsilon^{2-\alpha}\Big(\big\{\|\zeta\tau^m\partial_t c\partial_x\beta^\varepsilon\|_{L^2(O_t)}+\varepsilon^\alpha\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_{xx}c\|_{L^2(O_t)}\big\}\|\varphi\|_{L^2(O_t)}\\ &\quad+\varepsilon^{-\alpha/2}\big\{\|\zeta\tau^m\partial_t c\beta^\varepsilon\|_{L^2(O_t)}+\varepsilon^\alpha\|\tau^m\zeta\partial_{xx}c\|_{L^2(O_t)}\big\}\|\varepsilon^{\alpha/2}\partial_x\varphi\|_{L^2(O_t)}\Big)\\ &\le C\varepsilon^{m+1/4-\alpha}\big(\|\varphi\|_{L^2(O_t)}+\|\varepsilon^{\alpha/2}\partial_x\varphi\|_{L^2(O_t)}\big);\end{aligned}\tag{131}$$

$$\begin{aligned}&\left|\int_0^t\int_{\Omega^+}\varepsilon^{2-\alpha}\zeta\tau^m\partial_{xx}c\varphi\left(-\beta^\varepsilon\frac{Q}{D}(1-y^2)+2\varepsilon^\alpha\partial_x\beta^\varepsilon\right)\ dxdyd\tau\right|\\ &\le C\varepsilon^{2-\alpha}\big(\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_{xx}c\|_{L^2(O_t)}+\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_{xx}c\|_{L^2(O_t)}\big)\|\varphi\|_{L^2(O_t)}\\ &\le C\varepsilon^{m-\alpha+1/4}\|\varphi\|_{L^2(O_t)};\end{aligned}\tag{132}$$

$$\begin{aligned}&\left|\int_0^t\int_{\Omega^+}\varepsilon^{2-\alpha}\zeta\tau^m\partial_x c\partial_x\beta^\varepsilon(1-y^2)\varphi\ dxdyd\tau\right|\\ &\le C\varepsilon^{2-\alpha}\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_x c\|_{L^2(O_t)}\|\varphi\|_{L^2(O_t)}\le C\varepsilon^{m-\alpha+1}\|\varphi\|_{L^2(O_t)};\end{aligned}\tag{133}$$

$$\begin{aligned}&\left|\int_0^t\int_0^{+\infty}\varepsilon^{2-\alpha}\zeta\tau^m\partial_x c\varphi|_{y=1}\beta^\varepsilon|_{y=1}\ dxd\tau\right|\\ &\le C\varepsilon^{2-\alpha}\|\tau^m\zeta\partial_x\beta^\varepsilon\partial_x c\|_{L^2(O_t)}\|\varphi\|_{L^2((0,t)\times\Gamma^+)}\le C\varepsilon^{m-\alpha+1}\|\varphi\|_{L^2((0,t)\times\Gamma^+)};\end{aligned}\tag{134}$$

$$\begin{aligned}&\left|\int_0^t\int_{\Omega^+}\varepsilon^{2-\alpha}\zeta'(\tau)\tau^m\partial_x c\varphi\beta^\varepsilon\ dxdyd\tau\right|\\ &\le C\varepsilon^{2-\alpha}\|\tau^m\zeta'\beta^\varepsilon\partial_x c\|_{L^2(O_t)}\|\varphi\|_{L^2(O_t)}\le C\varepsilon^{m-\alpha+3/4}\|\varphi\|_{L^2(O_t)}.\end{aligned}\tag{135}$$

At this point the application of Proposition 6 is straightforward. As a result we get the following.

THEOREM 4. *Let c be the solution of (77) and let c_1^{eff} and $\bar{c}_1^{eff}$ be given by (76) and (117), respectively. With $\Omega_K=(0,K)\times(0,1)$, $K>0$, we have*

$$\begin{aligned}&\|t^3(c^\varepsilon-c_1^{eff}(x,t;\varepsilon)-\bar{c}_1^{eff})\|_{L^\infty(0,T;L^2(\Omega_K))}\\ &\le C\big(\varepsilon^{3-9\alpha/4}H(1-\alpha)+\varepsilon^{3(1-\alpha/2)/2}H(\alpha-1)\big);\end{aligned}\tag{136}$$

$$\begin{aligned}&\|t^3\partial_y(c^\varepsilon-c_1^{eff}(x,t;\varepsilon)-\bar{c}_1^{eff})\|_{L^2(\Omega_K\times(0,T))}\\ &\le C\varepsilon^{1-\alpha/2}\big(\varepsilon^{3-9\alpha/4}H(1-\alpha)+\varepsilon^{3(1-\alpha/2)/2}H(\alpha-1)\big);\end{aligned}\tag{137}$$

$$\begin{aligned}&\|t^2\partial_x(c^\varepsilon-c_1^{eff}(x,t;\varepsilon)-\bar{c}_1^{eff})\|_{L^2(\Omega_K\times(0,T))}\\ &\le C\varepsilon^{-\alpha/2}\big(\varepsilon^{3-9\alpha/4}H(1-\alpha)+\varepsilon^{3(1-\alpha/2)/2}H(\alpha-1)\big).\end{aligned}\tag{138}$$

7. Error estimate involving the second order in expansion. The most important power in the Peclet number α is $\alpha = 1$ because it describes Taylor's scaling. In this case our approximation is of order $\varepsilon^{3/4}$ in L^2. It is of interest to reach the order ε at least in this case. Also, it is of interest to get the higher order estimates because ε is *frequently* not very small.

After the results of section 6, the leading order terms in the estimates are ζF_1^ε and ζF_3^ε, where F_j^ε, $j = 1, \dots, 3$, are defined in (78). When deriving formally the effective equation, we have seen that they could be eliminated by introducing the next order correction. Following the formal expansion, we find that c_1^{eff} should be replaced with $c_1^{eff} + c_2^{eff}$, where

$$
\begin{aligned}
c_2^{eff} = -\varepsilon^{4-2\alpha}\frac{Q}{D^2}\zeta(t)\Big\{ & Q\partial_{xx}c\Big(\frac{281}{453600} + \frac{23}{1512}y^2 - \frac{37}{2160}y^4 + \frac{1}{120}y^6 - \frac{1}{672}y^8 - \tilde\beta_1\Big) \\
& -(\partial_{xt}c - D\varepsilon^\alpha\partial_{xxx}c)\Big(-\frac{1}{360}y^6 + \frac{1}{72}y^4 - \frac{7}{360}y^2 - \frac{31}{7560} - \tilde\beta_2\Big)\Big\} \\
+\varepsilon^{4-2\alpha}\frac{k}{D^2}\zeta(t)\Big\{ & Q\partial_x c\Big(\frac{1}{60}y^6 - \frac{1}{18}y^4 + \frac{11}{180}y^2 - \frac{11}{810} - \tilde\beta_3\Big) \\
& + \frac{1}{2}(\partial_t c - D\varepsilon^\alpha\partial_{xx}c)\Big(-\frac{1}{12}y^4 + \frac{1}{6}y^2 - \frac{7}{180} - \tilde\beta_5\Big) \\
& + \frac{Q}{45}\partial_x c\Big(\frac{1}{3} - y^2 - \tilde\beta_4\Big) - \frac{k}{6}c\Big(\frac{1}{3} - y^2\Big)\Big\},
\end{aligned} \tag{139}
$$

where $\tilde\beta_j$, $j = 1, \dots, 5$, are boundary layers analogous to (72).

The application of this additional correction term gives the following.

THEOREM 5. *With the notation of Theorem* 4*, we have*

$$
\begin{aligned}
&\|t^5(c^\varepsilon - c_1^{eff}(x,t;\varepsilon) - \bar c_1^{eff})\|_{L^\infty(0,T;L^2(\Omega_K))} \\
&\le C\big(\varepsilon^{4-13\alpha/4}H(1-\alpha) + \varepsilon^{3(1-\alpha/2)/2}H(\alpha-1)\big);
\end{aligned} \tag{140}
$$

$$
\begin{aligned}
&\|t^5\partial_y(c^\varepsilon - c_1^{eff}(x,t;\varepsilon) - \bar c_1^{eff})\|_{L^2(0,T;L^2(\Omega_K))} \\
&\le C\varepsilon^{1-\alpha/2}\big(\varepsilon^{4-13\alpha/4}H(1-\alpha) + \varepsilon^{3(1-\alpha/2)/2}H(\alpha-1)\big);
\end{aligned} \tag{141}
$$

$$
\begin{aligned}
&\|t^5\partial_x(c^\varepsilon - c_1^{eff}(x,t;\varepsilon) - \bar c_1^{eff})\|_{L^2(0,T;L^2(\Omega_K))} \\
&\le C\varepsilon^{-\alpha/2}\big(\varepsilon^{4-13\alpha/4}H(1-\alpha) + \varepsilon^{3(1-\alpha/2)/2}H(\alpha-1)\big).
\end{aligned} \tag{142}
$$

Proof. Applying the operator $\mathcal{L}^\varepsilon$, given by (34), to $c^\varepsilon - c_1^{eff} - \bar c_1^{eff} - c_2^{eff}$ we obtain a forcing term Φ_2^ε, analogous to Φ^ε, given by (118). To control its behavior we are going to study the expression $\zeta\sum_{j=1}^5 F_j$. As we have seen in Proposition 8, Lemma 5, and Proposition 9, other terms are small. We have the following:

- F_1^ε and F_3^ε are replaced with $\tilde F_1^\varepsilon$ and $\tilde F_3^\varepsilon$, given by

$$
\begin{cases}
\tilde F_1^\varepsilon = (1-y^2)\dfrac{Q^2\varepsilon^{4-2\alpha}}{D^2}\Big\{-\partial_{xxx}cP_8(y)Q \\
\qquad\qquad + (\partial_{xxt}c - D\varepsilon^\alpha\partial_{xxxx}c)P_6(y)\Big\}, \\
\tilde F_3^\varepsilon = -\varepsilon^{4-2\alpha}P_8(y)\dfrac{Q^2}{D^2}\Big\{\partial_{xxt}c - \varepsilon^\alpha\partial_{xxxx}cD\Big\} \\
+\varepsilon^{4-2\alpha}P_6(y)\dfrac{Q}{D^2}\Big\{\partial_{xtt}c - 2D\varepsilon^\alpha\partial_{xxxt}c + \varepsilon^{2\alpha}\partial_{xxxxx}cD^2\Big\},
\end{cases} \tag{143}
$$

where

$$\begin{cases} P_8(y) = \dfrac{281}{453600} + \dfrac{23}{1512}y^2 - \dfrac{37}{2160}y^4 + \dfrac{1}{120}y^6 - \dfrac{1}{672}y^8, \\ P_4(y) = \dfrac{y^2}{6} - \dfrac{y^4}{12} - \dfrac{7}{180};\ P_6(y) = \dfrac{y^3}{18} - \dfrac{y^5}{60} - \dfrac{7y}{180} - \dfrac{31}{7560}. \end{cases}$$

Using (32) we find, analogous to (108)–(109), that

$$\begin{aligned} &\int_0^t \int_0^\infty \int_0^1 \tau^m \zeta(|\tilde{F}_1^\varepsilon| + |\tilde{F}_3^\varepsilon|)|\varphi|\ dxdyd\tau \\ &\le C\big(\varepsilon^{4-13\alpha/4} H(1-\alpha) + \varepsilon^{3/2-3\alpha/4} H(\alpha-1)\big)\|\varphi\|_{L^2(O_t)} \end{aligned} \tag{144}$$

for all $\varphi \in H^1(O_T)$, $\varphi = 0$ at $x = 0$ and $m > 2$.

- F_2^ε and F_4^ε are replaced with

$$\begin{cases} \tilde{F}_2^\varepsilon = (1-y^2)\dfrac{Qk\varepsilon^{4-2\alpha}}{D}\Big\{\partial_{xx}c\dfrac{Q}{D}\tilde{P}_6(y) \\ \quad + \Big(\partial_{xt}c\dfrac{1}{2D} - \varepsilon^\alpha\partial_{xxx}c\dfrac{1}{2}\Big)P_4(y) + \Big(\dfrac{Q}{45D}\partial_x c - \dfrac{k}{6D}c\Big)P_2(y)\Big\}, \\ \tilde{F}_4^\varepsilon = -\varepsilon^{4-2\alpha}\tilde{P}_6(y)\dfrac{Qk}{D^2}\Big\{\partial_{xt}c - \varepsilon^\alpha\partial_{xxx}cD\Big\} \\ \quad + \varepsilon^{4-2\alpha}P_4(y)\dfrac{k}{2D^2}\Big\{\partial_{tt}c - 2D\varepsilon^\alpha\partial_{xxt}c + D^2\varepsilon^{2\alpha}\partial_{xxxx}c\Big\} \\ \quad + \varepsilon^{4-2\alpha}P_2(y)\dfrac{k}{3D^2}\Big\{\dfrac{Q}{15}\partial_{xt}c - \dfrac{k}{2}\partial_t c \\ \qquad\qquad - \dfrac{DQ\varepsilon^\alpha}{15}\partial_{xxx}c + \dfrac{Dk\varepsilon^\alpha}{2}\partial_{xx}c\Big\}, \end{cases} \tag{145}$$

where $P_2(y) = \frac{1}{3} - y^2$ and $\tilde{P}_6 = \frac{y^6}{60} - \frac{y^4}{18} + \frac{11y^2}{180} - \frac{11}{810}$. Using (32) we find, analogous to (110)–(111), that

$$\begin{aligned} &\int_0^t \int_0^\infty \int_0^1 \tau^m \zeta(|\tilde{F}_2^\varepsilon| + |\tilde{F}_4^\varepsilon|)|\varphi|\ dxdyd\tau \\ &\le C\big(\varepsilon^{4-11\alpha/4} H(1-\alpha) + \varepsilon^{5/2-5\alpha/4} H(\alpha-1)\big)\|\varphi\|_{L^2(O_t)} \end{aligned} \tag{146}$$

for all $\varphi \in H^1(O_T)$, $\varphi = 0$ at $x = 0$ and $m > 2$.

- It should be noted that the means of the polynomials in y, contained in $\tilde{F}_1$ and $\tilde{F}_3$, are no longer zero. Hence we cannot gain powers of ε using the derivative with respect to y of the test function.
- F_5 and the boundary term $k\zeta(t)\varepsilon^{2-\alpha}\big(\partial_x c\frac{2Q}{45D} - c\frac{k}{3D}\big)$ are canceled. At the boundary $y = 1$ we have a new nonhomogeneous term

$$\hat{g}^\varepsilon = (1-\zeta)kc - \zeta\varepsilon^{4-2\alpha}\Big(\frac{2Qk^2}{45D^2}\partial_x c\tilde{P}_6|_{y=1} + \Big(\frac{k}{2D^2}\partial_t c - \varepsilon^\alpha\frac{k}{2D}\partial_{xx}c\Big)P_4|_{y=1}\Big), \tag{147}$$

and the principal boundary contribution is given by

$$\begin{aligned}&\left|\int_0^t\int_0^\infty\int_0^1 \tau^m\zeta\varepsilon^{4-2\alpha}\left(\frac{2Qk^2}{45D^2}\partial_x c\tilde{P}_6|_{y=1}\right.\right.\\&\left.\left.+\left(\frac{k}{2D^2}\partial_t c-\varepsilon^\alpha\frac{k}{2D}\partial_{xx}c\right)P_4|_{y=1}\right)\varphi|_{y=1}\;dxdyd\tau\right|\\&\le C\big(\varepsilon^{4-9\alpha/4}H(1-\alpha)+\varepsilon^{7/2-7\alpha/4}H(\alpha-1)\big)\|\varphi|_{y=1}\|_{L^2((0,t)\times(0,+\infty))}.\end{aligned}\tag{148}$$

- Other terms are much smaller and do not have to be discussed.

After collecting the powers of ε and applying Proposition 6 we obtain the estimates (140)–(142). ∎

THEOREM 6. *Let c_2^{eff} be given by* (139). *Then, with the notation of Theorem* 4, *we have*

$$\begin{aligned}&\|t^5(c^\varepsilon-c_1^{eff}(x,t;\varepsilon)-\bar{c}_1^{eff}-c_2^{eff})\|_{L^\infty(0,T;L^1(\Omega_K))}\\&\le C\big(\varepsilon^{4-3\alpha}H(1-\alpha)+\varepsilon^{2-\alpha}H(\alpha-1)\big),\end{aligned}\tag{149}$$

$$\begin{aligned}&\|t^5(c^\varepsilon-c_1^{eff}(x,t;\varepsilon)-\bar{c}_1^{eff}-c_2^{eff})\|_{L^2(0,T;L^2(\Omega_K))}\\&\le C\big(\varepsilon^{4-3\alpha}H(1-\alpha)+\varepsilon^{2-\alpha}H(\alpha-1)\big).\end{aligned}\tag{150}$$

Proof. First we prove the $L^\infty(L^1)$ estimate (149). We test the equation for $\xi=c^\varepsilon-c_1^{eff}(x,t;\varepsilon)-\bar{c}_1^{eff}-c_2^{eff}$ by a regularized sign of ξ, multiplied by Ψ^2, and get

$$\begin{aligned}&t^m\int_{\Omega^+}\Psi(x)^2|\xi|(t)\;dxdy+k\int_0^t\int_0^{+\infty}\tau^m|\xi|_{y=1}|\Psi^2(x)\;dxd\tau\\&\le C_1\int_0^t\int_{\Omega^+}\tau^m\Psi(x)^2|\Phi_2^\varepsilon|\;dxdyd\tau+\int_0^t\int_0^{+\infty}\tau^m|\hat{g}^\varepsilon|_{y=1}|\Psi^2(x)\;dxd\tau|\\&+C_2\varepsilon^\alpha\int_0^t\int_{\Omega^+}\tau^m\Psi(x)^2|\xi|\;dxdyd\tau+m\int_0^t\int_{\Omega^+}\tau^{m-1}|\xi|\Psi^2\;dxdyd\tau\end{aligned}\tag{151}$$

for all $m\ge 3$. As before, the L^1 norm of $\Psi^2\xi$ is Hölder-continuous in time with some exponent $\alpha_0>0$. Arguing as in the proof of Proposition 6, we obtain

$$\sup_{0\le t\le T}\|t^m\Psi^2\xi(t)\|_{L^1(\Omega^+)}\le C(\|\Psi^2\Phi_2^\varepsilon\|_{L^1(\Omega^+\times(0,T))}+\|\Psi^2\hat{g}^\varepsilon|_{y=1}\|_{L^1((0,+\infty)\times(0,T))}),\tag{152}$$

and the estimate (149) is proved.

The improved $L^2(L^2)$ estimate (150) follows from (149), (141), and Poincaré's inequality in H^1 (see, e.g., [6]). ∎

At this point we use Moser's iterations to obtain an $L^\infty(L^\infty)$ error estimate.

THEOREM 7. *Let $O_T=\Omega^+\times(0,T)$. Then, with the notation of Theorem* 6, *we have*

$$\begin{aligned}&\|t^5(c^\varepsilon-c_1^{eff}(x,t;\varepsilon)-\bar{c}_1^{eff}-c_2^{eff})\|_{L^\infty(O_T)}\\&\le C(\delta)\big(\varepsilon^{4-7\alpha/2-\delta}H(1-\alpha)+\varepsilon^{3/2-\alpha-\delta}H(\alpha-1)\big)\quad\forall\delta>0.\end{aligned}\tag{153}$$

Remark 6. From the proof we see that $C(\delta)$ has an exponential growth when $\delta\to 0$.

Proof. Let $M > 0$, $\xi = c^\varepsilon - c_1^{eff}(x,t;\varepsilon) - \bar{c}_1^{eff} - c_2^{eff}$, and $\xi_M = \sup\{t^m\xi - M, 0\}$. We test the equation for ξ by $\Psi^2\xi_M$ and get

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega^+}\Psi(x)^2\xi_M^2(t)\ dxdy + D\varepsilon^\alpha\int_0^t\int_{\Omega^+}\Psi(x)^2|\partial_x\xi_M(\tau)|^2\ dxdyd\tau\\
&+ D\varepsilon^{\alpha-2}\int_0^t\int_{\Omega^+}\Psi(x)^2|\partial_y\xi_M(\tau)|^2\ dxdyd\tau + k\int_0^t\int_0^{+\infty}(\xi_M|_{y=1}\\
&+ M\tau^m)\xi_M|_{y=1}\Psi^2(x)\ dxd\tau \le C_1\left|\int_0^t\int_{\Omega^+}\tau^m\Psi(x)^2|\Phi_3^\varepsilon|\xi_M\ dxdyd\tau\right.\\
&\left.+\int_0^t\int_0^{+\infty}\tau^m|\hat{g}^\varepsilon|_{y=1}\xi_M|_{y=1}\Psi^2(x)\ dxd\tau\right| + C_2\varepsilon^\alpha\int_0^t\int_{\Omega^+}\tau^{2m}\Psi(x)^2\xi_M^2\ dxdyd\tau
\end{aligned} \tag{154}$$

for all $m \ge 3$, where $\tau^m\Phi_3^\varepsilon = -\tau^m\Phi_2^\varepsilon + m\tau^{m-1}\xi$. We suppose that

$$kM \ge \sup_{0\le\tau\le T}\tau^m\|\Psi\hat{g}^\varepsilon(\tau)|_{y=1}\|_{L^\infty(0,+\infty)} = c_0\big(\varepsilon^{4-5\alpha/2}H(1-\alpha) + \varepsilon^{3(1-\alpha/2)}H(\alpha-1)\big). \tag{155}$$

As in the classical derivation of the Nash–Moser estimate (see [10, pp. 181–186]) we introduce

$$\mu(M) = \int_0^T\int_{\Omega^+\cap\{t^m\xi - M > 0\}}\Psi^2\ dxdydt. \tag{156}$$

Now in exactly the same way as in [10, pp. 181–186], on a time interval which could be smaller than $[0,T]$, but assumed equal to it without losing the generality, we get

$$\begin{aligned}
&\|\xi_M\|_{V_2}^2 = \sup_{0\le t\le T}\int_{\Omega^+}\Psi(x)^2\xi_M^2(t)\ dxdy + D\varepsilon^\alpha\int_0^T\int_{\Omega^+}\Psi(x)^2|\partial_x\xi_M(\tau)|^2\ dxdyd\tau\\
&+ D\varepsilon^{\alpha-2}\int_0^T\int_{\Omega^+}\Psi(x)^2|\partial_y\xi_M(\tau)|^2\ dxdyd\tau \le \beta_0^2\|\tau^m\Phi_3^\varepsilon\Psi\|_{L^q(O_T)}^2\mu(M)^{1-2/q},\\
&q > 2.
\end{aligned} \tag{157}$$

Next, the estimate (157) is iterated in order to conclude that $\xi_M = 0$. Here we modify the classical argument from [10, pp. 102–103] and adapt it to our situation.

We note that, after making appropriate extensions, one finds

$$\|\Psi\varphi\|_{L^4(O_T)} \le c_0\|\Psi\varphi\|_{L^2(O_T)}^{1/2}\|\Psi\varphi\|_{H^1(O_T)}^{1/2} \le c_0\varepsilon^{-\alpha/4}\|\varphi\|_{V_2} \tag{158}$$

for all $\varphi \in V_2$, such that $\varphi|_{x=0} = 0$. As in [10, p. 102], at this point we take the sequence of levels $k_h = M(2-2^{-h})$, $h = 0,1,\dots$. Then

$$(k_{h+1} - k_h)\mu^{1/4}(k_{h+1}) \le \|\Psi\xi_{k_h}\|_{L^4(O_T)} \le \frac{\bar{\beta}\varepsilon^{-\alpha/4}}{k_{h+1}-k_h}\|\xi_{k_h}\|_{V_2} \tag{159}$$

and, with $\kappa = 1 - 2/q > 0$,

$$\mu^{1/4}(k_{h+1}) \le 2^h\frac{2\bar{\beta}\beta_0\|\tau^m\Phi_3^\varepsilon\Psi\|_{L^q(O_T)}\varepsilon^{-\alpha/4}}{M}\mu^{(1+\kappa)/4}(k_h). \tag{160}$$

$\mu^{1/4}(k_{h+1})$ will tend to zero for $h \to \infty$ if $\mu^{1/4}(M)$ satisfies

$$\mu^{1/4}(M) \le \left(\frac{2\bar{\beta}\beta_0 \|\tau^m \Phi_3^\varepsilon \Psi\|_{L^q(\Omega^+ \times (0,T))} \varepsilon^{-\alpha/4}}{M} \right)^{-1/\kappa} 2^{-1/\kappa^2}. \tag{161}$$

Equation (161) is satisfied if M equals the right-hand side of estimate (153). ∎

Our next result concerns higher order norms. It it not very satisfactory for large α and we state it without giving the proof, which follows from the demonstrations given above.

THEOREM 8. *With the notation of Theorem* 4, *we have*

$$\begin{aligned} &\|t^5 \partial_x (c^\varepsilon - c_1^{eff}(x,t;\varepsilon) - \bar{c}_1^{eff})\|_{L^\infty(0,T;L^2(\Omega_K))} \\ &\le C\big(\varepsilon^{4-15\alpha/4} H(1-\alpha) + \varepsilon^{(1-\alpha/2)/2} H(\alpha-1)\big), \end{aligned} \tag{162}$$

$$\begin{aligned} &\|t^5 \partial_t (c^\varepsilon - c_1^{eff}(x,t;\varepsilon) - \bar{c}_1^{eff})\|_{L^2(0,T;L^2(\Omega_K))} \\ &\le C\big(\varepsilon^{4-15\alpha/4} H(1-\alpha) + \varepsilon^{(1-\alpha/2)/2} H(\alpha-1)\big). \end{aligned} \tag{163}$$

Our final improvement concerns the error estimate in the $L^\infty(L^2)$ norm for small values of α. As mentioned in the proof of Theorem 5, $\tilde{F}_1^\varepsilon$ and $\tilde{F}_3^\varepsilon$ do not have zero means with respect to y, and consequently, the proof does not give the expected precision for the second-order correction. Nevertheless, when computing the term c^2 in the asymptotic expansion, there is a liberty of adding an arbitrary function C_2 of x and t. This function can be chosen such that the appropriate means are zero and the left-hand sides of the estimates (140)–(142) are multiplied by $\varepsilon^{1-\alpha/2}$. Unfortunately, there is a simultaneous new contribution of the form $QP_2(y)\partial_x C_2$. Its norm deteriorates the estimate for $\alpha \ge 4/5$. Consequently, this amelioration is not of real importance and we just give the result. The proof is completely analogous to the preceding ones.

COROLLARY 4. *Let the polynomials $P_j(y)$ be defined by* (143)–(145) *and let C_2 be given by the initial/boundary value problem*

(164)

$$\begin{aligned} &\frac{\partial C_2}{\partial t} + \frac{2Q}{3}\frac{\partial C_2}{\partial x} - \varepsilon^\alpha D \frac{\partial^2 C_2}{\partial x^2} \\ &= -\frac{Qk}{D}\zeta(t)\bigg\{ \partial_{xx} c \frac{Q}{D} \int_0^1 (1-y^2)\tilde{P}_6(y)\, dy + \left(\partial_{xt} c \frac{1}{2D} - \varepsilon^\alpha \partial_{xxx} c \frac{1}{2} \right) \int_0^1 (1-y^2) P_4(y)\, dy \\ &\qquad + \left(\frac{Q}{45D} \partial_x c - \frac{k}{6D} c \right) \int_0^1 (1-y^2) P_2(y)\, dy \bigg\} \\ &\quad - \frac{Q^2}{D^2}\zeta(t)\bigg\{ -\partial_{xxx} c Q \int_0^1 (1-y^2) P_8(y)\, dy \\ &\qquad + (\partial_{xxt} c - D\varepsilon^\alpha \partial_{xxxx} c) \int_0^1 (1-y^2) P_6(y)\, dy \bigg\} \quad in \quad (0,+\infty) \times (0,T), \end{aligned}$$

$$\partial_x C_2 \in L^2((0,+\infty)\times(0,T)), \quad C_2|_{t=0} = 0, \quad C_2|_{x=0} = 0. \tag{165}$$

Then, with the notation of Theorem 6 *and for $\alpha \in [0, 4/5]$ we have*

$$\|t^5 (c^\varepsilon - c_1^{eff} - \bar{c}_1^{eff} - c_2^{eff} - C_2)\|_{L^\infty(0,T;L^2(\Omega_K))} \le C\varepsilon^{5-17\alpha/4}, \tag{166}$$

$$\|t^5 \partial_y (c^\varepsilon - -c_1^{eff} - \bar{c}_1^{eff} - c_2^{eff})\|_{L^2(0,T;L^2(\Omega_K))} \le C\varepsilon^{6-19\alpha/4}, \tag{167}$$

$$\|t^5 \partial_x (c^\varepsilon - c_1^{eff} - \bar{c}_1^{eff} - c_2^{eff} - C_2)\|_{L^2(0,T;L^2(\Omega_K))} \le \varepsilon^{5-19\alpha/4}. \tag{168}$$

Acknowledgments. The authors are grateful to I. S. Pop for his assistance and they acknowledge the referees for their valuable comments.

REFERENCES

[1] R. Aris, *On the dispersion of a solute in a fluid flowing through a tube*, Proc. Roy. Soc. London Ser. A, 235 (1956), pp. 67–77.
[2] L. V. Berlyand and M. V. Goncharenko, *The averaging of the diffusion equation in porous medium with weak absorption*, J. Soviet Math., 53 (1990), pp. 3428–3435.
[3] A. Bourgeat, M. Jurak, and A. L. Piatnitski, *Averaging a transport equation with small diffusion and oscillating velocity*, Math. Methods Appl. Sci., 26 (2003), pp. 95–117.
[4] R. E. Caflisch and J. Rubinstein, *Lectures on the Mathematical Theory of Multiphase-Flow*, Courant Institute of Mathematical Sciences, New York, 1984.
[5] C. J. van Duijn and I. S. Pop, *Crystal dissolution and precipitation in porous media: Pore scale analysis*, J. Reine Angew. Math., 577 (2004), pp. 171–211.
[6] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[7] E. Grenier, *Boundary layers for parabolic regularization of totally characteristic quasilinear parabolic equations*, J. Math. Pures Appl. (9), 76 (1997), pp. 965–990.
[8] E. Grenier and O. Guès, *Boundary layers for viscous perturbations of noncharacteristic quasilinear hyperbolic problems*, J. Differential Equations, 143 (1998), pp. 110–146.
[9] E. Grenier and F. Rousset, *Stability of one-dimensional boundary layers by using Green's functions*, Comm. Pure Appl. Math., 54 (2001), pp. 1343–1385.
[10] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasi-Linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[11] A. J. Majda and P. R. Kramer, *Simplified models for turbulent diffusion: Theory, numerical modelling, and physical phenomena*, Phys. Rep., 314 (1999), pp. 237–574.
[12] R. Mauri, *Dispersion, convection, and reaction in porous media*, Phys. Fluids A, 3 (1991), pp. 743–756.
[13] G. N. Mercer and A. J. Roberts, *A centre manifold description of contaminant dispersion in channels with varying flow properties*, SIAM J. Appl. Math., 50 (1990), pp. 1547–1565.
[14] O. A. Oleinik and G. A. Iosif'jan, *On the behavior at infinity of solutions of second-order elliptic equations in domains with noncompact boundary*, Math. USSR Sb., 40 (1981), pp. 527–548.
[15] M. A. Paine, R. G. Carbonell, and S. Whitaker, *Dispersion in pulsed systems* I. *Heterogeneous reaction and reversible adsorption in capillary tubes*, Chem. Eng. Sci., 38 (1983), pp. 1781–1793.
[16] A. I. Piatniski, *Averaging singularly perturbed equation with rapidly oscillating coefficients in a layer*, Math. USSR-Sb., 49 (1984), pp. 19–40.
[17] A. I. Piatniski, *Personal communication*, 2005.
[18] J. Rubinstein and R. Mauri, *Dispersion and convection in periodic porous media*, SIAM J. Appl. Math., 46 (1986), pp. 1018–1023.
[19] G. I. Taylor, *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. London Ser. A, 219 (1953), pp. 186–203.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1288–1315

THE KELLER–SEGEL MODEL FOR CHEMOTAXIS WITH PREVENTION OF OVERCROWDING: LINEAR VS. NONLINEAR DIFFUSION*

MARTIN BURGER†, MARCO DI FRANCESCO‡, AND YASMIN DOLAK-STRUSS§

Abstract. The aim of this paper is to discuss the effects of linear and nonlinear diffusion in the large time asymptotic behavior of the Keller–Segel model of chemotaxis with volume filling effect. In the linear diffusion case we provide several sufficient conditions for the diffusion part to dominate and yield decay to zero solutions. We also provide an explicit decay rate towards self–similarity. Moreover, we prove that no stationary solutions with positive mass exist. In the nonlinear diffusion case we prove that the asymptotic behavior is fully determined by whether the diffusivity constant in the model is larger or smaller than the threshold value $\varepsilon = 1$. Below this value we have existence of nondecaying solutions and their convergence (along subsequences) to stationary solutions. For $\varepsilon > 1$ all compactly supported solutions are proved to decay asymptotically to zero, unlike in the classical models with linear diffusion, where the asymptotic behavior depends on the initial mass.

Key words. chemotaxis, nonlinear diffusion, asymptotic behavior, Keller–Segal model, overcrowding

AMS subject classifications. 92C17, 35K55, 35K57, 35K65, 35B40

DOI. 10.1137/050637923

1. Introduction.

This paper focuses on the mathematical analysis of a chemotaxis model in the cases of linear and nonlinear diffusion. The general structure of the model we consider is

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = \nabla \cdot (M(\rho) \nabla \mu(\rho, S)) \\ -\Delta S + S = \rho, \end{cases} \tag{1.1}$$

posed on $\mathbb{R}^d \times \mathbb{R}^+$ subject to the initial condition

$$\rho(x,0) = \rho_0(x), \quad \rho_0 \in L^1(\mathbb{R}^d), \quad 0 \le \rho_0(x) \le 1, \quad x \in \mathbb{R}^d. \tag{1.2}$$

The mobility term M is given by $M(\rho) = \rho(1-\rho)$. Such a choice takes into account the prevention of the overcrowding effect, sometimes also referred to as "volume filling" effect (see the motivations and references below). The potential μ reads

$$\mu(\rho, S) = \frac{\delta E}{\delta \rho}(\rho, S), \tag{1.3}$$

*Received by the editors August 10, 2005; accepted for publication (in revised form) June 21, 2006; published electronically December 15, 2006.
http://www.siam.org/journals/sima/38-4/63792.html

†Industrial Mathematics Institute, Johannes Kepler University, Altenbergerstr. 69, A-4040 Linz, Austria (martin.burger@jku.at). This author has been supported by the Austrian National Science Foundation FWF through project SFB F 013 / 08 and by the Johann Radon Institute for Computational and Applied Mathematics (Austrian Academy of Sciences ÖAW).

‡Sezione di Matematica per l'Ingegneria, Dipartimento di Matematica Pura ed Applicata, Università di L'Aquila, Piazzale E. Pontieri, 2, Monteluco di Roio, 67040 L'Aquila, Italy (difrance@univaq.it). This author has been partially supported by the Wittgenstein 2000 Award of Peter A. Markowich and by the EU–funded IHP-project "HYKE."

§Faculty of Mathematics, University of Vienna, Nordbergstr. 15, 1090 Vienna, Austria (yasmin.dolak-struss@univie.ac.at).

where $\frac{\delta E}{\delta \rho}$ denotes the functional derivative of some energy functional with respect to ρ. If we model the energy as a combination of a logarithmic entropy and an aggregation part, i.e.,

$$\begin{aligned} E(\rho, S) = \varepsilon \int_{\mathbb{R}^d} (\rho \log \rho + (1-\rho)\log(1-\rho))\ dx \\ - \int_{\mathbb{R}^d} \rho S\ dx + \frac{1}{2} \int_{\mathbb{R}^d} \left(|\nabla S|^2 + S^2\right)\ dx, \end{aligned} \tag{1.4}$$

with $\varepsilon > 0$, then system (1.1) becomes

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = \nabla \cdot (\varepsilon \nabla \rho - \rho(1-\rho)\nabla S) \\ -\Delta S + S = \rho, \end{cases} \tag{1.5}$$

which is a special case of the Keller–Segel model for chemotaxis, describing the behavior of a cell population ρ under the influence of the chemical S produced by the cells themselves. Introduced in 1970 [KS70] to describe aggregation of slime mold amoebae, this model has become one of the most widely studied models in mathematical biology. The cell flux on the right-hand side of (1.5) comprises two counteracting phenomena: random motion of cells described by Fick's law and cell movement in the direction of the gradient of the chemical S. In contrast to the equations presented here, the gradient of S is multiplied by a linear instead of a nonlinear function of ρ in the classical version of the model. An interesting feature of this choice is the fact that solutions can become unbounded in finite time, thus giving rise to concentration phenomena. Whether this blow-up of solutions occurs or not depends typically on the initial data and the space dimension d, and conditions for the blow-up of solutions have been derived by many authors (see [Hor03, Hor04] and the references therein). Most studies focus on models where the evolution of the chemical S is governed by a parabolic equation (as in the original Keller–Segel model). Typical alternative formulations for the evolution of S are given either by an elliptic equation of a more general form than in (1.1) or by the Poisson equation. In the majority of the cases, the model is considered on bounded domains, typically with Neumann boundary conditions. Concerning the case of an unbounded domain, an extensive analysis of the model on $\mathbb{R}^d$, $d \geq 2$, has been performed recently in [CPZ04], where the authors prove that solutions exist globally when the $L^{d/2}$-norm of the initial data does not exceed a critical value. In the recent contribution [DP04] it is proved that the critical value 8π of the initial mass produces an optimal threshold between blow-up and global existence when $d = 2$.

Although of great mathematical interest, models allowing for the infinite growth of solutions have often been criticized because their biological interpretation is not fully understood. Several generalizations of the Keller–Segel model, where the formation of singularities is prevented *a priori*, have been studied recently (see, e.g., [HP01]). These models superimpose a maximal value for the cell density, which yields the global existence of solutions. Such an assumption is reasonable in certain physical situations where cells stop aggregating after a certain size of the aggregate has been reached. The model presented in [HP01] is of type (1.5) but with a parabolic equation for S. Note that the main difference from a standard derivation of the model without prevention of overcrowding is the additional factor $(1-\rho)$ in the mobility term $M(\rho)$ in (1.1) and the additional entropy term depending on $(1-\rho)$ in the energy. Intuitively, this change of mobility seems obvious, since the possibility of cells to move freely

is limited by the cells around them. Therefore, in the case of overcrowding (which happens in our scaling for $\rho \geq 1$) their motion would be stopped. It is less obvious how to interpret the additional entropy term depending on $(1-\rho)$—it somehow forces cells to diffuse even in the case of overcrowding. On the other hand if the volume-filling mechanism is thought of as a finite size effect really blocking the cells, then also no diffusion would be allowed and the volume-filling effect would have consequences only upon mobility but not on the energy. Such a possibility is supported by results in [BO04], where the authors observe that many chemicals appear both to stimulate directed motion up the chemotactic gradient and to alter the random mobility (i.e., the nonlinear diffusion coefficient). As a consequence of that, they establish a direct relationship between the chemotaxis and random motion coefficients. In particular, they both may vanish at the threshold value of the density.

This argument is the main reason why we shall also discuss a different choice for the energy in this paper. More precisely, we shall use a quadratic energy of the form

$$E(\rho, S) = \frac{\varepsilon}{2} \int_{\mathbb{R}^d} \rho^2 \, dx - \int_{\mathbb{R}^d} \rho S \, dx + \frac{1}{2} \int_{\mathbb{R}^d} \left(|\nabla S|^2 + S^2 \right) \, dx, \tag{1.6}$$

yielding the degenerate parabolic-elliptic problem

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = \nabla \cdot (\rho(1-\rho)\nabla(\varepsilon\rho - S)) \\ -\Delta S + S = \rho. \end{cases} \tag{1.7}$$

Several reformulations of the Keller–Segel model have been studied recently, with a nonlinear diffusion term in the evolution of the cell density replacing the linear one (see [Hor03, Chapter 6]). An interesting issue is whether it is possible to avoid blow-up of solutions by introducing nonlinear diffusion effects without the help of the volume-filling term in the mobility (see also [CC05]). We stress that the main feature in our nonlinear model (1.7) is that the random mobility vanishes at the threshold value $\rho = 1$. Such a property is not usually covered in the literature concerning chemotaxis models with nonlinear diffusion.

A further reason to study (1.7) relates to the asymptotic behavior of its solutions, in particular to the possibility of achieving either existence of nontrivial stationary solutions or a large time decay to zero in L^∞. Whether the former or the latter phenomenon takes place depends upon the size of the diffusivity ε. We recall that the long-time asymptotics of the linear case (1.5) in bounded intervals (under Neumann boundary conditions) have been recently studied in [DS05] and, with a parabolic equation for S, in [PH05]. The observed behavior is a coarsening process reminiscent of phase change models, where plateau-like peaks of the cell density form after a short transition period and then merge exponentially slowly. Numerical studies indicate that in most situations the only stable stationary states are single plateaus located at the boundary of the domain. It is therefore not surprising that the behavior of solutions on the whole space is different. Roughly speaking, the cells are not stopped by any boundaries and therefore the linear model would allow them to spread out rather than aggregate. We shall make this statement more rigorous by several results on the decay of the cell density ρ for large time. For now we emphasize that, to our knowledge, system (1.7) is the only one known to admit stationary profiles on the whole space.

In order to make the comparison between the linear and nonlinear diffusion case more concrete, we provide a guideline through the paper for readers who are interested

in the results rather than on the detailed proofs. The first step towards the large time behavior of solutions is an investigation of possible stationary solutions. Here we find a first difference between the two cases. More precisely, no finite mass stationary solutions different from zero exist for linear diffusion (subsection 2.2.1), while finite mass nontrivial stationary solutions do exist for nonlinear diffusion in the case $\varepsilon < 1$ (subsection 3.4.2). A partly numerical construction of stationary solutions even indicates their existence for arbitrary mass if $\varepsilon < 1$ (section 3.5). The detailed large-time behavior is described by:

- Decay to zero of solutions in the linear diffusion case in the following two cases:
  - with arbitrary initial mass for $\varepsilon > \frac{1}{4}$ in 1–d (section 2.2),
  - with small initial mass for arbitrary $\varepsilon > 0$ (section 2.2).

  In both cases, solutions converge in L^1 towards the self-similar Gaussian solution of the heat equation with variance ε (section 2.3).
- Existence of nondecaying solutions for $\varepsilon < 1$ in the nonlinear case (subsection 3.4.1). These solutions converge (along subsequences) to stationary solutions (subsection 3.4.2). For $\varepsilon > 1$ all compactly supported solutions decay and their support must become unbounded as $t \to \infty$ (subsection 3.4.3).

2. Linear diffusion. In this section, we first cover the existence and uniqueness theory for weak solutions of model (1.5). We then turn our attention to the asymptotic behavior for large time.

2.1. Existence theory and preliminaries. We study the Cauchy problem for the following parabolic-elliptic system

$$\begin{cases} \frac{\partial \rho}{\partial t} = \varepsilon \Delta \rho - \nabla \cdot (\rho(1-\rho)\nabla S) \\ -\Delta S + S = \rho, \\ \rho(x,0) = \rho_0(x), \end{cases} \tag{2.1}$$

where $x \in \mathbb{R}^d$, $d \geq 1$, $t \geq 0$, $\varepsilon > 0$, ρ, and S are scalar functions, and ρ_0 belongs in $L^1(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$. We recall that the above system can be decoupled in order to get a nonlocal parabolic equation for ρ by means of the convolution representation formula

$$S(x,t) = \int_{\mathbb{R}^d} \mathcal{B}(x-y)\rho(y,t)dy, \tag{2.2}$$

where $\mathcal{B}$ is the *Bessel potential*

$$\mathcal{B}(x) = \frac{1}{(4\pi)^{d/2}} \int_0^{+\infty} \frac{e^{-t-\frac{|x|^2}{4t}}}{t^{d/2}} dt. \tag{2.3}$$

For further reference we recall the formula for the heat convolution kernel

$$\mathcal{G}(x,t) = \frac{1}{(4\pi\varepsilon t)^{d/2}} e^{-\frac{|x|^2}{4\varepsilon t}}, \quad x \in \mathbb{R}^d,\ t > 0. \tag{2.4}$$

The proof of the following lemma follows by straightforward computations.

LEMMA 2.1. *The following estimates are satisfied:*

$$\|\mathcal{B}\|_{L^1(\mathbb{R}^d)} = 1, \tag{2.5}$$

$$\|\nabla \mathcal{B}\|_{L^1(\mathbb{R}^d)} < +\infty, \tag{2.6}$$

$$\|\mathcal{G}(\cdot,t)\|_{L^p(\mathbb{R}^d)} \leq Ct^{-\frac{d(p-1)}{2p}}, \quad p \geq 1, \tag{2.7}$$

$$\|\nabla \mathcal{G}(\cdot,t)\|_{L^p(\mathbb{R}^d)} \leq Ct^{-\frac{d(p-1)}{2p}-\frac{1}{2}}, \quad p \geq 1, \tag{2.8}$$

The estimates provided by the previous lemma enable us to prove local existence of $L^1 \cap L^\infty$ solutions of (2.1) by means of the implicit representation formula

$$\rho(x,t) = (\mathcal{G} * \rho_0)(x,t) + \int_0^t \int_{\mathbb{R}^d} \nabla \mathcal{G}(x-y, t-s)\, (\rho(1-\rho)\nabla(\mathcal{B} * \rho))(y,s) dsdy \tag{2.9}$$

and by a standard fixed point argument. This leads to the following theorem, whose proof is rather standard and will be omitted (see also [Hor03] and the references therein for the local existence theory of other models closely related to (2.1)).

THEOREM 2.2 (local existence). *Let $\rho_0 \in L^1(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ (resp., $\rho_0 \in L^\infty(\mathbb{R}^d)$). Then, there exists a unique solution $\rho(x,t)$ to* (2.1) *such that*

$$\rho \in L^\infty\left([0,T], L^1(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)\right)$$

(resp., $\rho \in L^\infty\left([0,T], L^\infty(\mathbb{R}^d)\right)$) for a small enough positive time T.

Again by standard fixed point technique, the local-in-time solution $\rho(x,t)$ provided by Theorem 2.2 can be endowed with the same regularity (with respect to x) as the initial datum. Some regularity for ρ without the help of any further requirements on the initial datum can be obtained at least in 1–d.

PROPOSITION 2.3 (regularizing effect). *Let the initial datum $\rho_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. Then, at any positive time t, the solution $\rho(t)$ of* (2.1) *is continuous with respect to x.*

Proof. Since $\mathcal{G} * \rho_0$ is a C^∞ function, we only need to prove that $\rho - \mathcal{G} * \rho_0$ is continuous. By formula (2.9), we have for small $h > 0$

$$\begin{aligned}&(\rho - \mathcal{G} * \rho_0)(x+h,t) - (\rho - \mathcal{G} * \rho_0)(x,t)\\ &= \int_0^t \int_{-\infty}^{+\infty} \left[\mathcal{G}_x(x+h-y, t-s) - \mathcal{G}_x(x-y,t-s)\right] (\rho(1-\rho)\nabla(\mathcal{B} * \rho))(y,s) dsdy.\end{aligned}$$

We recall that the term $(\rho(1-\rho)\nabla(\mathcal{B} * \rho))$ is (locally) bounded. Moreover, thanks to (2.8) we can find nonnegative functions $H(x,y,t,s)$ and $K \in L^1$ such that

$$|\mathcal{G}_x(x+h-y,t-s) - \mathcal{G}_x(x-y,t-s)| \le H(x,y,t,s) \le (t-s)^{-1} K\left(\frac{|x-y|^2}{t-s}\right).$$

Therefore, the limit as $h \to 0$ in the above integral is zero due to Lebesgue's dominated convergence theorem. ∎

The representation formula (2.9) is also the basis of the result in the following proposition, the proof of which is again omitted.

PROPOSITION 2.4 (continuity with respect to the initial data). *Let ρ and $\overline{\rho}$ be two local-in-time solutions to* (2.1)*, with initial data $\rho_0, \overline{\rho}_0 \in L^1 \cap L^\infty$, respectively. Then, for a small T we have*

$$\|\rho(t) - \overline{\rho}(t)\|_{L^1} \le C(T) \|\rho_0 - \overline{\rho}_0\|_{L^1} \tag{2.10}$$

for all $t \in [0,T]$.

As suggested by the divergence form, L^1 solutions of (2.1) preserve the total mass, i.e.,

$$\int_{\mathbb{R}^d} \rho(x,t) dx = \int_{\mathbb{R}^d} \rho_0(x) dx. \tag{2.11}$$

The proof of the above identity is standard (see, for instance, [Váz92, Chapter III–section 2]).

In order to achieve global existence of solutions, from now on in this section we shall restrict our analysis to the case of initial data ρ_0 in (2.1) satisfying the assumption

$$0 \le \rho_0(x) \le 1, \quad x \in \mathbb{R}^d. \tag{2.12}$$

Our next aim is to prove that condition (2.12) is invariant under the flow induced by the model (2.1). Such a property together with the conservation of the total mass will imply the global existence of the solution of (2.1). Similar properties are proved in [HP01] for the fully parabolic model and in [DS05] in a bounded domain.

THEOREM 2.5 (global existence). *Assume the initial datum $\rho_0 \in L^1(\mathbb{R}^d)$ satisfies* (2.12)*. Then there exists a unique global solution to the Cauchy problem* (2.1)*, satisfying*

$$0 \le \rho(x,t) \le 1 \quad \textit{for any} \quad (x,t) \in \mathbb{R}^d \times [0,\infty). \tag{2.13}$$

In particular,

$$0 < \rho(x,t) < 1 \quad \textit{if} \quad 0 < \rho_0(x) < 1. \tag{2.14}$$

Proof. Writing (2.1) as $\rho_t + \nabla\rho \cdot \nabla S(1-2\rho) + \rho(1-\rho)(S-\rho) = \varepsilon\Delta\rho$, it can be seen immediately that $\underline{\rho} \equiv 0$ and $\overline{\rho} \equiv 1$ are lower and upper solutions, respectively. By the mean value theorem of multidimensional calculus, the function $\underline{w}(x,t) = \rho - \underline{\rho}$ satisfies the inequality

$$\underline{w}_t + A(x,t)\nabla\underline{w} + B(x,t)\underline{w} - \varepsilon\Delta\underline{w} \ge 0,$$

with bounded coefficients $A(x,t)$ and $B(x,t)$. For $\overline{w}(x,t) = \rho - \overline{\rho}$, the same equations with a reversed inequality sign holds, and the boundedness of ρ follows from the Phragmèn–Lindelöf principle for parabolic equations [PW84]. Therefore the proof is complete in view of the conservation of the total mass. ∎

2.2. Decay of solutions as $t \to +\infty$. In this subsection we determine sufficient conditions on the diffusivity constant ε in (2.1) and on the initial datum ρ_0 such that the $L^\infty(\mathbb{R}^d)$ norm of the corresponding solution $\rho(x,t)$ tends to zero as $t \to +\infty$. We prove here that this is the case either when the total mass is smaller than a threshold value depending on ε or without any condition on the total mass in case $\varepsilon > 1/4$ and $d = 1$. It is already known in case of models without prevention of overcrowding that, when the mass of the initial datum is much smaller than some constant depending on ε, the diffusion term becomes dominant and produces a long time decay of the solution (see [CPZ04, DP04]). For the sake of completeness we shall reproduce the same result for our model in the following proposition, the proof of which is partly the same as in [CPZ04].

PROPOSITION 2.6. *Let $\varepsilon > 0$ and let $\rho_0 \in L^1(\mathbb{R}^d)$ satisfy* (2.12)*. Then, there exists a constant $C(d)$ depending only on the dimension d such that, for total mass satisfying*

$$\int_{\mathbb{R}^d} \rho_0 dx < \left(\frac{4\varepsilon}{C(d)}\right)^{1/\beta} \qquad \beta = \min\{1, 2/d\}. \tag{2.15}$$

The solution $\rho(x,t)$ to (2.1) satisfies the decay estimates

$$\|\rho(t)\|_{L^p(\mathbb{R}^d)} \le C(t+1)^{-\frac{d(p-1)}{2p}} \quad \text{if } 2 \le p < +\infty, \tag{2.16}$$

$$\|\rho(t)\|_{L^\infty(\mathbb{R}^d)} \le C(t+1)^{-\frac{d}{2}}. \tag{2.17}$$

Proof. Let us start with the L_p estimates for finite p. By multiplying the equation in (2.1) by $p\rho^{p-1}$ and after integration by parts we obtain

$$\begin{aligned}
&\frac{d}{dt}\int_{\mathbb{R}^d} \rho^p(x,t)dx = \varepsilon p \int_{\mathbb{R}^d} \rho^{p-1}\Delta\rho dx + p(p-1)\int_{\mathbb{R}^d}(\rho^{p-1}-\rho^p)\nabla S\cdot\nabla\rho dx\\
&\le -\varepsilon\frac{4(p-1)}{p}\int_{\mathbb{R}^d}|\nabla\rho^{p/2}|^2dx + (p-1)\int_{\mathbb{R}^d}\nabla\rho^p\cdot\nabla S dx\\
&\le -\varepsilon\frac{4(p-1)}{p}\int_{\mathbb{R}^d}|\nabla\rho^{p/2}|^2dx + (p-1)\int_{\mathbb{R}^d}\rho^{p+1}dx,
\end{aligned}$$

where we have used the a priori estimate $0 \le \rho \le 1$ and $S \ge 0$. By means of the Gagliardo–Nirenberg inequality (see also [CPZ04])

$$\int_{\mathbb{R}^d}\rho^{p+1}dx \le C(d)\|\rho\|_{L^\alpha(\mathbb{R}^d)}\int_{\mathbb{R}^d}|\nabla\rho^{p/2}|^2dx,$$

where $\alpha = 1$ for $d = 1, 2$, $\alpha = d/2$ for $d > 2$, we easily get

$$\frac{d}{dt}\int_{\mathbb{R}^d}\rho^p(x,t)dx \le -(p-1)\left(\frac{4\varepsilon}{p} - C(d)\left(\int_{\mathbb{R}^d}\rho_0 dx\right)^\beta\right)\int_{\mathbb{R}^d}|\nabla\rho^{p/2}|^2dx,$$

where $\beta = \min\{1, 2/d\}$. Hence, for

$$\int_{\mathbb{R}^d}\rho_0 dx < \left(\frac{4\varepsilon}{pC(d)}\right)^{1/\beta}, \tag{2.18}$$

we can write, for some $C > 0$,

$$\frac{d}{dt}\int_{\mathbb{R}^d}\rho^p(x,t)dx + C\int_{\mathbb{R}^d}|\nabla\rho^{p/2}|^2dx \le 0.$$

Thanks to the following interpolation inequality (see, e.g., [EZ91]):

$$\|\rho\|_{L^p(\mathbb{R}^d)}^{\frac{(d(p-1)+2)p}{d(p-1)}} \le \overline{C}(p,d)\|\nabla\rho^{p/2}\|_{L^2(\mathbb{R}^d)}^2\|\rho\|_{L^1(\mathbb{R}^d)}^{\frac{2p}{d(p-1)}},$$

we have

$$\frac{d}{dt}\int_{\mathbb{R}^d}\rho^p(x,t)dx + Cm\left(\int_{\mathbb{R}^d}\rho^p dx\right)^{\frac{d(p-1)+2}{d(p-1)}} \le 0,$$

which implies the desired polynomial time-decay in L^p

$$\|\rho(t)\|_{L^p(\mathbb{R}^d)} \le C(p,d)(t+1)^{-d(p-1)/2p}, \tag{2.19}$$

where the term $(t+1)$ instead of t is justified by the global-in-time control of all the L^p norms proved in the previous subsection. We remark that the previous L^p decay

has been obtained for solutions with initial data satisfying (2.18) where the constant on the right-hand side depends on p, unlike in the statement of Theorem 2.5. We shall recover the threshold condition (2.15) later on in this proof. In order to obtain the L^∞ estimate (2.17), we employ the implicit representation of the solution ρ provided by the Duhamel principle

$$\begin{aligned}\rho(x,2t) &= \mathcal{G}(t) * \rho(t) + \int_t^{2t} \nabla\mathcal{G}(2t-\sigma) * (\rho(1-\rho)\nabla\mathcal{B} * \rho)(\sigma)d\sigma \\ &= \mathcal{G}(2t) * \rho(t) + \int_0^t \nabla\mathcal{G}(t-s) * (\rho(1-\rho)\nabla\mathcal{B} * \rho)(t+s)ds. \qquad (2.20)\end{aligned}$$

By taking the L^∞ norm in (2.20), we obtain

$$\begin{aligned}&\|\rho(2t)\|_{L^\infty(\mathbb{R}^d)} \le \|\mathcal{G}(t)\|_{L^\infty(\mathbb{R}^d)}\|\rho(t)\|_{L^1(\mathbb{R}^d)} \\ &+ \int_0^t \|\nabla\mathcal{G}(t-s)\|_{L^r(\mathbb{R}^d)}\|\rho(1-\rho)\nabla\mathcal{B} * \rho(t+s)\|_{L^{r'}(\mathbb{R}^d)}ds \\ &\le Ct^{-\frac{d}{2}} + C\int_0^t (t-s)^{-\frac{d(r-1)}{2r}-\frac{1}{2}}\|\rho(t+s)\|^2_{L^{2r'}(\mathbb{R}^d)}ds, \qquad r' = \frac{r}{r-1}.\end{aligned}$$

Here $r > 1$ can be chosen arbitrarily with the only restriction $r < \frac{d}{d-2}$ in case $d > 2$. Once r and r' are fixed, we require that ρ_0 satisfies (2.18) with $p = 2r'$. Hence, (2.19) implies

$$\|\rho(2t)\|_{L^\infty(\mathbb{R}^d)} \le Ct^{-\frac{d}{2}} + \int_0^t (t-s)^{-\frac{d(r-1)}{2r}-\frac{1}{2}}(t+s)^{-\frac{d(2r'-1)}{2r'}}ds = Ct^{-\frac{d}{2}} + Ct^{\frac{1}{2}-d}.$$

Since $\|\rho(t)\|_{L^\infty(\mathbb{R}^d)}$ is uniformly bounded and since $d \ge 1$, we obtain the estimate

$$\|\rho(t)\|_{L^\infty(\mathbb{R}^d)} \le C(t+1)^{-\frac{d}{2}}$$

and the L^p decay rates (2.16) can be easily obtained by interpolation. □

In the next proposition we prove that solutions to (2.1) enjoy a time decay rate as in (2.16) no matter how large the mass is, provided $\varepsilon > 1/4$ and $d = 1$. This result constitutes an essential difference of the present model with respect to the classical Keller–Segel type models.

PROPOSITION 2.7. *Let $\varepsilon > 1/4$ and $d = 1$. Let $\rho_0 \in L^1(\mathbb{R})$ satisfying (2.12). Then, the solution $\rho(x,t)$ to (2.1) satisfies the decay estimates*

$$\|\rho(t)\|_{L^p(\mathbb{R})} \le C(t+1)^{-\frac{(p-1)}{2p}}, \quad 2 \le p \le \infty. \qquad (2.21)$$

Proof. We start with the L^2 estimate

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}} \rho^2(x,t)dx &= 2\varepsilon\int_{\mathbb{R}} \rho\rho_{xx}dx + 2\int_{\mathbb{R}} \rho(1-\rho)S_x\rho_x dx \\ &\le -2\varepsilon\int_{\mathbb{R}} \rho_x^2 dx + \frac{1}{2}\int_{\mathbb{R}} |\rho_x S_x|dx.\end{aligned}$$

As a consequence of the Young inequality for convolutions we have (recalling (2.2))

$$\frac{d}{dt}\int_{\mathbb{R}} \rho^2(x,t)dx \le -2\left(\varepsilon - \frac{1}{4}\right)\int_{\mathbb{R}} \rho_x^2 dx.$$

Then, by means of the Gagliardo–Nirenberg inequality, as in the previous proposition, we obtain the decay in L^2

$$\|\rho(t)\|_{L^2(\mathbb{R})} \leq C(t+1)^{-1/4}.$$

By taking the L^4 norm in the representation (2.20), in a similar fashion as in the previous proposition, we obtain

$$\begin{aligned}\|\rho(2t)\|_{L^4(\mathbb{R})} &\leq C(t+1)^{-3/8} + \int_0^t \|\nabla \mathcal{G}(t-s)\|_{L^4} \|\rho \nabla \mathcal{B} * \rho(t+s)\|_{L^1} ds \\ &\leq C(t+1)^{-3/8} + \int_0^t (t-s)^{-7/8}(t+s)^{-1/2} ds \leq C(t+1)^{-3/8} + Ct^{-3/8}.\end{aligned}$$

Finally,

$$\begin{aligned}\|\rho(2t)\|_{L^\infty(\mathbb{R})} &\leq C(t+1)^{-1/2} + \int_0^t \|\nabla \mathcal{G}(t-s)\|_{L^2} \|\rho \nabla \mathcal{B} * \rho(t+s)\|_{L^2} ds \\ &\leq C(t+1)^{-1/2} + \int_0^t (t-s)^{-3/4}(t+s)^{-3/4} ds \leq C(t+1)^{-1/2}.\end{aligned}$$

The remaining L^p estimates are easily obtained by interpolation. ∎

2.2.1. Some remarks and the nonexistence of stationary solutions. Clearly, an open question is whether solutions to (2.1) decay for any ε and for arbitrarily large masses. In bounded domains, for $d = 1$ and Neumann boundary conditions, solutions of the system (2.1) has been shown to decay to the constant solution if $\varepsilon > \frac{1}{4}$, but if ε is small enough, stationary, periodic solutions in $L^1((0, L))$ exist (see [DS05] and [PH05]). However, one can easily prove that there exist no nonzero stationary solutions to (2.1) in $L^1(\mathbb{R}^d)$ in the case of unbounded domains: We define the energy functional $E(\rho, S)$ of system (2.1) by

$$E = \frac{1}{2}\int (|\nabla S|^2 + S^2) dx - \int \rho S dx + \varepsilon \int [\rho \log \rho + (1-\rho)\log(1-\rho)] dx, \tag{2.22}$$

with

$$\frac{\partial E}{\partial \rho} = -S + \varepsilon \log \frac{\rho}{(1-\rho)} \quad \text{and} \quad \frac{\partial E}{\partial S} = -\Delta S + S - \rho. \tag{2.23}$$

Rewriting the first equation of (2.1) as

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho(1-\rho) \nabla \frac{\partial E}{\partial \rho} \right), \tag{2.24}$$

and differentiating the energy with respect to time, we obtain

$$\begin{aligned}\frac{dE}{dt} =& \frac{\partial E}{\partial S}\frac{\partial S}{\partial t} + \frac{\partial E}{\partial \rho}\frac{\partial \rho}{\partial t} = \int \nabla \cdot \left(\rho(1-\rho) \nabla \frac{\partial E}{\partial \rho} \right) \frac{\partial E}{\partial \rho} dx \\ =& -\int \rho(1-\rho) \left| \nabla \frac{\partial E}{\partial \rho} \right|^2 dx \leq 0.\end{aligned}$$

Hence, the energy is decreasing in time, and the stationary state $\frac{dE}{dt} = 0$ is only reached if $\rho = 0$, $\rho = 1$ or $\frac{\partial E}{\partial \rho} =$const., the latter implying that the stationary solution (ρ, S) should satisfy, for some positive constant C,

$$\frac{\rho}{1-\rho} = e^{\frac{S+C}{\varepsilon}},$$

in some open set $\Omega \subset \mathbb{R}^d$, with $\rho = 0$ at some point of $\partial\Omega$. However, this is incompatible with S being bounded, because of the continuity of ρ stated in Proposition 2.3.

2.3. Self-similar long time behavior. The aim of this subsection is to prove that, under suitable assumptions on the initial datum ρ_0, the solution of (2.1) decays towards the fundamental solution of the heat equation with a polynomial rate. To perform this task we employ the entropy dissipation method (see [AMTU01]). The long time decay properties of the solution $\rho(x,t)$ are a crucial ingredient in the arguments below. We shall prove our result by assuming *a priori* that the solution $\rho(x,t)$ satisfies the decay estimate

$$\|\rho(t)\|_{L^\infty(\mathbb{R}^d)} \le C(t+1)^{-\frac{d}{2}} \tag{2.25}$$

which we proved to be fulfilled under the assumptions in Propositions 2.6 and 2.7. Our first step is the following standard time dependent rescaling:

$$\begin{cases} \rho(x,t) = R(t)^{-\frac{d}{2}} v(y,s), \\ y = R(t)^{-\frac{1}{2}} x, \\ s = \frac{1}{2}\log R(t), \\ R(t) = 2t+1. \end{cases} \tag{2.26}$$

Then, it is easy to see that $v(y,s)$ solves the Cauchy problem

$$\begin{cases} \frac{\partial v}{\partial s} = \nabla\cdot(\varepsilon\nabla v + yv) - e^{-ds}\nabla\cdot(v(1-e^{-s}v)\mathcal{B}_s * \nabla v) \\ v(y,0) = \rho_0(y), \end{cases} \tag{2.27}$$

where $\mathcal{B}_s(y) = e^s\mathcal{B}(e^s y)$. For future reference we write (2.27) as follows:

$$\frac{\partial v}{\partial s} = \varepsilon\nabla\cdot\left(v\nabla\left(\log v + \frac{|y|^2}{2\varepsilon}\right)\right) - e^{-ds}\nabla\cdot(v(1-e^{-s}v)\mathcal{B}_s * \nabla v). \tag{2.28}$$

We remark that the fundamental solution $\mathcal{G}(x,t)$ of the heat equation in rescaled variables depends only on y; more precisely, it is given by $\mathcal{U}_m(y) = Ce^{-\frac{|y|^2}{2\varepsilon}}$, where C depends on the mass m of ρ_0. Moreover, we recall that $\mathcal{U}_m$ satisfies the elliptic equation $\nabla\cdot(\varepsilon\nabla\mathcal{U}_m + y\mathcal{U}_m) = 0$. We shall make use of the classical entropy functional

$$\mathrm{E}(v) = \int_{\mathbb{R}^d} v(y)\log v(y)dy + \frac{1}{2\varepsilon}\int_{\mathbb{R}^d}|y|^2 v(y)dy, \tag{2.29}$$

and of the *relative entropy* $\mathrm{RE}(v|\mathcal{U}_m) = \mathrm{E}(v) - \mathrm{E}(\mathcal{U}_m)$. Once the mass m is fixed, the relative entropy functional attains zero as minimum value at the ground state $\mathcal{U}_m$. The following inequality (see [AMTU00] for the proof) establishes a connection between the convergence in relative entropy and the convergence in L^1.

THEOREM 2.8 (Csiszár–Kullback inequality). *Let $v \in L^1(\mathbb{R}^d)$ having mass m. Then, there exists a fixed constant C (depending on m) such that*

$$\|v - \mathcal{U}_m\|^2 \le C\mathrm{RE}(v|\mathcal{U}_m). \tag{2.30}$$

For future reference, we recall the following logarithmic Sobolev inequality (see, e.g., [AMTU01]).

THEOREM 2.9 (logarithmic Sobolev inequality). *Let $v \in L^1(\mathbb{R}^d)$ having mass m and such that $\mathrm{E}(v) < +\infty$. Then, the following inequality is satisfied:*

$$\mathrm{RE}(v|\mathcal{U}_m) \leq \frac{\varepsilon}{2}\mathrm{I}(v|\mathcal{U}_m), \tag{2.31}$$

where

$$\mathrm{I}(v|\mathcal{U}_m) = \int_{\mathbb{R}^d} v \left|\nabla\left(\log\frac{v}{\mathcal{U}_m}\right)\right|^2 dy = \int_{\mathbb{R}^d} v \left|\nabla\left(\log v + \frac{|y|^2}{2\varepsilon}\right)\right|^2 dy$$

is called (relative) Fisher information.

Finally, we remark that assumption (2.25) for ρ in the new variables $v(y,s)$ reads

$$\|v(s)\|_{L^\infty(\mathbb{R}^d)} \leq C \tag{2.32}$$

for some fixed $C > 0$ depending only on the initial datum. We are ready to state the main theorem of this subsection.

THEOREM 2.10. *Let $\rho \in L^1(\mathbb{R}^d)$ satisfy* (2.12) *and $\mathrm{E}(\rho_0) < \infty$, where* E *is defined in* (2.29). *Suppose that the corresponding solution $\rho(x,t)$ to* (2.1) *satisfies the time decay condition* (2.25). *Then,*

$$\|\rho(t) - \mathcal{G}(t)\|_{L^1(\mathbb{R}^d)} = \begin{cases} o(t^{-\frac{1}{2}+\delta}) & \text{for arbitrary } 0 < \delta \ll 1 \quad \text{if } d = 1 \\ O(t^{-\frac{1}{2}}) & \text{if } d > 1, \end{cases} \tag{2.33}$$

where $\mathcal{G}$ is the Gaussian solution (2.4) *of the heat equation with same mass as ρ_0.*

Proof. In what follows, we shall denote a generic positive constant independent on s by C. Let us multiply (2.27) by $\log v(y) + \frac{|y|^2}{2\varepsilon}$ and integrate over $\mathbb{R}^d$. Then, integration by parts and conservation of the total mass yield

$$\begin{aligned}
&\frac{d}{ds}\mathrm{E}(v(s)|\mathcal{U}_m) = -\varepsilon\int_{\mathbb{R}^d} v\left|\nabla\left(\log\frac{v}{\mathcal{U}_m}\right)\right|^2 dy + Ce^{-ds}\int_{\mathbb{R}^d} v\nabla\mathcal{B}_s * v\cdot\nabla\left(\log v + \frac{|y|^2}{2\varepsilon}\right)dy \\
&\leq -\varepsilon\mathrm{I}(v|\mathcal{U}_m) + Ce^{-ds}\left(\int_{\mathbb{R}^d} v\left|\nabla\mathcal{B}_s * v\right|^2 dy\right)^{1/2}\left(\int_{\mathbb{R}^d} v\left|\nabla\left(\log v + \frac{|y|^2}{2\varepsilon}\right)\right|^2 dy\right)^{1/2} \\
&=: -\varepsilon\mathrm{I}(v|\mathcal{U}_m) + J.
\end{aligned} \tag{2.34}$$

We now estimate the term J by using (2.32) and Young's inequality,

$$\begin{aligned}
J &\leq Ce^{-ds}\left(\int_{\mathbb{R}^d}\frac{1}{v}\left|\nabla v\right|^2 dy\right)^{1/2} I(v|\mathcal{U}_m)^{1/2} \\
&= Ce^{-ds}\left(I(v|\mathcal{U}_m) + 2d\int_{\mathbb{R}^d} v - \frac{1}{\varepsilon}\int_{\mathbb{R}^d} v|y|^2 dy\right)^{1/2} I(v|\mathcal{U}_m)^{1/2} \\
&\leq Ce^{-ds}\left(I(v|\mathcal{U}_m) + I(v|\mathcal{U}_m)^{1/2}\right).
\end{aligned}$$

Therefore, inequality (2.34) implies

$$\frac{d}{ds}\mathrm{E}(v(s)|\mathcal{U}_m) \leq -(\varepsilon - Ce^{-ds} - Ce^{-2\delta s})I(v|\mathcal{U}_m) + Ce^{-2ds+2\delta s},$$

for an arbitrarily small $\delta > 0$ and for a fixed $C > 0$. For $s \geq s_0(\varepsilon)$ we have $\varepsilon - Ce^{-ds} - Ce^{-2\delta s} > 0$ and we can use inequality (2.31) to obtain

$$\frac{d}{ds}\mathrm{E}(v(s)|\mathcal{U}_m) \leq -\frac{2(\varepsilon - Ce^{-ds} - Ce^{-2\delta s})}{\varepsilon}\mathrm{E}(v(s)|\mathcal{U}_m) + Ce^{-2ds+2\delta s}.$$

Hence, due to the variation of constants formula we obtain the following decay rates as $s \to +\infty$ (here $\delta > 0$ is arbitrarily small):

$$\mathrm{E}(v(s)|\mathcal{U}_m) = \begin{cases} O(e^{-2(1-\delta)s}) & \text{if } d = 1 \\ O(e^{-2s}) & \text{if } d > 1. \end{cases} \tag{2.35}$$

Finally, by means of the Csiszár–Kullback inequality (2.30) and by using the original variables $\rho(x,t)$ according to (2.26), we obtain the desired estimate (2.33). ∎

3. Nonlinear diffusion. In this section we focus our attention on the nonlinear model

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = \nabla \cdot (\rho(1-\rho)\nabla(\varepsilon\rho - S)) \\ -\Delta S + S = \rho, \end{cases} \tag{3.1}$$

subject to the initial condition

$$\rho(x,0) = \rho_0(x), \quad \rho_0 \in L^1(\mathbb{R}^d), \quad 0 \leq \rho_0(x) \leq 1, \quad x \in \mathbb{R}^d. \tag{3.2}$$

3.1. Existence of weak solutions. We start by providing a suitable definition of weak solutions. In order to simplify the notation, we fix $\varepsilon = 1$. We denote

$$A(\rho) = \int_0^\rho \xi(1-\xi)d\xi = \frac{\rho^2}{2} - \frac{\rho^3}{3}, \qquad \mathcal{A}(\rho) = \int_0^\rho A(\xi)d\xi = \frac{\rho^3}{6} - \frac{\rho^4}{12}.$$

Definition 3.1. *A function $\rho \in L^2([0,+\infty) \times \mathbb{R}^d)$ is called a* weak solution *of the Cauchy problem* (3.1)–(3.2) *on $\mathbb{R}^d \times [0,T]$ (T eventually $+\infty$) if the following conditions are satisfied:*

(i) $\mathcal{A}(\rho) \in L^\infty([0,T]; L^1(\mathbb{R}^d))$
(ii) $\nabla A(\rho) \in L^2([0,T] \times \mathbb{R}^d)$
(iii) $0 \leq \rho(x,t) \leq 1$ *almost everywhere in* $[0,T] \times \mathbb{R}^d$.
(iv) *For all $\varphi \in C_c^\infty([0,T) \times \mathbb{R}^d)$, the following relation holds:*

$$\int_0^T \int_{\mathbb{R}^d} \rho\varphi_t dxdt - \int_0^T \int_{\mathbb{R}^d} \nabla A(\rho) \cdot \nabla\varphi dxdt + \int_0^T \int_{\mathbb{R}^d} \rho(1-\rho)\nabla S \cdot \nabla\varphi dxdt + \int_{\mathbb{R}^d} \rho(x,0)\varphi(x,0)dx = 0,$$

where $S = S[\rho]$ is the unique H^1 solution to $-\Delta S + S = \rho$.

The existence of weak solutions in the sense of the above definition is stated in the following theorem.

Theorem 3.2 (global existence of solutions). *There exists at least a global weak solution $\rho(x,t)$ to the Cauchy problem* (3.1)–(3.2) *in the sense of Definition* 3.1.

The proof of the previous theorem can be performed by a standard approximation procedure. We refer in particular to [Car99, CP03] for existence theory of degenerate convection diffusion equations and to [BCM03, CMV03] for problems including

nonlocal terms. The main technical problem in our case is related to the degeneracy of the nonlinear diffusion coefficient at the threshold value $\rho = 1$. We can overcome this difficulty by including the invariant domain property $0 \leq \rho \leq 1$ in Definition 3.1. This property prevents the mobility coefficient $\rho(1-\rho)$ from turning into the backward diffusion range. For the sake of completeness, let us give an outline of the proof. As a first step we consider a nondegenerate approximation of our problem on a smooth bounded domain with Dirichlet boundary conditions. If S is given, the solution ρ is provided by the classical theory of parabolic equations. Hence, the existence of the solution to the coupled problem can be provided by compactness of the solution operator $S = S(\rho)$ and by Schauder's fixed point theorem. The nondegenerate mobility coefficients can be carefully chosen in order to preserve the conditions in (3.2) in time. This property can be proved in the same way as in Theorem 2.5 due to the nondegeneracy of the parabolic equation. The second step consists in taking the limit in the mobility coefficients in order to prove existence for the degenerate problem in a closed ball. Finally one can send the radius of the ball to infinity and complete the proof. In the last two steps one can use the classical energy estimate technique from [Váz92] in order to achieve sufficient compactness.

Remark 3.3. The conservation of the total mass can be proved in the same way as in the case of linear diffusion.

3.2. entropy solutions. In the following we turn our attention to the problem of uniqueness of suitable solutions. For equations with an interaction of nonlocal fluxes and degenerate diffusions there is no straightforward way to prove the uniqueness of weak solutions (cf. also [BCM03, Car99]) and one might even expect nonuniqueness, as for nonlocal transport equations (cf. [DGT00]). We therefore turn our attention to a rather natural restriction of weak solutions, the so-called *entropy solutions*. Apart from uniqueness, the main motivation for considering entropy solutions is to obtain the correct dissipation of entropy functionals, which will be discussed in the subsections below. Due to the fact that the convolution $\mathcal{B} * \rho$ is smooth anyway, the behavior of dissipation functionals on this part seems unimportant and we shall therefore adapt the definition of entropy solutions for fixed flux $\nabla S(x,t)$ (cf. [BCM03] for a more detailed discussion).

Definition 3.4 (entropy solutions). *We shall say that a nonnegative function* $\rho \in L^1([0,T] \times \mathbb{R}^d) \cap C(0,T;L^1(\mathbb{R}^d))$ *is an* entropy solution *of the Cauchy problem* (3.1)–(3.2) *on* $\mathbb{R}^d \times [0,T]$ *if the following conditions are satisfied:*

(i) *For all* $c \in \mathbb{R}$ *and all nonnegative test functions* $\varphi \in C_c^\infty([0,T) \times \mathbb{R}^d)$*, the following entropy inequality holds:*

$$\int_0^T \int_{\mathbb{R}^d} \Big[|\rho - c| \varphi_t + \text{sign } (\rho - c) \left(\rho(1-\rho) - c(1-c)\right) \nabla S \cdot \nabla \varphi + \varepsilon |A(\rho) - A(c)| \Delta \varphi - \text{sign } (\rho - c) c(1-c) \Delta S \varphi \Big] \, dx \, dt \geq 0, \tag{3.3}$$

where $S = S[\rho]$ *is the unique* H^1 *solution to* $-\Delta S + S = \rho$.

(ii) $A(\rho) \in L^2(0,T;H^1(\mathbb{R}^d))$

(iii) $0 \leq \rho(x,t) \leq 1$ *almost everywhere in* $[0,T] \times \mathbb{R}^d$.

(iv) *Essentially, as* $t \downarrow 0$,

$$\int_{\mathbb{R}^d} |\rho(x,t) - \rho_0(x)| \, dx \to 0.$$

In order to prove the existence of entropy solutions, one can follow the same nondegenerate approximation procedure briefly discussed in the previous subsection and even obtain a unique classical solution. The classical solution clearly satisfies the corresponding entropy condition (with smoothed nonlinearities on the bounded domain). Due to *a priori* bounds on ρ, $A(\rho)$, and $S = \mathcal{B} * \rho$ (see once again the aforementioned reference [Váz92]) we always extract suitably convergent subsequences when passing from the nondegenerate approximation to (3.1), so that the entropy inequality (3.3) carries over to the limit. Hence, we obtain the following result.

THEOREM 3.5. *Let $\rho_0 \in L^1(\mathbb{R}^d)$ satisfy $0 \le \rho_0 \le 1$. Then there exists an entropy solution of* (3.1), (3.2) *according to Definition* (3.4).

In order to prove uniqueness of the entropy solution, we shall use the continuous dependence of entropy solutions on the flux in the L^1-norm.

LEMMA 3.6. *Let $S^1, S^2 \in C(0,T;H^1(\mathbb{R}^d)) \cap L^\infty([0,T]\times\mathbb{R}^d)$, with*

$$\nabla S^j \in C(0,T;W^{1,1}_{loc}(\mathbb{R}^d)) \cap C([0,T]\times\mathbb{R}^d), \qquad \Delta S^j \in L^\infty([0,T]\times\mathbb{R}^d)$$

be given and let $u^j \in L^\infty(0,T;BV(\mathbb{R}^d))$ be entropy solutions of

$$u^j_t = \nabla\cdot\big(u^j(1-u^j)\nabla(\varepsilon u^j - S^j)\big)$$

with initial values $u^j_0 \in BV(\mathbb{R}^d)$ for $j = 1, 2$.

$$\begin{aligned}\|u^1(t)-u^2(t)\|_{L^1(\mathbb{R}^d)} \le \|u^1_0 - u^2_0\|_{L^1(\mathbb{R}^d)} + \frac{t}{4}\|\nabla S^1 - \nabla S^2\|_{L^\infty(0,t;BV(\mathbb{R}^d))}\\ + t\max\{V_1,V_2\}\|\nabla S^1 - \nabla S^2\|_{L^\infty([0,t]\times\mathbb{R}^d)},\end{aligned} \tag{3.4}$$

where $V_j = \|\nabla S^j\|_{L^\infty(0,t;BV(\mathbb{R}^d))}$.

Proof. The proof can be carried out as the proof of Theorem 1.3 in [KR03] by appropriately using the time-dependence of the flux in the estimates and the fact that the function $p \mapsto p(1-p)$ has Lipschitz constant 1 and supremum $\frac{1}{4}$ on the interval $[0,1]$. □

Below we shall prove a uniqueness result in the smaller class of entropy solutions of bounded variation in the case of spatial dimension one. Before proving their uniqueness, we verify the existence of such entropy solutions.

PROPOSITION 3.7 (regularity of entropy solutions). *Let $\rho_0 \in BV(\mathbb{R}^1;[0,1])$, then an entropy solution of* (3.1), (3.2) *satisfies*

$$\rho \in L^\infty(0,T;BV(\mathbb{R}^1)).$$

Proof. We construct the BV-solution by smooth approximation. Let ρ be an L^1 viscosity solution of (3.1), (3.2) and let $\rho^\delta \in C(0,T;C^1(\mathbb{R}^1))$ such that $0 \le \rho^\delta \le 1$ almost everywhere and $\rho^\delta \to \rho$ in $C(0,T;BV(\mathbb{R}^1))$. Then we also have

$$S^\delta = \mathcal{B} * \rho \in C(0,T;BV(\mathbb{R}^1))$$

and consequently

$$S^\delta_{xx} = S^\delta - \rho^\delta \in C(0,T;BV(\mathbb{R}^d)).$$

The results of [KR01] imply the existence of an entropy solution u^δ of

$$u_t - \big(u(1-u)(\delta u - S^\delta)_x\big) = 0$$

with initial value $u(0) = \rho_0$, u^δ belonging in the space $\in L^\infty(0,T;BV(\mathbb{R}^1)$. Moreover, since $u_h^\delta(x,t) = u^\delta(x+h,t)$ is an entropy solution of the same equation with $S^\delta(x.t)$ replaced by $S_h^\delta(x,t) := S^\delta(x+h,t)$ and initial value $\rho_0(.+h)$, we may apply the continuous dependence estimate to deduce

$$\begin{aligned}\|u_h^\delta(t) - u^\delta(t)\|_{L^1(\mathbb{R}^1)} \le \|\rho_0(.+h) - \rho^0\|_{L^1(\mathbb{R}^1)} + \frac{t}{4}\|S_x^\delta - (S_h^\delta)_x\|_{L^\infty(0,t;BV(\mathbb{R}^1))} \\ + tV^\delta \|S_x^\delta - (S_h^\delta)_x\|_{L^\infty([0,t]\times\mathbb{R}^1)},\end{aligned}$$

with $V^\delta = \|S_x^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))}$. After division by h we obtain in the limit $h \to 0$,

$$\|u^\delta(t)\|_{BV(\mathbb{R}^1)} \le \|\rho^0\|_{BV(\mathbb{R}^1)} + \frac{t}{4}\|S_{xx}^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))} + tV^\delta\|S_{xx}^\delta\|_{L^\infty([0,t]\times\mathbb{R}^1)}.$$

From the uniform bounds for ρ^δ in $L^\infty(\mathbb{R}^1) \cap L^1(\mathbb{R}^1)$ one can easily deduce uniform estimates for V^δ and $\|S_{xx}^\delta\|_{L^\infty([0,t]\times\mathbb{R}^1)}$. Moreover, there exists a constant $c > 0$ (independent of δ and t) such that

$$\|S_{xx}^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))} \le c\|\rho^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))}.$$

Hence, we deduce

$$\|u^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))} \le \|\rho^0\|_{BV(\mathbb{R}^1)} + \frac{ct}{4}\|\rho^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))} + Ct.$$

As $\delta \to 0$, one can prove in a standard way that $u^\delta \to \rho$. Now let t be such that $2ct < 1$, then by lower semicontinuity

$$\begin{aligned}\frac{1}{2}\|\rho\|_{L^\infty(0,t;BV(\mathbb{R}^1))} &\le \limsup\left(\|u^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))} - \frac{1}{2}\|\rho^\delta\|_{L^\infty(0,t;BV(\mathbb{R}^1))}\right) \\ &\le \|\rho^0\|_{BV(\mathbb{R}^1)} + Ct,\end{aligned}$$

and hence, $\rho \in L^\infty(0,t;BV(\mathbb{R}^1))$.

By applying the same argument consecutively to time intervals of length smaller than $\frac{1}{2c}$ we finally obtain that $\rho \in L^\infty(0,T;BV(\mathbb{R}^1))$. ∎

Finally, from the continuous dependence it is a small step to prove the main uniqueness result.

THEOREM 3.8 (uniqueness). *The entropy solution ρ of* (3.1), (3.2) *is unique in the space $L^\infty(0,T;BV(\mathbb{R}^1))$.*

Proof. Let ρ^1 and ρ^2 be two different entropy solutions belonging in the space $L^\infty(0,T;BV(\mathbb{R}^1))$. Then we can assume without restriction of generality that $\|\rho^1(t) - \rho^2(t)\| \neq 0$ for $t > 0$ is arbitrarily small (otherwise we can take τ as the maximal time before which the solutions are equal and rescale time to $t - \tau$).

One can verify in a straightforward way that S^1 and S^2 satisfy the assumptions of Lemma 3.6 and that there exists a constant $C > 0$ such that

$$\|S_x^1 - S_x^2\|_{L^\infty(0,t;BV(\mathbb{R}^1))} \le C\|\rho^1 - \rho^2\|_{L^\infty(0,t;L^1(\mathbb{R}^1))}$$

and

$$\|S_x^1 - S_x^2\|_{L^\infty([0,t]\times\mathbb{R}^1)} \le C\|\rho^1 - \rho^2\|_{L^\infty(0,t;L^1(\mathbb{R}^1))}.$$

Hence, by Lemma 3.6 the estimate

$$\|\rho^1(t) - \rho^2(t)\|_{L^1(\mathbb{R}^1)} \le \tilde{C}t\|\rho^1(t) - \rho^2(t)\|_{L^1(\mathbb{R}^1)}$$

holds for some constant $\tilde{C}$. Since we can choose t small enough such that $Ct < 1$, this yields a contradiction. ∎

3.3. Finite speed of propagation in one space dimension. In this subsection we focus our attention on the Cauchy problem in one space dimension

$$\begin{cases} \partial_t \rho = \partial_x \left(\rho(1-\rho) \partial_x \left(\varepsilon \rho - S \right) \right), \\ \rho(x,0) = \rho_0(x), \end{cases} \tag{3.5}$$

where ρ_0 is compactly supported and satisfies the usual condition $0 \leq \rho_0 \leq 1$. Our aim is to prove that the solution $\rho(t)$ at any time $t > 0$ is still compactly supported. This feature is usually referred to as the *the finite rate of propagation property*, and it is typically satisfied by nonlinear diffusion equations of the form

$$\rho_t = A(\rho)_{xx},$$

when $A : \mathbb{R}_+ \to \mathbb{R}_+$ is a smooth nondecreasing function such that $A'(0) = 0$. Our approach in proving such a property is closely related with the estimates of the Wasserstein distances between any two solutions to a nonlinear diffusion equation developed in [CGT04]. For a positive integer n, we define the $2n$th moment of a nonnegative, integrable function ρ as

$$M_{2n}(\rho) := \int_{-\infty}^{+\infty} x^{2n} \rho(x) dx.$$

The distribution function of ρ is given by

$$F(x) = \int_{-\infty}^{x} \rho(y) dy.$$

The *pseudoinverse* function of F, defined on the interval $[0, m]$, $m = \int_{-\infty}^{+\infty} \rho(x) dx$, is given by

$$F^{-1}(\xi) = \inf \left\{ x \in \mathbb{R} \,\middle|\, F(x) > \xi \right\}.$$

If $\rho(x) > 0$ almost everywhere, then F^{-1} is a real inverse, and the change of variables

$$x \in \operatorname{supp}(\rho) \mapsto \xi = F(x) \in [0, m]$$

is a bijection. Therefore, one can change variables in the definition of the moments above and get

$$M_{2n}(\rho) = \int_{-\infty}^{+\infty} x^{2n} \rho(x) dx = \int_0^m F^{-1}(\xi)^{2n} d\xi.$$

Obviously we have

$$\left[\frac{1}{m} \int_0^m F^{-1}(\xi)^{2n} d\xi \right]^{1/2n} \to \| F^{-1} \|_{L^\infty([0,m])},$$

as $n \to +\infty$. Moreover, it is clear that

$$\operatorname{meas}(\operatorname{supp}(\rho)) \leq 2 \| F^{-1} \|_{L^\infty([0,m])}. \tag{3.6}$$

Thus, a uniform estimate with respect to n of $M_{2n}^{1/2n}(\rho(t))$, where $\rho(t)$ is the solution to (3.5), at time t automatically ensures the finiteness of the size of the support of

$\rho(t)$. We shall prove that such an estimate is true. For future reference we briefly recall the following result in [Kne77] for a nonlinear friction equation.

THEOREM 3.9. *Let* $A(\rho) = \frac{\rho^2}{2} - \frac{\rho^3}{3}$. *Let* $\overline{\rho}(t)$ *be a nonnegative solution to the equation*

$$\overline{\rho}_t = A(\overline{\rho})_{xx} \tag{3.7}$$

with initial datum having compact support. Let $\overline{F}^{-1}(t)$ *be the pseudoinverse of the distribution function of* $\overline{\rho}(t)$. *Then, there exists a continuous (increasing) function of time* $t \mapsto C(t)$ *such that*

$$\|\overline{F}^{-1}(t)\|_{L^{2n}} \leq C(t), \tag{3.8}$$

where $C(t)$ *does not depend on* n.

We now state our result for (3.5).

THEOREM 3.10 (finite speed of propagation). *Let* $\rho(x,t)$ *be the unique entropy solution to* (3.5) *with compactly supported initial datum* ρ_0. *Then, the profile* $\rho(t)$ *has compact support at any positive time* t.

Proof. As pointed out before, to pursue our aim we only need to control $M_{2n}^{1/2n}$ uniformly in n in any finite time interval $[0,T]$. We shall perform this task by a formal computation of the L^{2n} norm of the pseudoinverse of the distribution function of ρ. This formal computation can be easily made rigorous by an approximation procedure that allows us both to cancel out the boundary terms due to integration by parts and to treat the pseudoinverse of the distribution function of ρ as a real inverse. This procedure is well explained in [CGT04] and we shall omit the details about it. We only observe that classical energy estimates on (3.5) in the spirit of [Váz92] provide the necessary compactness needed to extend such approximation argument to our case.

We denote by ρ the solution of (3.5) with an initial datum ρ_0 having mass $m = \int \rho_0$. Let $u(t) : [0,m] \to \mathbb{R}$ be the pseudoinverse of the distribution function of ρ. Let us also consider the solution $\overline{\rho}(t)$ of the nonlinear diffusion equation (3.7) having ρ_0 as initial datum. Let $\overline{u}(t)$ be the pseudoinverse of the distribution function of $\overline{\rho}(t)$. A standard computation with pseudoinverses (see, e.g., [CGT04]) shows that u and $\overline{u}$ satisfy the following equations:

$$\begin{aligned}
\partial_t u &= -\partial_\xi A\left((\partial_\xi u)^{-1}\right) + \left(1 - (\partial_\xi u)^{-1}\right) \int_0^m B'(u(\xi) - u(\eta))d\eta, \\
\partial_t \overline{u} &= -\partial_\xi A\left((\partial_\xi \overline{u})^{-1}\right).
\end{aligned}$$

For fixed n we have, after integration by parts,

$$\begin{aligned}
&\frac{d}{dt}\int_0^m [u(t) - \overline{u}(t)]^{2n} d\xi \\
&= +2n(2n-1)\int_0^m (u-\overline{u})^{2n-2}\left[\partial_\xi u - \partial_\xi \overline{u}\right]\left[A((\partial_\xi u)^{-1}) - A((\partial_\xi \overline{u})^{-1})\right] d\xi \\
&\quad + 2n\int_0^m (u-\overline{u})^{2n-1}\left(1 - (\partial_\xi u)^{-1}\right)\int_0^m B'(u(\xi) - u(\eta))d\eta d\xi.
\end{aligned}$$

Hence, since the function $t \to A(t^{-1})$ is nonincreasing we can get rid of the first addend on the right-hand side above. Therefore, due to the uniform bound of ρ in

L^∞ and to the definition of $\mathcal{B}$, there exists a fixed constant $C > 0$ such that

$$\frac{d}{dt}\int_0^m [u(t) - \overline{u}(t)]^{2n} d\xi \le Cn\left[\int_0^m [u(t) - \overline{u}(t)]^{2n} d\xi + 1\right],$$

and, in view of (3.8) and of the Gronwall inequality,

$$\|u(t)\|_{L^{2n}[0,m]} \le e^{Ct}$$

for some positive constant C independent of n. We can now send $n \to \infty$ to get a uniform estimate for $\|u(t)\|_{L^\infty[0,m]}$ and the proof is complete in view of (3.6). ∎

3.4. Asymptotic behavior. In the following we investigate the asymptotic behavior of weak solutions to (3.1), (3.2) for large time. The main idea in this case is the analysis of the behavior of the associated energy functional

$$\tilde{E}(\rho) := \int_{\mathbb{R}^d} \rho(\varepsilon\rho - S(\rho))\ dx, \tag{3.9}$$

where $S(\rho)$ is the unique solution to $-\Delta S + S = \rho$ decaying at infinity. This energy functional is to be considered on the admissible set

$$\mathcal{K} := \{\rho \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d) \mid 0 \le \rho \le 1 \text{ a.e. }\}. \tag{3.10}$$

For the nonlinear diffusion case, the change from $\varepsilon > 1$ to $\varepsilon < 1$ is of particular interest, since $\tilde{E}(\rho)$ changes from a strictly convex (for $\varepsilon > 1$) to a nonconvex functional, which we verify in the following lemma.

LEMMA 3.11. *The functional $\tilde{E} : \mathcal{K} \to \mathbb{R}$ is bounded below by $-\int_{\mathbb{R}^d} \rho\ dx$ for $\varepsilon > 0$. Moreover, $\tilde{E}$ is positive and strictly convex for $\varepsilon > 1$.*

Proof. First of all, since $\rho \le 1$ and $S(\rho) \ge 0$, we have

$$\tilde{E}(\rho) := \int_{\mathbb{R}^d} \rho(\varepsilon\rho - S(\rho))\ dx \ge -\int_{\mathbb{R}^d} S(\rho)\ dx.$$

The property $\int_{\mathbb{R}^d} S(\rho)\ dx = \int_{\mathbb{R}^d} \rho\ dx$ can be deduced immediately from the elliptic equation satisfied by $S(\rho)$ and hence,

$$\tilde{E}(\rho) \ge -\int_{\mathbb{R}^d} \rho\ dx.$$

Since $\tilde{E}$ is a quadratic functional, convexity is equivalent to strict positivity. From the Cauchy–Schwarz inequality we have

$$\tilde{E}(\rho) = \int_{\mathbb{R}^d} \rho(\varepsilon\rho - S(\rho))\ dx \ge (\varepsilon - 1)\int_{\mathbb{R}^d} \rho^2\ dx + \frac{1}{2}\int_{\mathbb{R}^d} \left(\rho^2 - S(\rho)^2\right)\ dx.$$

Finally, a standard energy estimate for the elliptic equation satisfied by $S(\rho)$ shows that the second term is nonnegative, and hence,

$$\tilde{E}(\rho) \ge (\varepsilon - 1)\int_{\mathbb{R}^d} \rho^2\ dx > 0$$

for $\rho \ne 0$ and $\varepsilon > 1$. ∎

A fundamental property of the model is the dissipation of the energy $\tilde{E}$. Formally, if we compute the time derivative of $\tilde{E}(\rho(t))$, insert the equations and apply Gauss' theorem, then we obtain

$$\begin{aligned}\frac{d}{dt}\tilde{E}(\rho(t)) &= \int_{\mathbb{R}^d} \rho_t(2\varepsilon\rho - \mathcal{B}*\rho)\ dx - \int_{\mathbb{R}^d} \rho\mathcal{B}*\rho_t\ dx \\ &= 2\int_{\mathbb{R}^d} \rho_t(\varepsilon\rho - \mathcal{B}*\rho)\ dx \\ &= 2\int_{\mathbb{R}^d} \nabla\cdot[\rho(1-\rho)\nabla(\varepsilon\rho - \mathcal{B}*\rho)]\,(\varepsilon\rho - \mathcal{B}*\rho)\ dx \\ &= -2\int_{\mathbb{R}^d} \rho(1-\rho)\left|\nabla(\varepsilon\rho - \mathcal{B}*\rho)\right|^2\ dx := -2I(\rho,S) < 0. \end{aligned} \tag{3.11}$$

This estimate can be made rigorous by standard smooth approximation techniques, and thus, we have derived the following result.

PROPOSITION 3.12 (energy dissipation). *Let ρ be a weak solution of* (3.1), (3.2). *Then the functional*

$$e : \mathbb{R}^+ \to \mathbb{R}, t \mapsto e(t) := \tilde{E}(\rho(t))$$

is nonincreasing. Moreover, $e(s) = e(t)$ for $s > t$ if and only if ρ is stationary in the interval $[s,t]$ and

$$\rho(1-\rho)\nabla(\varepsilon\rho - S(\rho)) = 0 \qquad \text{a.e. in } \mathbb{R}^d \times [s,t].$$

3.4.1. Nondecaying solutions in 1-*d* for moderate diffusivity. It is an important consequence of the previous proposition that for an initial value ρ_0 for (3.1) with $\tilde{E}(\rho_0) < 0$, the existence of nondecaying solutions follows: in this case, we have $\tilde{E}(\rho(t)) \le \tilde{E}(\rho_0) < 0$ due to the energy dissipation. Hence, such solutions cannot decay for $t \to \infty$, since otherwise one would have $\liminf \tilde{E}(\rho(t)) \ge \tilde{E}(0) = 0$.

In one space dimension and for every $\varepsilon < 1$, we can verify that the functional $\tilde{E}$ is not positive—and hence nondecaying solutions do exist—by an explicit construction of appropriate densities.

PROPOSITION 3.13. *Let $d = 1$ and $\varepsilon < 1$, then for each $m > 0$ there exists $\rho \in \mathcal{K}$ satisfying*

$$\tilde{E}(\rho) < 0, \qquad \text{and} \qquad \int_{\mathbb{R}^d} \rho\ dx = m. \tag{3.12}$$

Proof. Let $0 < \alpha^2 < \frac{1}{\varepsilon} - 1$ and let ψ^α be a continuous function satisfying

$$\begin{aligned} -\tfrac{d^2\psi^\alpha}{dx^2} + \psi^\alpha &= \tfrac{1}{c}, && 0 \le x < a, \\ -\tfrac{d^2\psi^\alpha}{dx^2} - \alpha^2\psi^\alpha &= 0, && a \le x \le b, \\ -\tfrac{d^2\psi^\alpha}{dx^2} + \psi^\alpha &= 0, && x > b, \end{aligned}$$

with boundary conditions $\psi^\alpha(0) = 1$, $\frac{d\psi^\alpha}{dx}(0) = 0$ and $\psi^\alpha(x) \to 0$ as $x \to \infty$, and some constant c satisfying

$$0 < \varepsilon < c < \frac{1}{\alpha^2 + 1} < 1. \tag{3.13}$$

We find that if $a < \ln \frac{c}{1-c}$, a continuously differentiable, nonnegative solution exists and is given by

$$\psi^\alpha(x) = \begin{cases} \frac{1}{c}(1-(c-1)\cosh x), & 0 \le x < a, \\ c_1 \sin(\alpha(x-a)) + c_2 \cos(\alpha(x-a)), & a \le x \le b, \\ c_3 e^{b-x}, & b < x, \end{cases}$$

where the constants satisfy

$$\begin{aligned} c_1 &= \tfrac{c-1}{\alpha c} \sinh a < 0, \\ c_2 &= (1+(c-1)\cosh a)/c, \\ c_3 &= c_1 \sin(\alpha(b-a)) + c_2 \cos(\alpha(b-a)), \end{aligned}$$

and the length of the second interval is fixed by the relation

$$b = a + \frac{1}{\alpha} \arctan\Big(-\frac{c_2 + \alpha c_1}{c_1 - \alpha c_2}\Big).$$

If we choose S as the symmetric extension of $c\psi^\alpha$ to $\mathbb{R}$ and

$$\rho(x) = \begin{cases} 1, & -a < x < a, \\ c(\alpha^2+1)\psi^\alpha(x), & a \le x \le b, \\ c(\alpha^2+1)\psi^\alpha(-x), & -b \ge x \ge -a, \\ 0, & |x| > b, \end{cases}$$

then S satisfies $-\frac{d^2S}{dx^2} + S = \rho$ and $\rho \in \mathcal{K}$. Moreover, we have

$$\tilde{E}(\rho) = \int_{\mathbb{R}^d} \rho(\varepsilon\rho - S)\, dx = 2\Big[\int_0^a (\varepsilon - S)\, dx + (\varepsilon(\alpha^2+1)-1)(\alpha^2+1)\int_a^b S^2\, dx\Big].$$

This expression is negative if $a < \operatorname{arccosh}\Big(\frac{1-\varepsilon}{1-c}\Big)$. The mass m is given by

$$\frac{1}{2}m = a + (\alpha^2+1)\int_a^b S\, dx = a + c(\alpha^2+1)\int_a^b c_1 \sin(\alpha(x-a)) + c_2\cos(\alpha(x-a))\, dx.$$

Because of (3.13), we have that $\frac{1}{2}m < a + \int_a^b c_1 \sin(\alpha(x-a)) + c_2\cos(\alpha(x-a))\, dx$. If α is chosen small enough, we can choose the constant c to be so close to 1 that the constraints on a, and hence on m, become arbitrarily large. Thus, for any given mass m, we can always construct a solution $\rho \in \mathcal{K}$ such that (3.12) holds. ∎

3.4.2. Attractors of the semigroup in 1-*d*. As another interesting consequence of the energy estimate (3.11), for any positive ε we can characterize the set of attractors of the semigroup (3.1), (3.2) as the set of its stationary entropy solutions, i.e., the set of all $\rho \in L^1$, $0 \le \rho \le 1$ such that

$$\rho(1-\rho)\nabla(\varepsilon\rho - S(\rho)) = 0 \qquad \text{a.e. in } \mathbb{R}\times[0,+\infty), \tag{3.14}$$

where $S(\rho)$ is the unique solution to $-\Delta S + S = \rho$ decaying at infinity.

THEOREM 3.14 (attractors of the semigroup). *Let $\rho(t)$ be the solution to* (3.1), (3.2) *in one space dimension. Then, any sequence of times admits a subsequence t_k such that $\rho(t_k) \to \rho^\infty$ almost everywhere. Moreover, ρ^∞ is a solution to* (3.14).

Proof. We set

$$B(\rho) = \int_0^\rho \sqrt{r(1-r)}dr$$

and we observe that (3.11) implies that the quantity

$$\int_0^{+\infty} I(\rho(t), S(t))dt = \int_0^{+\infty} \int_{-\infty}^{+\infty} \left[\varepsilon^2 B(\rho)_x^2 - 2\varepsilon A(\rho)_x S_x + \rho(1-\rho)S_x^2\right] dxdt$$

is uniformly bounded. Therefore, any sequence of times tending to infinity has a subsequence $t_k \to +\infty$ such that $I(\rho(t_k), S(t_k)) \to 0$ as $k \to \infty$. Since

$$\int_{-\infty}^{+\infty} \rho(1-\rho)S_x^2 dx \le \frac{1}{4}\int_{\mathbb{R}^d} \rho^2 dx,$$

and $\rho(t)$ is uniformly bounded in $L^1 \cap L^\infty$, and because of

$$-2\varepsilon \int_{\mathbb{R}^d} A(\rho)_x S_x dx = 2\varepsilon \int_{\mathbb{R}^d} A(\rho)(\rho - S)dx,$$

we easily get a uniform (with respect to time) bound for

$$\int_{-\infty}^{+\infty} B(\rho(t_k))_x^2 dx.$$

Thanks to the uniform bound in $L^1 \cap L^\infty$ for the solution $\rho(t)$, by Sobolev embedding we can extract a new subsequence of times (still denoted by t_k for simplicity) such that

$$\begin{cases} \rho(t_k) \to \rho^\infty \ \text{ a.e. in } \mathbb{R} \times [0+\infty) \\ \rho(t_k) \to \rho^\infty \ \text{ in } \ L^2_{loc}(\mathbb{R}) \\ B(\rho(t_k))_x \to v^\infty \ \text{ weakly in } L^2(\mathbb{R}) \\ I(\rho(t_k), S(t_k)) \to 0, \end{cases}$$

as $k \to \infty$. Now, for any test function $\phi \in C_0^\infty(\mathbb{R})$ we have

$$\begin{aligned} \int_{-\infty}^{+\infty} v^\infty \phi \ dx &= \lim_{k\to\infty} \int_{-\infty}^{+\infty} B(\rho(t_k))_x \phi \ dx \\ &= -\lim_{k\to\infty} \int_{-\infty}^{+\infty} B(\rho(t_k))\phi_x dx = -\int_{-\infty}^{+\infty} B(\rho^\infty)\phi_x dx \end{aligned}$$

and therefore $v^\infty = B(\rho^\infty)_x$ almost everywhere. We denote by S^∞ the solution to $-S_{xx} + S = \rho^\infty$. By weak lower semicontinuity of the L^2 norm and by the strong compactness of $S_x = \mathcal{B}' * \rho$ in L^2 we can extract another subsequence such that

$$\begin{aligned} I(\rho^\infty, S^\infty) &= \varepsilon^2 \int_{-\infty}^{+\infty} \varepsilon B(\rho^\infty)_x^2 dx - 2\varepsilon \int_{-\infty}^{+\infty} A(\rho^\infty)_x S_x^\infty dx \\ &+ \int_{-\infty}^{+\infty} \rho^\infty(1-\rho^\infty)(S_x^\infty)^2 dx \le \liminf_{k\to\infty} \left\{ \int_{-\infty}^{+\infty} B(\rho(t_k))_x dx \right. \\ &\left. -2\varepsilon \int_{-\infty}^{+\infty} A(\rho(t_k))_x S(t_k)_x dx + \int_{-\infty}^{+\infty} \rho(t_k)(1-\rho(t_k))S(t_k)_x^2 dx \right\} \\ &= \lim_{k\to\infty} I(\rho(t_k), S(t_k)) = 0. \end{aligned}$$

Recalling that $I(\rho,S)=\int_{-\infty}^{+\infty}\rho(1-\rho)(\varepsilon\rho-S)_x^2dx$, the assertion follows. ∎

Remark 3.15. In the arguments above we have implicitly supposed existence of stationary solutions. The convergence of subsequences in the above theorem is strong enough to prove existence of stationary, weak, entropy solutions of (3.1), (3.2) in one space dimension. Due to the nondecay result in subsection 3.4.1, we deduce existence of stationary solutions which are not identically zero in case of $\varepsilon<1$.

3.4.3. Characterization of the attractors for large diffusivity in 1-*d*. In this subsection we prove that the only attractor of the semigroup, and hence the only solution to the stationary problem (3.14) is the constant solution $\rho^\infty\equiv 0$ in the case of large diffusivity $\varepsilon>1$. To perform this task, we need an additional energy estimate, i.e., we compute the evolution of the logarithmic functional

$$L(\rho)=\int_{-\infty}^{+\infty}[\rho\log\rho+(1-\rho)\log(1-\rho)]dx.$$

As for the energy estimate in the previous subsections, we compute the evolution of $L(\rho)$ by means of a formal computation which can be made rigorous by approximation. Integration by parts and conservation of the total mass yield

$$\begin{aligned}L(\rho(t))-L(\rho(0))&=\int_0^t\int(\log\rho-\log(1-\rho))\rho_t dxd\tau\\&=-\int_0^t\int\left(\frac{\rho_x}{\rho}+\frac{\rho_x}{1-\rho}\right)(\varepsilon\rho(1-\rho)\rho_x-\rho(1-\rho)S_x)\,dxd\tau\\&=-\varepsilon\int_0^t\int\rho_x^2dxd\tau+\int_0^t\int\rho_xS_xdxd\tau\le-(\varepsilon-1)\int_0^t\int\rho_x^2dxd\tau.\end{aligned}\tag{3.15}$$

The logarithmic functional L cannot be used directly to discern the asymptotic behavior of the solution ρ. However, the estimate performed above can be used to characterize the stationary solutions for $\varepsilon>1$. We have the following theorem.

THEOREM 3.16 (attractors for large diffusivity in 1–*d*). *Let ρ,S be a solution to* (3.1) *with $\varepsilon>1$ such that ρ has finite support at any time. Then, the support of ρ is not uniformly bounded with respect to t. Consequently, there exist no compactly supported stationary solutions ρ,S to* (3.1) *if $\varepsilon>1$ is different from zero.*

Proof. Suppose that $\rho(t)$ is a solution with uniformly bounded support. Since the function

$$[0,1]\ni\rho\mapsto\rho\log\rho+(1-\rho)\log(1-\rho)$$

is bounded, $L(\rho(t))$ is uniformly bounded in time. Therefore, as in the proof of Theorem 3.14, we can handle the right-hand side of estimate (3.15) in a clever way in order to derive strong compactness. More precisely, there exists a divergent sequence of times t_k such that $\rho(t_k)$ converges to some ρ^∞ almost everywhere and strongly in L^2_{loc}, and such that $\rho(t_k)_x$ converges to zero strongly in L^2. As in Theorem 3.14, we can easily prove that $\rho_x^\infty=0$ and, by Fatou's lemma, we conclude that $\rho^\infty=0$. By the Sobolev interpolation lemma we get $\rho(t_k)\to 0$ uniformly, and this contradicts $\rho(t)$ having uniformly bounded support because of the conservation of the mass. This proves the first assertion of the theorem. In particular, we have also proven that any compactly supported stationary solution must equal zero. ∎

As a consequence of the previous theorem, we have the following asymptotic decay result in the case of large diffusivity.

COROLLARY 3.17 (decay of solutions in 1-d for large diffusivity). *Let ρ be the solution to* (3.1) *with compactly supported initial datum ρ_0 satisfying* (3.2). *Then,*

$$\lim_{t\to\infty} \|\rho(t)\|_{L^\infty(\mathbb{R})} = 0.$$

Proof. From Theorem 3.14, any divergent sequence of times admits a subsequence t_k such that $\rho(t_k)$ converges almost everywhere to a stationary solutions satisfying (3.14). Thanks to the results in Theorems 3.10 and 3.16, such a solution must be the constant solution $\rho \equiv 0$. Moreover, the convergence to zero holds in L^∞ in view of $B(\rho(t_k)) \to 0$ in L^2 and by the Sobolev interpolation lemma. ∎

3.5. Stationary solutions. As stated in Proposition 3.12, stationary solutions of (3.1), (3.2) have to satisfy $\rho = 0$, $\rho = 1$ or $\varepsilon\nabla\rho = \nabla S$. In one space dimension, this means that we can construct nontrivial stationary solutions by arranging subintervals on $\mathbb{R}$ such that ρ is in the admissible set $\mathcal{K}$ and satisfies one of these conditions in every interval.

PROPOSITION 3.18. *Let $d = 1$ and $\varepsilon < 1$, then for each $m > 0$ small enough there exists a stationary solution of* (3.1) *satisfying $\rho \in \mathcal{K}$.*

Proof. Let ρ and S be the symmetric extension to $\mathbb{R}$ of

$$\bar{S}(x) = \begin{cases} \frac{\varepsilon}{\varepsilon-1}c_1 + c_2\cos\left(\sqrt{\frac{1-\varepsilon}{e}}x\right), & 0 \le x \le a, \\ c_3 e^{a-x}, & a \le x, \end{cases} \qquad \bar{\rho}(x) = \begin{cases} \frac{1}{\varepsilon}S + c_1, & 0 \le x \le a, \\ 0, & a \le x, \end{cases}$$

and let the constants satisfy

$$c_1 = c\frac{\sqrt{\varepsilon}-1}{\sqrt{\varepsilon}} < 0, \quad c_2 = \frac{c(\varepsilon\sqrt{\varepsilon}-\varepsilon)}{\varepsilon-1} > 0, \quad c_3 = -\varepsilon c_1,$$

where c is the maximal value of ρ and a is given by

$$a = \sqrt{\frac{\varepsilon}{1-\varepsilon}}\arccos(-\sqrt{\varepsilon}).$$

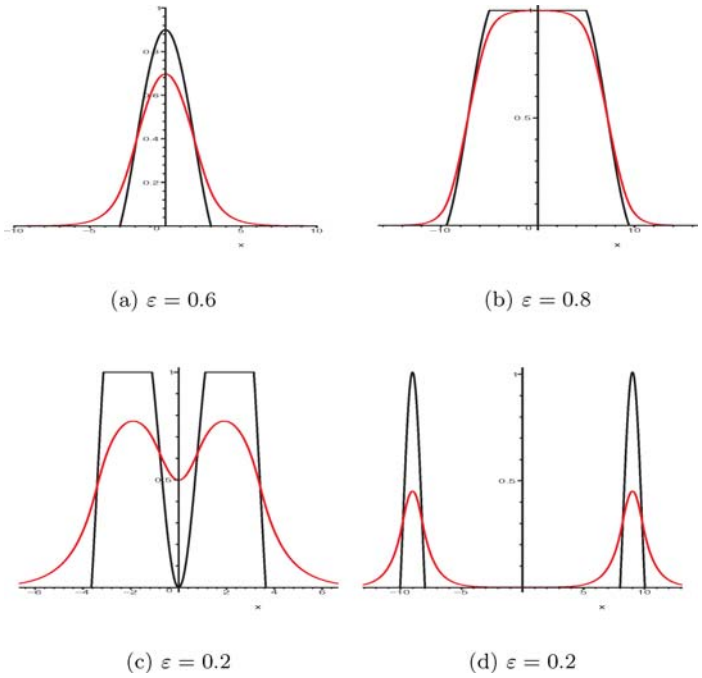
Then, a simple calculation shows that for any given values of $\varepsilon < 1$ and $0 \le c \le 1$, a nonnegative solution $\rho \in C(\mathbb{R})$ and $S \in C^1(\mathbb{R})$ exists. Moreover, S and ρ are decreasing functions on $[0, a]$, implying the assertion. ∎

An example of this type of solution is shown in Figure 1(a), where we set $c = 0.9$ and $\varepsilon = 0.6$.

In general, more complicated stationary solutions can also be constructed, for instance, solutions with

$$\bar{\rho}(x) = \begin{cases} 1, & 0 \le x \le a, \\ \frac{1}{\varepsilon}S + c_1, & a \le x \le b, \\ 0, & b \le x, \end{cases} \tag{3.16}$$

or solutions with several peaks (see Figure 1(b)–1(d)). It is no longer straightforward to show that these solutions exist for any choice of $\varepsilon < 1$, but there is strong numerical evidence. As an example, we chose a stationary solution of type (3.16): it seems that for any mass m large enough and $\varepsilon < 1$, a solution can be uniquely determined. Figure 2 shows the interval a as a function of the mass for different values of ε. Depending on ε, there exists a minimal value for m, which is due to the fact that the slope of ρ, and hence also the minimal distance between intervals where $\rho = 1$ and $\rho = 0$, is proportional to ε.



(a) $\varepsilon = 0.6$ (b) $\varepsilon = 0.8$

(c) $\varepsilon = 0.2$ (d) $\varepsilon = 0.2$

FIG. 1. *Stationary solutions ρ (dark lines) and S (light lines) of* (3.1), (3.2).

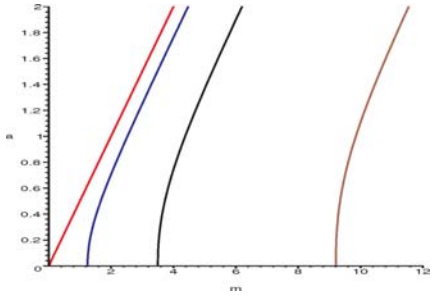
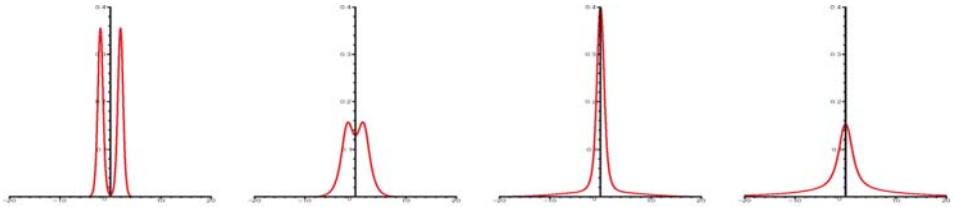


FIG. 2. *Dependence on the parameter a on m for different values of ε (left to right: $\varepsilon = 10^{-3}$, $\varepsilon = 0.2$, $\varepsilon = 0.6$, and $\varepsilon = 0.9$).*

4. Numerical simulation. In the following we discuss the numerical simulation of the linear and nonlinear model in one and two space dimensions. We discretize (2.1) and (3.1) numerically by applying a straightforward finite difference method: for each time step, S is computed using an implicit discretization of the elliptic equation, then the cell density ρ is calculated using the updated value for S. This is done by standard operator splitting of the equation for ρ. First, the advection term is calculated with an upwind scheme (cf. [LeV90]). The solution obtained from this step is subsequently used as an initial value for a time step in the diffusion problem $\rho_t = \varepsilon\Delta\rho$ and $\rho_t = \varepsilon\Delta A(\rho)$ for the linear and the nonlinear case, respectively. If ε is relatively large and the parabolic CFL condition $\Delta t \leq C\frac{\Delta x^2}{\varepsilon}$ becomes too expensive compared to the hyperbolic CFL condition for the advection term, we use a Crank–Nicholson scheme (an implicit, weighted average method, cf. [Qua03]) for the time



(a) $t = 4$ (b) $t = 40$ (c) $t = 400$ (d) $t = 2000$

FIG. 3. *Linear problem,* $\varepsilon = 0.05$.

integration of the linear diffusion term.

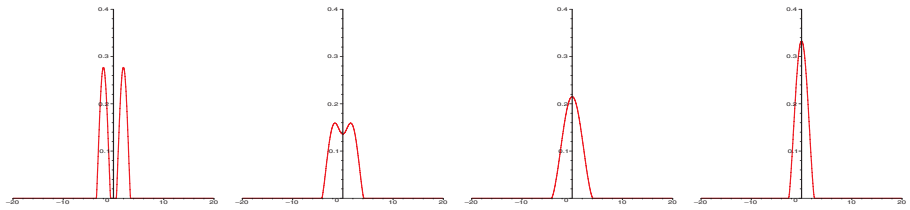
This numerical approach requires the approximation of the solution on an unbounded domain by (artificial) boundary conditions on a bounded domain. For the numerical simulations presented below, we take Dirichlet boundary conditions for S and ρ and perform the computations on large domains.

As a test whether this method is reliable, we compared our results for the nonlinear model to the numerical solution obtained by a so-called Lagrangian approach, where the equation for the pseudoinverse of the distribution function is discretized (see, for instance, [GT05]). Since this method is posed on the interval $(0, 1)$, it does not require any artificial cut of the computational domain. Our comparisons showed that the maximal value of the error is approximately Δx (data not shown) and thus of the same order as the as the error caused by the discretization itself. Hence, the cutting of the domain in our numerical method seems to introduce only a negligible error.

Since the distribution function and its pseudoinverse do not exist in multiple dimensions, we could not compare our numerical solutions to those obtained by the Lagrangian approach for the two-dimensional case. However, we checked our results by successively doubling the size of the domain and comparing the corresponding numerical solutions, which also produced negligible variations for reasonably large domain size.

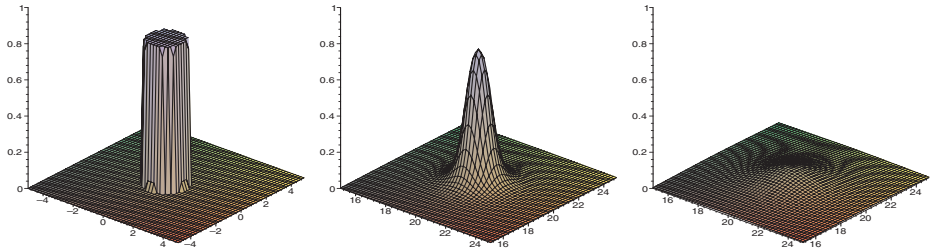
4.1. One-dimensional simulations. The evolution of the cell density in (1.1) is illustrated in Figures 3 and 4 for the linear and nonlinear case, respectively: starting with a symmetrical initial condition for ρ consisting of two peaks both with mass $\frac{1}{2}$, we compute the solutions with the numerical scheme described above, using $\Delta x = 10^{-2}$, $\Delta t = 5 \times 10^{-3}$ in the linear case and $\Delta x = 10^{-2}$, $\Delta t = 5 \times 10^{-4}$ in the nonlinear case. In order to reduce the computation time, we consider symmetric initial data with compact support on $0 \le x \le a$ and prescribe Neumann boundary conditions at $x = 0$, so that we only have to compute on half of the domain.

In Figure 3 one observes that in the linear case the two initial peaks merge into a single peak that eventually decays with time. In Figure 4 the time evolution of the nonlinear diffusion problem (3.1) is illustrated. In order to have a more fair comparison with the linear case we set $\varepsilon = 0.5$, i.e., ten times the value we took for the linear case (but simulations with smaller ε showed a very similar behavior). Starting with the same initial conditions, the solution first behaves as in the linear case: there is attraction between the two peaks and they merge to a single one. In contrast to the linear problem, however, this peak does not decay, but approaches a nontrivial stationary solution as characterized in Proposition 3.18.



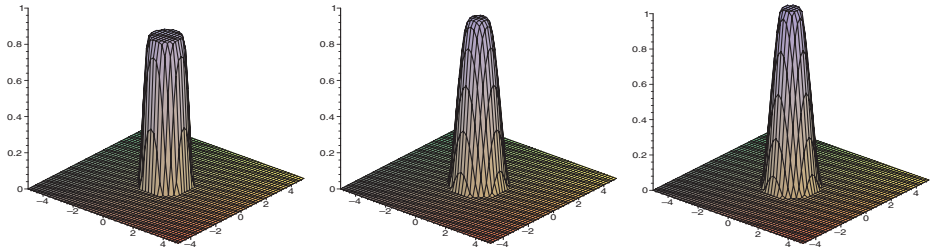
(a) $t = 4$ (b) $t = 40$ (c) $t = 80$ (d) $t = 400$

FIG. 4. *Nonlinear problem,* $\varepsilon = 0.5$.



(a) $t = 0$ (b) $t = 1200$ (c) $t = 1500$

FIG. 5. *Linear problem,* $\varepsilon = 0.04$.



(a) $t = 0.25$ (b) $t = 2$ (c) $t = 35$

FIG. 6. *Nonlinear problem,* $\varepsilon = 0.2$.

4.2. Two-dimensional simulations. In two space dimensions, we perform numerical experiments on a rectangular grid with $\Delta x = \Delta t = 5 \times 10^{-2}$ for the linear model and $\Delta x = 5 \times 10^{-2}$, $\Delta t = 4 \times 10^{-4}$ for the nonlinear model. The diffusivities are $\varepsilon = 0.04$ for the linear and $\varepsilon = 0.2$ for the nonlinear model. Figure 5 illustrates the temporal evolution of the solution ρ of the linear problem, starting from the initial condition shown in the first picture. We see that the maximal value of the density remains close to 1 for a long time interval, but for even larger time diffusion dominates aggregation and the density starts to decay. In Figure 6 the behavior of the nonlinear model with the same initial conditions is shown. One observes that the solution is not decaying, and eventually a stationary state with finite support is approached.

Acknowledgments. The authors would like to thank Kenneth Karlsen (University of Oslo), Josè A. Carrillo (Universitat Autónoma de Barcelona), Chiara Simeoni (Laboratoire Dieudonne, CNRS Nice), Christian Schmeiser (University of Vienna), and Benoit Perthame (ENS, Paris) for useful discussions and suggestions.

REFERENCES

[AMTU00] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter, *On generalized Csiszár-Kullback inequalities*, Monatsh. Math., 131 (2000), pp. 235–253.

[AMTU01] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter, *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations*, Comm. Partial Differential Equations, 26 (2001), pp. 43–100.

[BCM03] M. Burger, V. Capasso, and D. Morale, *On an Aggregation Model with Long and Short Range Interactions*, CAM-Report 03-30, UCLA, Los Angeles, CA, 2003.

[BO04] H. M. Byrne and M. R. Owen, *A new interpretation of the Keller-Segel model based on multiphase modelling*, J. Math. Biol., 49 (2004), pp. 604–626.

[Car99] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[CC05] V. Calvez and J. A. Carrillo, *Volume effects in the Keller–Segel model, energy preventing blow up*, J. Math. Pures Appl. (9), 86 (2006), pp. 155–175.

[CGT04] J. A. Carrillo, M. P. Gualdani, and G. Toscani, *Finite speed of propagation in porous media by mass transportation methods*, C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 815–818.

[CMV03] J. A. Carrillo, R. J. McCann, and C. Villani, *Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates*, Rev. Mat. Iberoamericana, 19 (2003), pp. 971–1018.

[CP03] G.-Q. Chen and B. Perthame, *Well-posedness for non-isotropic degenerate parabolic-hyperbolic equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 645–668.

[CPZ04] L. Corrias, B. Perthame, and H. Zaag, *Global solutions of some chemotaxis and angiogenesis systems in high space dimensions*, Milan J. Math., 72 (2004), pp. 1–28.

[DGT00] O. Diekmann, M. Gyllenberg, and H. R. Thieme, *Lack of uniqueness in transport equations with a nonlocal nonlinearity*, Math. Models Methods Appl. Sci., 10 (2000), pp. 581–591.

[DP04] J. Dolbeault and B. Perthame, *Optimal critical mass in the two-dimensional Keller-Segel model in* $\mathbb{R}^2$, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 611–616.

[DS05] Y. Dolak and C. Schmeiser, *The Keller-Segel model with logistic sensitivity function and small diffusivity*, SIAM J. Appl. Math., 66 (2005), pp. 286–308.

[EZ91] M. Escobedo and E. Zuazua, *Large time behavior for convection-diffusion equations in* $\mathbf{R}^N$, J. Funct. Anal., 100 (1991), pp. 119–161.

[GT05] L. Gosse and G. Toscani, *Lagrangian numerical approximations to one-dimensional convolution-diffusion equations*, SIAM J. Sci. Comput., 28 (2006), pp. 1203–1227.

[Hor03] D. Horstmann, *From* 1970 *until present: The Keller-Segel model in chemotaxis and its consequences, Part* I, Jahresber. Deutsch. Math.-Verein., 105 (2003), pp. 103–165.

[Hor04] D. Horstmann, *From* 1970 *until present: The Keller-Segel model in chemotaxis and its consequences, Part* II, Jahresber. Deutsch. Math.-Verein., 106 (2004), pp. 51–69.

[HP01] T. Hillen and K. Painter, *Global existence for a parabolic chemotaxis model with prevention of overcrowding*, Adv. in Appl. Math., 26 (2001), pp. 280–301.

[Kne77] B. F. Knerr, *The porous medium equation in one dimension*, Trans. Amer. Math. Soc., 234 (1977), pp. 381–415.

[KR01] K. H. Karlsen and N. H. Risebro, *Convergence of finite difference schemes for viscous and inviscid conservation laws with rough coefficients*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 239–269.

[KR03] K. H. Karlsen and N. H. Risebro, *On the uniqueness and stability of entropy solutions of nonlinear degenerate parabolic equations with rough coefficients*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 1081–1104.

[KS70] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an*

instability, J. Theor. Biol., 26 (1970), pp. 399–415.

[LeV90] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, Switzerland, 1990.

[PH05] A. B. Potapov and T. Hillen, *Metastability in chemotaxis models*, J. Dynam. Differential Equations, 17 (2005), pp. 293–330.

[PW84] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Spinger-Verlag, New York, 1984.

[Qua03] A. Quarteroni, *Numerical Modelling for Differential Problems*, Springer, Milano, Italy, 2003.

[Váz92] J. L. Vázquez, *An introduction to the mathematical theory of the porous medium equation*, Shape Optimization and Free Boundaries (Montreal, PQ, 1990), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., Vol. 380, Kluwer Acad. Publ., Dordrecht, The Netherlands, 1992, pp. 347–389.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1316–1328

UNIQUENESS, RENORMALIZATION, AND SMOOTH APPROXIMATIONS FOR LINEAR TRANSPORT EQUATIONS*

FRANÇOIS BOUCHUT† AND GIANLUCA CRIPPA‡

Abstract. Transport equations arise in various areas of fluid mechanics, but the precise conditions on the vector field for them to be well-posed are still not fully understood. The renormalized theory of DiPerna and Lions for linear transport equations with an unsmooth coefficient uses the tools of approximation of an arbitrary weak solution by smooth functions, and also uses the renormalization property; that is, the possibility of writing an equation on a nonlinear function of the solution. Under some $W^{1,1}$ regularity assumption on the coefficient, well-posedness holds. In this paper, we establish that these properties are indeed equivalent to the uniqueness of weak solutions to the Cauchy problem, without any regularity assumption on the coefficient. Coefficients with unbounded divergence but with bounded compression are also considered.

Key words. linear transport equations with an unsmooth coefficient, renormalized solutions, approximation by smooth functions, coefficients of bounded compression

AMS subject classifications. 35R05, 35F99, 35L99

DOI. 10.1137/06065249X

1. Introduction.

In this paper we consider linear transport equations

$$\partial_t u + \operatorname{div}(bu) = 0 \qquad \text{in } (0,T)\times\mathbb{R}^d, \tag{1}$$

where $b(t,x)\in\mathbb{R}^d$ is the coefficient and u is scalar. Such equations arise in many areas of fluid mechanics, and a precise analysis of them is a key issue for the understanding of the particle flows in applications. In the present work, we give sharp results characterizing the well-posedness of transport equations. The question of well-posedness for the associated Cauchy problem for (1) has a well-known answer when b is continuous and Lipschitz continuous with respect to x, because of the Cauchy–Lipschitz theorem and the relation between (1) and the ordinary differential equation $dX/ds = b(s,X(s))$. When b is not smooth, the well-posedness is much more delicate. A general theory has been developed in [13] in the case when $b\in L^1((0,T),W^{1,1}_{loc}(\mathbb{R}^d))$, $\operatorname{div} b\in L^\infty$, and under some growth conditions on b. After some intermediate results (see in particular [5], [9], and [10]), the theory has been generalized in [2] to the case of only BV regularity for b instead of $W^{1,1}$. However, some recent counterexamples (as in [11] and [12], both inspired by [1]) show that there is not much room in which to weaken the regularity assumptions. Nevertheless, some questions remain open, such as the case of BD regularity for b (the symmetric part of $\nabla_x b$ is a measure, instead of the full matrix as in the BV case); see [8] and [4] for some partial results in this direction. For a detailed exposition and for a wider bibliography, the reader is referred to [3].

In this paper, we intend to give results of a different type that do not give directly the answer to the well-posedness problem, but rather give equivalent conditions for it to hold, without regularity assumptions on b. For simplicity we shall always assume

*Received by the editors February 20, 2006; accepted for publication (in revised form) July 3, 2006; published electronically December 15, 2006.
http://www.siam.org/journals/sima/38-4/65249.html

†DMA, CNRS, & École Normale Supérieure, 45 Rue d'Ulm, F-75230 Paris cedex 05, France (Francois.Bouchut@ens.fr).
‡Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy (g.crippa@sns.it).

that $b \in L^\infty((0,T)\times\mathbb{R}^d)$, and consider an L^2 framework. The approach of [13] and [2] relies on an approximation by convolution of a given weak solution to (1) and on the renormalized property; that is, if u solves (1) and if $\mathrm{div} b = 0$ (to simplify), then $\beta(u)$ also solves (1) for any suitable nonlinearity β. Theorem 2.1 states that such properties are indeed equivalent to the well-posedness of both forward and backward Cauchy problems, up to the fact that the smooth approximate solution (in the sense of the norm of the graph of the transport operator) is not necessarily given by convolution. Then, one can consider separately the two different issues of forward and backward uniqueness. Theorem 3.1 states that a characterization of backward uniqueness is the existence of a solution to the forward Cauchy problem that is approximable by smooth functions in the sense of the norm of the graph of the transport operator. Finally, we also consider the case of a coefficient b with unbounded divergence, but with *bounded compression*. We show that the previous results extend naturally to this case.

2. Forward-backward formulation.

THEOREM 2.1. *Let $b \in L^\infty((0,T)\times\mathbb{R}^d;\mathbb{R}^d)$ such that $\mathrm{div} b = 0$. Then the following statements are equivalent:*

(i) *b has the uniqueness property for weak solutions in $C([0,T];L^2(\mathbb{R}^d)-w)$ for both the forward and the backward Cauchy problems starting, respectively, from 0 and T; i.e., the only solutions in $C([0,T];L^2(\mathbb{R}^d)-w)$ to the problems*

$$\begin{cases} \partial_t u_F + \mathrm{div}(bu_F) = 0, \\ u_F(0,\cdot) = 0, \end{cases} \qquad \begin{cases} \partial_t u_B + \mathrm{div}(bu_B) = 0, \\ u_B(T,\cdot) = 0 \end{cases}$$

are $u_F \equiv 0$ and $u_B \equiv 0$.

(ii) *The Banach space*

$$\mathcal{F} := \left\{ \begin{array}{c} u \in C([0,T];L^2(\mathbb{R}^d)-w) \text{ such that} \\ \partial_t u + \mathrm{div}(bu) \in L^2((0,T)\times\mathbb{R}^d) \end{array} \right\} \tag{2}$$

with norm

$$\|u\|_{\mathcal{F}} := \|u\|_{B([0,T];L^2(\mathbb{R}^d))} + \|\partial_t u + \mathrm{div}(bu)\|_{L^2((0,T)\times\mathbb{R}^d)} \tag{3}$$

has the property that the space of functions in $C^\infty([0,T]\times\mathbb{R}^d)$ with compact support in x is dense in $\mathcal{F}$.

(iii) *Every weak solution in $C([0,T];L^2(\mathbb{R}^d)-w)$ of $\partial_t u + \mathrm{div}(bu) = 0$ lies in $C([0,T];L^2(\mathbb{R}^d)-s)$ and is a renormalized solution, i.e., for every function $\beta \in C^1(\mathbb{R};\mathbb{R})$ such that $|\beta'(s)| \le C(1+|s|)$ for some constant $C \ge 0$, one has $\partial_t\big(\beta(u)\big) + \mathrm{div}\big(b\beta(u)\big) = 0$ in $(0,T)\times\mathbb{R}^d$.*

In the statement of the theorem we used the notation $C([0,T];L^2(\mathbb{R}^d)-w)$ and $C([0,T];L^2(\mathbb{R}^d)-s)$ for the spaces of maps which are continuous from $[0,T]$ into $L^2(\mathbb{R}^d)$, endowed with the weak or the strong topology, respectively. We recall the classical fact that, up to a redefinition in a negligible set of times, every solution to (1) belongs to $C([0,T];L^2(\mathbb{R}^d)-w)$ (see, for example, Remark 3 in [3]).

Proof of Theorem 2.1.

(i) $\Rightarrow$ (ii).

Step 1. *Cauchy problem in $\mathcal{F}$.* It is easy to check that $\mathcal{F}$ is a Banach space, since L^2 and $B([0,T];L^2(\mathbb{R}^d))$ are Banach spaces (the latter denotes the space of bounded functions, with the supremum norm). We preliminarily show that for any $f \in L^2((0,T)\times\mathbb{R}^d)$ and $u^0 \in L^2(\mathbb{R}^d)$, the Cauchy problem

$$\begin{cases} \partial_t u + \mathrm{div}(bu) = f, \\ u(0,\cdot) = u^0 \end{cases} \tag{4}$$

has a unique solution in $\mathcal{F}$. We proceed by regularization. Consider a sequence of smooth vector fields $\{b_n\}_n$, with $b_n \to b$ a.e., b_n uniformly bounded in L^∞, and $\mathrm{div} b_n = 0$ for every n. Let u_n be the solution to the problem

$$\begin{cases} \partial_t u_n + \mathrm{div}(b_n u_n) = f, \\ u_n(0,\cdot) = u^0. \end{cases}$$

Then, by standard results on the smooth theory of transport equations (see, for example, [6]), we know that the solution u_n is unique in $C([0,T];L^2(\mathbb{R}^d))$ and is given by

$$u_n(t,x) = u^0(X_n(0,t,x)) + \int_0^t f(\tau, X_n(\tau,t,x))\,d\tau,$$

where $X_n(s,t,x)$ is the flow of b_n at time s, starting at the point x at time t, i.e., the solution to the ordinary differential equation

$$\begin{cases} \dfrac{dX_n}{ds}(s,t,x) = b_n(s, X_n(s,t,x)), \\ X_n(t,t,x) = x. \end{cases}$$

Recalling that $\mathrm{div} b_n = 0$, so that $X_n(s,t,\cdot)_\# \mathcal{L}^d = \mathcal{L}^d$ for every s and t (we denote by $\mathcal{L}^d$ the d-dimensional Lebesgue measure on $\mathbb{R}^d$), we can estimate the L^2 norm of $u_n(t,\cdot)$ as follows:

$$\begin{aligned} \|u_n(t,\cdot)\|_{L^2} &\le \|u^0(X_n(0,t,\cdot))\|_{L^2} + \int_0^t \|f(\tau, X_n(\tau,t,\cdot))\|_{L^2}\,d\tau \\ &\le \|u^0\|_{L^2} + \int_0^t \|f(\tau,\cdot)\|_{L^2}\,d\tau \\ &\le \|u^0\|_{L^2} + \sqrt{T}\|f\|_{L^2}. \end{aligned}$$

This implies that the sequence $\{u_n\}_n$ is equibounded in $C([0,T];L^2(\mathbb{R}^d))$. From the equation on u_n, we have also that for any $\varphi \in C_c^\infty(\mathbb{R}^d)$, $d/dt(\int u_n \varphi dx)$ is bounded in $L^2(0,T)$. We deduce that for any $\varphi \in L^2(\mathbb{R}^d)$, $\int u_n \varphi dx$ is uniformly in n equicontinuous in $[0,T]$. Thus, up to the passage to a subsequence (which does not depend on t), we can suppose that $u_n(t,\cdot) \rightharpoonup u(t,\cdot)$ in $L^2(\mathbb{R}^d) - w$, with $u \in C([0,T];L^2(\mathbb{R}^d) - w)$. By the semicontinuity of the norm with respect to weak convergence we also obtain that

$$\|u(t,\cdot)\|_{L^2} \le \|u^0\|_{L^2} + \sqrt{T}\|f\|_{L^2}. \tag{5}$$

Passing to the limit in the transport equation, we obtain that u solves the Cauchy problem

$$\begin{cases} \partial_t u + \mathrm{div}(bu) = f, \\ u(0,\cdot) = u^0. \end{cases}$$

Noticing that $\partial_t u + \mathrm{div}(bu) = f \in L^2$, we conclude that $u \in \mathcal{F}$. Uniqueness is clear: every solution to the Cauchy problem (4) is by definition a weak solution in $C([0,T];L^2(\mathbb{R}^d) - w)$ of the forward Cauchy problem with right-hand side, and thus by linearity, uniqueness is guaranteed by the forward part of assumption (i).

Step 2. *Density of smooth functions.* Define a linear operator

$$A: \begin{array}{ccc} \mathcal{F} & \to & L^2(\mathbb{R}^d)\times L^2((0,T)\times\mathbb{R}^d), \\ u & \mapsto & \big(u(0,\cdot),\ \partial_t u+\operatorname{div}(bu)\big). \end{array}$$

This operator in clearly bounded by the definition of the norm we have taken on $\mathcal{F}$. It is also a bijection because of Step 1, with continuous inverse because of (5). This means that A is an isomorphism, and thus we can identify $\mathcal{F}$ with the space $L^2(\mathbb{R}^d)\times L^2((0,T)\times\mathbb{R}^d)$, and its dual $\mathcal{F}^*$ with $L^2(\mathbb{R}^d)\times L^2((0,T)\times\mathbb{R}^d)$. Therefore, for every functional $L\in\mathcal{F}^*$, we can uniquely define $v_0\in L^2(\mathbb{R}^d)$ and $v\in L^2((0,T)\times\mathbb{R}^d)$ in such a way that

$$Lu=\int_{(0,T)\times\mathbb{R}^d}\big(\partial_t u+\operatorname{div}(bu)\big)v\,dtdx+\int_{\mathbb{R}^d}u(0,\cdot)v_0\,dx\qquad\text{for every } u\in\mathcal{F}.$$

We recall the classical fact that a subspace of a Banach space is dense if and only if every functional which is zero on the subspace is in fact identically zero. Then the density of smooth functions is equivalent to the following implication:

$$\begin{array}{c} \int_{(0,T)\times\mathbb{R}^d}\big(\partial_t u+\operatorname{div}(bu)\big)v\,dtdx+\int_{\mathbb{R}^d}u(0,\cdot)v_0\,dx=0 \\ \text{for every } u\in C^\infty([0,T]\times\mathbb{R}^d) \text{ with compact support in } x \\ \Longrightarrow v_0=0 \text{ and } v=0. \end{array} \tag{6}$$

If we first take u arbitrary but with compact support also in time, we obtain that

$$\int_{(0,T)\times\mathbb{R}^d}\big(\partial_t u+\operatorname{div}(bu)\big)v\,dtdx=0,$$

and since $\operatorname{div}b=0$, this is precisely the weak form of

$$\partial_t v+\operatorname{div}(bv)=0.$$

This implies that $v\in C([0,T];L^2(\mathbb{R}^d)-w)$. Now let χ be a cut-off function on $\mathbb{R}$, i.e., $\chi\in C_c^\infty(\mathbb{R})$, $\chi(z)=1$ for $|z|\le 1$, and $\chi(z)=0$ for $|z|\ge 2$. For every function $\varphi\in C_c^\infty(\mathbb{R}^d)$, take a function $\tilde u\in C^\infty([0,T]\times\mathbb{R}^d)$ with compact support in x such that $\tilde u(T,\cdot)=\varphi$. Then, testing in (6) with $u(t,x)=\tilde u(t,x)\chi\big((T-t)/\varepsilon\big)$, we obtain for $0<\varepsilon<T/2$

$$\begin{aligned} 0&=\int_{(0,T)\times\mathbb{R}^d}\left[\partial_t\left(\tilde u(t,x)\chi\left(\frac{T-t}{\varepsilon}\right)\right)+\operatorname{div}\left(b(t,x)\tilde u(t,x)\chi\left(\frac{T-t}{\varepsilon}\right)\right)\right]v(t,x)\,dtdx \\ &=\int_{(0,T)\times\mathbb{R}^d}\left[\partial_t\tilde u(t,x)+\operatorname{div}(b(t,x)\tilde u(t,x))\right]v(t,x)\chi\left(\frac{T-t}{\varepsilon}\right)dtdx \\ &\quad-\int_{(0,T)\times\mathbb{R}^d}\frac{1}{\varepsilon}\chi'\left(\frac{T-t}{\varepsilon}\right)\tilde u(t,x)v(t,x)\,dtdx. \end{aligned} \tag{7}$$

Letting $\varepsilon\to 0$, we observe that the first integral clearly converges to 0 since the support of $\chi\big((T-t)/\varepsilon\big)$ is contained in $[T-2\varepsilon,\,T+2\varepsilon]$. The second integral can be rewritten as

$$-\int_0^T\frac{1}{\varepsilon}\chi'\left(\frac{T-t}{\varepsilon}\right)\left[\int_{\mathbb{R}^d}\tilde u(t,x)v(t,x)\,dx\right]dt.$$

Now, since $\tilde{u}$ is smooth and $v \in C([0,T];L^2(\mathbb{R}^d)-w)$, the integral over $\mathbb{R}^d$ is a continuous function of t. Moreover, it is easy to check that

$$-\int_0^T \frac{1}{\varepsilon}\chi'\left(\frac{T-t}{\varepsilon}\right)dt = 1.$$

Therefore, coming back to (7) and letting $\varepsilon \to 0$ we get

$$0 = \int_{\mathbb{R}^d} \tilde{u}(T,x)v(T,x)\,dx = \int_{\mathbb{R}^d} \varphi(x)v(T,x)\,dx.$$

Since $\varphi \in C_c^\infty(\mathbb{R}^d)$ is arbitrary, we obtain $v(T,\cdot) = 0$. We conclude that $v \in C([0,T];L^2(\mathbb{R}^d)-w)$ solves the Cauchy problem

$$\begin{cases} \partial_t v + \operatorname{div}(bv) = 0, \\ v(T,\cdot) = 0. \end{cases}$$

Thus, by the backward part of the uniqueness assumption (i), we get that $v = 0$. Substituting in (6), we get that $\int_{\mathbb{R}^d} u(0,\cdot)v_0\,dx = 0$ for every $u \in C^\infty([0,T]\times\mathbb{R}^d)$ with compact support in space, and this implies that $v_0 = 0$. This concludes the proof of the implication (6), which ensures that (ii) holds.

(ii) $\Rightarrow$ (iii). Let $u \in C([0,T],L^2(\mathbb{R}^d)-w)$ satisfy $\partial_t u + \operatorname{div}(bu) = 0$. Then by (ii), there exists a sequence $\{u_n\}$ of functions in $C^\infty([0,T]\times\mathbb{R}^d)$ with compact support in space such that $\|u_n - u\|_{\mathcal{F}} \to 0$. In particular this gives that $u_n \to u$ in $B([0,T];L^2(\mathbb{R}^d))$; thus $u \in C([0,T],L^2(\mathbb{R}^d)-s)$. Then, define $f_n = \partial_t u_n + \operatorname{div}(bu_n) \in L^2((0,T)\times\mathbb{R}^d)$. By the definition of convergence in $\mathcal{F}$ we have that $f_n \to 0$ strongly in $L^2((0,T)\times\mathbb{R}^d)$. For every function β with the regularity stated we can apply the classical chain-rule, giving

$$\partial_t(\beta(u_n)) + \operatorname{div}(b\beta(u_n)) = \beta'(u_n)f_n.$$

The left-hand side clearly converges to $\partial_t(\beta(u)) + \operatorname{div}(b\beta(u))$ in the sense of distributions. According to the assumed bound on β', we have that the sequence $\beta'(u_n)$ is equibounded in $L^2_{loc}((0,T)\times\mathbb{R}^d)$; hence with the strong convergence of f_n we deduce that the right-hand side converges strongly in $L^1_{loc}((0,T)\times\mathbb{R}^d)$ to zero. This implies that $\partial_t\big(\beta(u)\big) + \operatorname{div}\big(b\beta(u)\big) = 0$.

(iii) $\Rightarrow$ (i). This step is classical. Let $u \in C([0,T],L^2(\mathbb{R}^d)-w)$ satisfy $\partial_t u + \operatorname{div}(bu) = 0$. According to (iii), u lies in $C([0,T],L^2(\mathbb{R}^d)-s)$, and applying the renormalization property with $\beta(u) = u^2$, we get

$$\partial_t(u^2) + \operatorname{div}(bu^2) = 0,$$

with $u^2 \in C([0,T],L^1(\mathbb{R}^d)-s)$. Testing this equation against smooth functions of the form $\psi(t)\varphi_R(x)$, where $\psi \in C_c^\infty((0,T))$ and $\varphi_R(x) = \varphi(x/R)$ with $\varphi \in C_c^\infty(\mathbb{R}^d)$ is a cut-off function equal to 1 on the ball of radius 1 and equal to 0 outside the ball of radius 2, we get

$$\int_0^T \left[\int_{\mathbb{R}^d} u^2\varphi\left(\frac{x}{R}\right)dx\right]\psi'(t)\,dt + \int_0^T \left[\int_{\mathbb{R}^d} bu^2\frac{1}{R}\nabla\varphi\left(\frac{x}{R}\right)dx\right]\psi(t)\,dt = 0.$$

Thus, we get in the sense of distributions in $(0,T)$

$$\frac{d}{dt}\int_{\mathbb{R}^d} u^2\varphi\left(\frac{x}{R}\right)dx = \int_{\mathbb{R}^d} bu^2\frac{1}{R}\nabla\varphi\left(\frac{x}{R}\right)dx.$$

Since the right-hand side is in $L^\infty(0,T)$ and since for every t it is bounded by $\frac{1}{R}\|b\|_{L^\infty_{t,x}}\|\nabla\varphi\|_{L^\infty_x}\|u(t,\cdot)\|^2_{L^2_x}$, letting $R\to+\infty$ we obtain

$$\frac{d}{dt}\int_{\mathbb{R}^d} u(t,x)^2\,dx = 0 \quad \text{in } (0,T).$$

Recalling that $u^2\in C([0,T],L^1(\mathbb{R}^d)-s)$, this yields $\int u(t,x)^2dx = cst$ on $[0,T]$, which implies uniqueness for both forward and backward Cauchy problems, proving (i). ∎

Remark 2.2 (well-posedness). The space $\mathcal{F}$ defined in (2) is a natural space for the study of the Cauchy problem (4). Whenever one of the statements of Theorem 2.1 is true, we have existence and uniqueness in $\mathcal{F}$ with the estimate (5), as shown in the proof. Moreover, every solution is renormalized and strongly continuous with respect to time, i.e., $u\in C([0,T];L^2(\mathbb{R}^d)-s)$. Overall, the following weak stability holds: If $\{f_n\}_n$ is a bounded sequence in $L^2((0,T)\times\mathbb{R}^d)$ which converges weakly to f, $\{u_n^0\}_n$ is a bounded sequence in $L^2(\mathbb{R}^d)$ which converges weakly to u^0, and $\{b_n\}_n$ is a bounded sequence in $L^\infty((0,T)\times\mathbb{R}^d)$ which converges strongly in L^1_{loc} to b and such that $\mathrm{div}b_n = 0$ for every n, then the solutions $\{u_n\}_n$ to

$$\partial_t u_n + \mathrm{div}(b_n u_n) = f_n, \qquad u_n(0,\cdot) = u_n^0$$

converge in $C([0,T];L^2(\mathbb{R}^d)-w)$ to the solution u to the Cauchy problem (4).

Remark 2.3 (L^p case). We can modify the summability exponent in the definition of the space $\mathcal{F}$. For every $p\in\,]1,\infty[$, define $\mathcal{F}_p$ as the space containing those functions $u\in C([0,T];L^p(\mathbb{R}^d)-w)$ that satisfy $\partial_t u+\mathrm{div}(bu)\in L^p((0,T)\times\mathbb{R}^d)$ and define the norm $\|\cdot\|_{\mathcal{F}_p}$ in the obvious way, which makes $\mathcal{F}_p$ a Banach space. Denoting by p' the conjugate exponent of p, i.e., $\frac{1}{p}+\frac{1}{p'}=1$, the following statements are equivalent:

(i) Smooth functions with compact support in x are dense in $\mathcal{F}_p$ and in $\mathcal{F}_{p'}$.

(ii) The vector field b has the forward uniqueness property for weak solutions in $C([0,T];L^p(\mathbb{R}^d)-w)$ and the backward uniqueness property for weak solutions in $C([0,T];L^{p'}(\mathbb{R}^d)-w)$.

Remark 2.4 (equivalent norms). According to the proof of Theorem 2.1, if one of the properties (i), (ii), and (iii) holds, then the norm of $\mathcal{F}$ is equivalent to the norm

$$\|u\|_{\mathcal{F},0} = \|u(0,\cdot)\|_{L^2(\mathbb{R}^d)} + \|\partial_t u+\mathrm{div}(bu)\|_{L^2((0,T)\times\mathbb{R}^d)}$$

(see the estimate (5)). In the same spirit, it is easy to prove that $\|\cdot\|_{\mathcal{F}}$ is in fact equivalent to every norm of the form

$$\|u\|_{\mathcal{F},s} = \|u(s,\cdot)\|_{L^2(\mathbb{R}^d)} + \|\partial_t u+\mathrm{div}(bu)\|_{L^2((0,T)\times\mathbb{R}^d)}$$

for $s\in[0,T]$.

Remark 2.5 (Depauw's counterexample). A simple modification (translation in time) of the counterexample constructed in [12] shows that the renormalization property is really linked to the uniqueness in *both* the forward and the backward Cauchy problems. In fact, we can construct a divergence-free vector field $b\in L^\infty((0,1)\times\mathbb{R}^2;\mathbb{R}^2)$ and a function $\bar u\in L^\infty(\mathbb{R}^2)$ such that

1. the backward Cauchy problem with datum $\bar u$ at time $t=1$ has a unique solution, which is, however, *not* renormalized and *not* strongly continuous with respect to time;

2. the forward Cauchy problem with datum 0 at time $t=0$ has more than one solution;

3. the unique solution $u(t,x)$ to the backward Cauchy problem with datum $\bar{u}$ at time $t=1$ satisfies

$$\begin{cases} |u(t,x)| = 0 & \text{for } 0 \le t \le 1/2, \\ |u(t,x)| = 1 & \text{for } 1/2 < t \le 1; \end{cases}$$

hence the equivalence of the norms in Remark 2.4 does *not* hold.

Remark 2.6 (the Sobolev and the BV cases). In the case of a vector field with Sobolev regularity with respect to the space variable, $b \in L^1((0,T); W^{1,p'}_{loc}(\mathbb{R}^d))$ with $1 < p < \infty$, it is almost possible to prove that the natural regularization by convolution with respect to the space variable of $u \in \mathcal{F}_p$ (see Remark 2.3) converges to u with respect to $\|\cdot\|_{\mathcal{F}_p}$. Indeed, let η_ε be a standard convolution kernel in $\mathbb{R}^d$ and set $u_\varepsilon = u * \eta_\varepsilon$. We can compute

$$\begin{aligned} &\partial_t u + \operatorname{div}(bu) - \partial_t u_\varepsilon - \operatorname{div}(bu_\varepsilon) \\ &\quad = \big[\partial_t u + \operatorname{div}(bu)\big] - \big[\partial_t u + \operatorname{div}(bu)\big] * \eta_\varepsilon + \big[\operatorname{div}(bu) * \eta_\varepsilon - \operatorname{div}(bu_\varepsilon)\big]. \end{aligned}$$

Then the convergence of u_ε to u with respect to $\|\cdot\|_{\mathcal{F}_p}$ is equivalent to the strong convergence in $L^p((0,T)\times\mathbb{R}^d)$ to zero of the *commutator*

$$r_\varepsilon = \operatorname{div}(bu) * \eta_\varepsilon - \operatorname{div}(bu_\varepsilon).$$

The results of [13] ensure this strong convergence for every convolution kernel η_ε, except that it holds in L^1_{loc} instead of L^p. We need also a regularization with respect to time and a cut-off in order to get the density property in Theorem 2.1(ii), but this means that our strategy is more or less "equivalent" to the one of [13], in the framework of Sobolev vector fields. However, the situation is different in the BV case studied in [2]. In general, the commutator r_ε is not expected to converge strongly to zero; our result shows that, even in this case, there exists some smooth approximation of the solution, but it is less clear how to construct it in an explicit way.

Remark 2.7 (strong continuity condition). The condition of continuity with values in strong L^2 in Theorem 2.1(iii) cannot be removed; otherwise the equivalence with (i) fails. This can be seen again with Depauw's counterexample with singularity at time $t=0$. In this case all weak solutions are renormalized in $(0,T)\times\mathbb{R}^d$ since b is locally BV in x, but uniqueness of weak solutions does not hold. Another remark is that, in general, a renormalized solution does not need to be continuous with values in strong L^2, even inside the interval, as the following counterexample shows. On the interval $(-1,1)$, take for b the one of Depauw's counterexample in $(0,1)$ (with singularity at 0), and define on $(-1,0)$ the vector field as $b(t,x) = -b(-t,x)$. Consider then the weak solution u with value 0 at $t=0$, which we extend on $(-1,0)$ by $u(t,x) = u(-t,x)$. Then u is a renormalized solution on $(-1,1)$ but is not strongly continuous at $t=0$.

3. One-way formulation.

THEOREM 3.1. *Let $b \in L^\infty((0,T)\times\mathbb{R}^d;\mathbb{R}^d)$ such that* $\operatorname{div} b \in L^\infty((0,T)\times\mathbb{R}^d)$, *and let $c \in L^\infty((0,T)\times\mathbb{R}^d)$. Define the Banach space $\mathcal{F}$ and its norm $\|\cdot\|_{\mathcal{F}}$ as in* (2)–(3). *Moreover, define $\mathcal{F}^0 \subset \mathcal{F}$ as the closure (with respect to $\|\cdot\|_{\mathcal{F}}$) of the subspace of functions in $C^\infty([0,T]\times\mathbb{R}^d)$ with compact support in x. Then the following statements are equivalent:*

(i) *For every $u^0 \in L^2(\mathbb{R}^d)$ and every $f \in L^2((0,T)\times\mathbb{R}^d)$ there exists a solution $u \in \mathcal{F}^0$ to the Cauchy problem*

$$\begin{cases} \partial_t u + \operatorname{div}(bu) + cu = f, \\ u(0,\cdot) = u^0, \end{cases} \qquad u \in \mathcal{F}^0.$$

(ii) *There is uniqueness for weak solutions in* $C([0,T];L^2(\mathbb{R}^d)-w)$ *for the backward dual Cauchy problem starting from* T*; i.e., the only function* v *belonging to* $C([0,T];L^2(\mathbb{R}^d)-w)$ *which solves*

$$\begin{cases} \partial_t v + b\cdot\nabla v - cv = 0, \\ v(T,\cdot) = 0 \end{cases}$$

is $v \equiv 0$.

Here and further on, the advection term $b\cdot\nabla v$ is defined according to $b\cdot\nabla v \equiv \mathrm{div}(bv) - v\,\mathrm{div}b$, which makes sense since $\mathrm{div}b \in L^\infty$.

Remark 3.2. The two statements in Theorem 3.1 are really the "nontrivial" properties relative to the vector field b. In general, there is always uniqueness in $\mathcal{F}^0$ (see Step 1 in the proof) and there is always existence of weak solutions in $\mathcal{F}$ (this can be easily proved by regularization, as in the first step of the proof of Theorem 2.1).

Before proving the theorem, we recall the following standard result of functional analysis (see, for example, Theorems II.19 and II.20 of [7]).

LEMMA 3.3. *Let* E *and* F *be Banach spaces and let* $L : E \to F$ *be a bounded linear operator. Denote by* $L^* : F^* \to E^*$ *the adjoint operator, defined by*

$$\langle v, Lu\rangle_{F^*,F} = \langle L^*v, u\rangle_{E^*,E} \qquad \text{for every } u\in E \text{ and } v\in F^*.$$

Then

(i) L *is surjective if and only if* L^* *is injective and with closed image;*

(ii) L^* *is surjective if and only if* L *is injective and with closed image.*

Proof of Theorem 3.1.

Step 1. *An energy estimate in* $\mathcal{F}^0$. In this first step we prove that for every $u\in\mathcal{F}^0$ the following energy estimate holds:

(8)
$$\|u(t,\cdot)\|_{L^2_x} \le \left(\|u(0,\cdot)\|_{L^2_x} + \sqrt{T}\|\partial_t u + \mathrm{div}(bu) + cu\|_{L^2_{t,x}}\right)\exp\left(T\left\|c+\frac{1}{2}\mathrm{div}b\right\|_{L^\infty_{t,x}}\right).$$

Let us first prove the estimate for u smooth with compact support in x. We define

$$f = \partial_t u + \mathrm{div}(bu) + cu,$$

and we multiply this relation by u, giving

$$\partial_t\frac{u^2}{2} + \mathrm{div}\left(b\frac{u^2}{2}\right) + \left(c+\frac{1}{2}\mathrm{div}b\right)u^2 = fu.$$

For justifying the previous identity, we used the Leibnitz rule

$$\partial_i(H\psi) = \psi\partial_i H + H\partial_i\psi, \tag{9}$$

valid for $\psi\in C^\infty$ and H any distribution. Then, integrating over $x\in\mathbb{R}^d$ we get in the sense of distributions in $(0,T)$

$$\frac{d}{dt}\int_{\mathbb{R}^d} u(t,x)^2dx = 2\int_{\mathbb{R}^d} fu\,dx - 2\int_{\mathbb{R}^d}\left(c+\frac{1}{2}\mathrm{div}b\right)u^2dx.$$

Therefore, we get for a.e. $t\in(0,T)$

$$\left|\frac{d}{dt}\int_{\mathbb{R}^d} u(t,x)^2dx\right| \le 2\|f(t,\cdot)\|_{L^2_x}\|u(t,\cdot)\|_{L^2_x} + 2\left\|\left(c+\frac{1}{2}\mathrm{div}b\right)(t,\cdot)\right\|_{L^\infty_x}\|u(t,\cdot)\|^2_{L^2_x}.$$

This differential inequality can be easily integrated, obtaining

$$\|u(t,\cdot)\|_{L^2_x} \le \|u(0,\cdot)\|_{L^2_x} \exp\left(\int_0^t \left\|\left(c+\frac{1}{2}\mathrm{div}b\right)(s,\cdot)\right\|_{L^\infty_x} ds\right) + \int_0^t \|f(s,\cdot)\|_{L^2_x} \exp\left(\int_s^t \left\|\left(c+\frac{1}{2}\mathrm{div}b\right)(\tau,\cdot)\right\|_{L^\infty_x} d\tau\right) ds,$$

which clearly implies (8). In the general case of $u \in \mathcal{F}^0$, we can find approximations u_n smooth with compact support such that $\|u_n - u\|_{\mathcal{F}} \to 0$, and we obtain the estimate (8) at the limit.

Step 2. *The operator* A^0. As in the proof of Theorem 2.1, we consider the linear operator

$$A^0 : \begin{array}{ccc} \mathcal{F}^0 & \to & L^2(\mathbb{R}^d) \times L^2((0,T)\times\mathbb{R}^d), \\ u & \mapsto & \big(u(0,\cdot),\ \partial_t u + \mathrm{div}(bu) + cu\big). \end{array}$$

Since we can estimate

$$\begin{aligned} \|A^0 u\|_{L^2_x \times L^2_{t,x}} &= \|u(0,\cdot)\|_{L^2_x} + \|\partial_t u + \mathrm{div}(bu) + cu\|_{L^2_{t,x}} \\ &\le \|u\|_{B_t(L^2_x)} + \|\partial_t u + \mathrm{div}(bu)\|_{L^2_{t,x}} + \|c\|_{L^\infty_{t,x}}\sqrt{T}\|u\|_{B_t(L^2_x)} \\ &\le \big(1 + \|c\|_{L^\infty_{t,x}}\sqrt{T}\big)\|u\|_{\mathcal{F}}, \end{aligned}$$

we deduce that A^0 is a bounded operator. Next, the energy estimate established in the first step gives that for any $u \in \mathcal{F}^0$,

$$\|u\|_{B_t(L^2_x)} \le \exp\left(T\left\|c+\frac{1}{2}\mathrm{div}b\right\|_{L^\infty_{t,x}}\right)\max(1,\sqrt{T})\|A^0 u\|_{L^2_x\times L^2_{t,x}}.$$

But we have

$$\|\partial_t u + \mathrm{div}(bu)\|_{L^2_{t,x}} \le \|\partial_t u + \mathrm{div}(bu) + cu\|_{L^2_{t,x}} + \|c\|_{L^\infty_{t,x}}\sqrt{T}\|u\|_{B_t(L^2_x)},$$

and we conclude that

$$\|u\|_{\mathcal{F}} \le C\,\|A^0 u\|_{L^2_x\times L^2_{t,x}}, \qquad u \in \mathcal{F}^0. \tag{10}$$

This means that A^0 is injective and with closed image. Notice that the injectivity of A^0 is equivalent to the fact that the only solution $u \in \mathcal{F}^0$ to

$$\begin{cases} \partial_t u + \mathrm{div}(bu) + cu = 0, \\ u(0,\cdot) = 0 \end{cases}$$

is $u \equiv 0$.

Step 3. *Proof of the equivalence of the two statements.* Since by Step 2, A^0 is injective with closed image, we can apply Lemma 3.3(ii) to get the surjectivity of the adjoint operator $(A^0)^* : L^2(\mathbb{R}^d)\times L^2((0,T)\times\mathbb{R}^d) \to (\mathcal{F}^0)^*$. We recall that the adjoint operator is characterized by the condition

$$\begin{aligned} \langle (A^0)^*(v_0,v), u\rangle &= \langle (v_0,v), A^0 u\rangle \\ &= \int_{\mathbb{R}^d} v_0 u(0,\cdot)\,dx + \int_{(0,T)\times\mathbb{R}^d} v\big(\partial_t u + \mathrm{div}(bu) + cu\big)\,dtdx, \end{aligned} \tag{11}$$

for $(v_0, v) \in L^2(\mathbb{R}^d) \times L^2((0,T) \times \mathbb{R}^d)$ and $u \in \mathcal{F}^0$. Since $(A^0)^*$ is surjective, in particular it has closed image. Therefore, applying Lemma 3.3(i) we get the equivalence between surjectivity of A^0 and injectivity of $(A^0)^*$.

It is clear that the surjectivity of the operator A^0 is equivalent to the existence of solutions in $\mathcal{F}^0$ (statement (i)). Therefore, it remains only to characterize the injectivity of $(A^0)^*$. Recalling the definition of $\mathcal{F}^0$ as the closure of the set of smooth functions with compact support in x, and recalling the characterization of the adjoint operator given in (11), we obtain that the injectivity of $(A^0)^*$ is equivalent to the following implication:

$$\begin{aligned} &\textstyle\int_{(0,T)\times\mathbb{R}^d} \big(\partial_t u + \operatorname{div}(bu) + cu\big) v \, dtdx + \int_{\mathbb{R}^d} u(0,\cdot) v_0 \, dx = 0 \\ &\text{for every } u \in C^\infty([0,T] \times \mathbb{R}^d) \text{ with compact support in } x \\ &\qquad\qquad \Longrightarrow v_0 = 0 \text{ and } v = 0. \end{aligned} \tag{12}$$

Arguing as in Step 2 of the proof of Theorem 2.1, and eventually testing the integral condition with smooth functions of the form $u(t,x) = \chi(t/\varepsilon)\tilde{u}(t,x)$ (using the same notation as in the proof of Theorem 2.1), we obtain that the following two properties are equivalent for given $v_0 \in L^2(\mathbb{R}^d)$ and $v \in L^2((0,T) \times \mathbb{R}^d)$:

1. For every $u \in C^\infty([0,T] \times \mathbb{R}^d)$ with compact support in x we have

$$\int_{(0,T)\times\mathbb{R}^d} \big(\partial_t u + \operatorname{div}(bu) + cu\big) v \, dtdx + \int_{\mathbb{R}^d} u(0,\cdot) v_0 \, dx = 0.$$

2. $v \in C([0,T]; L^2(\mathbb{R}^d) - w)$, $v_0 = v(0,\cdot)$, and v is a weak solution of the backward dual Cauchy problem

$$\begin{cases} \partial_t v + b \cdot \nabla v - cv = 0, \\ v(T,\cdot) = 0. \end{cases}$$

Therefore we deduce that the implication (12) is equivalent to the uniqueness of weak solutions in $C([0,T]; L^2(\mathbb{R}^d) - w)$ of the backward dual Cauchy problem, i.e., statement (ii). □

Remark 3.4 (time inversion). By reversing the direction of time, we see that there is existence for the backward Cauchy problem in $\mathcal{F}^0$ if and only if there is uniqueness for weak solutions to the forward dual Cauchy problem.

Remark 3.5 (approximation by smooth functions and renormalization). Solutions in $\mathcal{F}^0$ lie in $C([0,T], L^2(\mathbb{R}^d) - s)$ and are renormalized: this can be seen as in the proof of the implication (ii) $\Rightarrow$ (iii) of Theorem 2.1, using the density of smooth functions in $\mathcal{F}^0$. Conversely, it is possible that some renormalized solutions do not belong to $\mathcal{F}^0$. This can be seen by noticing that one can have several renormalized solutions to the same Cauchy problem (see an example in [13]), while there is always uniqueness in $\mathcal{F}^0$. Another difference between the criterion of approximation by smooth functions and the renormalization property is that $\mathcal{F}^0$ is a vector space, while in general, renormalized solutions are not a vector space.

Remark 3.6 (Depauw's example again). We notice that forward uniqueness and backward uniqueness of weak solutions are really distinct properties: the example described in Remark 2.5 shows how to construct bounded divergence-free vector fields with backward uniqueness, but not forward uniqueness, and vice versa.

4. Vector fields of bounded compression. We shall say that a vector field $b \in L^\infty((0,T) \times \mathbb{R}^d; \mathbb{R}^d)$ has *bounded compression* if there exists a function

$\rho \in C([0,T];L^\infty(\mathbb{R}^d)-w^*)$, with $0<C^{-1}\le\rho\le C<\infty$ for some constant $C>0$, such that the identity

$$\partial_t\rho+\mathrm{div}(b\rho)=0 \tag{13}$$

holds in the sense of distributions in $(0,T)\times\mathbb{R}^d$. We remark that every vector field b with bounded divergence has bounded compression (if b is smooth, take for ρ the Jacobian determinant of the flow generated by b, $\rho(t,x)=\det\nabla_x X(0,t,x)$, which is bounded since $\rho(t,x)=\exp\big[-\int_0^t(\mathrm{div}b)(\sigma,X(\sigma,t,x))d\sigma\big]$, where $X(s,t,x)$ satisfies $dX(s,t,x)/ds=b(s,X(s,t,x))$, $X(t,t,x)=x$), but in general a vector field of bounded compression does not need to have absolutely continuous divergence.

THEOREM 4.1. *Let $b\in L^\infty((0,T)\times\mathbb{R}^d;\mathbb{R}^d)$ be a vector field of bounded compression, and fix an associated function $\rho\in C([0,T];L^\infty(\mathbb{R}^d)-w^*)$. We define the Banach space $\mathcal{F}$ and its norm $\|\cdot\|_{\mathcal{F}}$ as in* (2)–(3). *Let $\mathcal{F}^1\subset\mathcal{F}$ be the closure of*

$$\big\{\rho\varphi\ :\ \ \varphi\in C^\infty([0,T]\times\mathbb{R}^d)\ \textit{with compact support in } x\big\}$$

with respect to $\|\cdot\|_{\mathcal{F}}$. Then the following statements are equivalent:

(i) *For every $u^0\in L^2$ and every $f\in L^2$ there exists a solution $u\in\mathcal{F}^1$ to the Cauchy problem*

$$\begin{cases}\partial_t u+\mathrm{div}(bu)=f,\\ u(0,\cdot)=u^0,\end{cases}\qquad u\in\mathcal{F}^1.$$

(ii) *There is uniqueness for weak solutions in $C([0,T];L^2(\mathbb{R}^d)-w)$ for the backward dual Cauchy problem starting from T; i.e., the only function ρ belonging to $C([0,T];L^2(\mathbb{R}^d)-w)$ which solves*

$$\begin{cases}\partial_t(\rho v)+\mathrm{div}(b\rho v)=0,\\ \rho(T,\cdot)v(T,\cdot)=0\end{cases}$$

is $\rho v\equiv 0$.

Remark 4.2. In this context, the equation $\partial_t(\rho v)+\mathrm{div}(b\rho v)=0$ is dual to the equation $\partial_t u+\mathrm{div}(bu)=0$, since we can write (formally, since it is not possible to give a meaning to the product $b\cdot\nabla v$ without a condition of absolute continuity of $\mathrm{div}b$)

$$\partial_t(\rho v)+\mathrm{div}(b\rho v)=\rho\big(\partial_t v+b\cdot\nabla v\big).$$

Proof of Theorem 4.1. The proof is very close to that of Theorem 3.1; thus we shall sometimes omit the technical details.

Step 1. *An energy estimate in $\mathcal{F}^1$.* We preliminarily prove that for every $u\in\mathcal{F}^1$ the following estimate holds (C is the constant related to the function ρ):

$$\|u\|_{B_t(L^2_x)}\le C\|u(0,\cdot)\|_{L^2_x}+C\sqrt{T}\|\partial_t u+\mathrm{div}(bu)\|_{L^2_{t,x}}. \tag{14}$$

Fix a smooth function φ with compact support in $\mathbb{R}^d$, and define $f=\partial_t(\rho\varphi)+\mathrm{div}(b\rho\varphi)=\rho(\partial_t\varphi+b\cdot\nabla\varphi)$ (use the Leibniz rule (9) and formula (13)). We deduce with the same argument that $2\varphi f=\rho(\partial_t\varphi^2+b\cdot\nabla\varphi^2)=\partial_t(\rho\varphi^2)+\mathrm{div}(b\rho\varphi^2)$. Thus, we get the following estimate in the sense of distributions in $(0,T)$:

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^d}\rho(t,x)\varphi(t,x)^2\,dx &= 2\int_{\mathbb{R}^d}\varphi(t,x)f(t,x)\,dx\\ &\le 2\|f(t,\cdot)\|_{L^2_x}\|\varphi(t,\cdot)\|_{L^2_x}\\ &\le 2\sqrt{C}\|f(t,\cdot)\|_{L^2_x}\left[\int_{\mathbb{R}^d}\rho(t,x)\varphi(t,x)^2\,dx\right]^{1/2}.\end{aligned}$$

By integration with respect to time this implies

$$\left[\int_{\mathbb{R}^d}\rho(t,x)\varphi(t,x)^2\,dx\right]^{1/2}\le\left[\int_{\mathbb{R}^d}\rho(0,x)\varphi(0,x)^2\,dx\right]^{1/2}+\sqrt{C}\int_0^t\|f(s,\cdot)\|_{L^2_x}\,ds.$$

Using the fact that $C^{-1}\le\rho\le C$ we deduce

$$\frac{1}{\sqrt{C}}\|\rho(t,\cdot)\varphi(t,\cdot)\|_{L^2_x}\le\sqrt{C}\|\rho(0,\cdot)\varphi(0,\cdot)\|_{L^2_x}+\sqrt{C}\int_0^t\|f(s,\cdot)\|_{L^2_x}\,ds,$$

and thus

$$\|\rho(t,\cdot)\varphi(t,\cdot)\|_{L^2_x}\le C\|\rho(0,\cdot)\varphi(0,\cdot)\|_{L^2_x}+C\sqrt{T}\|\partial_t(\rho\varphi)+\operatorname{div}(b\rho\varphi)\|_{L^2_{t,x}}. \tag{15}$$

But by definition of $\mathcal{F}^1$, the validity of (15) for every smooth function φ with compact support in x implies the validity of (14) for every function $u\in\mathcal{F}^1$.

Step 2. *The operator* A^1. We define the linear operator

$$A^1:\begin{array}{ccc}\mathcal{F}^1 & \to & L^2(\mathbb{R}^d)\times L^2((0,T)\times\mathbb{R}^d),\\ u & \mapsto & \big(u(0,\cdot),\ \partial_t u+\operatorname{div}(bu)\big).\end{array}$$

It is immediate to see that the operator A^1 is bounded. Using the energy estimate (14) it is also immediate to check that $\|u\|_{\mathcal{F}}\le\tilde{C}\|A^1u\|$, and therefore that A^1 is injective with closed image. Applying Lemma 3.3(ii) we obtain that the adjoint operator

$$(A^1)^*:L^2(\mathbb{R}^d)\times L^2((0,T)\times\mathbb{R}^d)\to(\mathcal{F}^1)^*$$

is surjective. The adjoint operator is characterized by the identity

$$\begin{aligned}\langle(A^1)^*(v_0,v),u\rangle &= \langle(v_0,v),A^1u\rangle\\ &= \int_{\mathbb{R}^d}v_0u(0,\cdot)\,dx+\int_{(0,T)\times\mathbb{R}^d}v\big(\partial_tu+\operatorname{div}(bu)\big)\,dtdx\end{aligned} \tag{16}$$

for $(v_0,v)\in L^2(\mathbb{R}^d)\times L^2((0,T)\times\mathbb{R}^d)$ and $u\in\mathcal{F}^1$.

Step 3. *Proof of the equivalence of the two statements.* Statement (i) (existence of solutions in $\mathcal{F}^1$) is the surjectivity of the operator A^1, which is equivalent (applying Lemma 3.3(i) and using the surjectivity of $(A^1)^*$ proved in Step 2) to the injectivity of $(A^1)^*$. But recalling the characterization (16) and the definition of the space $\mathcal{F}^1$, we see that the injectivity of $(A^1)^*$ is equivalent to the following implication for $v_0\in L^2(\mathbb{R}^d)$ and $v\in L^2((0,T)\times\mathbb{R}^d)$:

$$\begin{aligned}&\textstyle\int_{(0,T)\times\mathbb{R}^d}\big(\partial_t(\rho\varphi)+\operatorname{div}(b\rho\varphi)\big)v\,dtdx+\int_{\mathbb{R}^d}\rho(0,\cdot)\varphi(0,\cdot)v_0\,dx=0\\ &\quad\text{for every }\varphi\in C^\infty([0,T]\times\mathbb{R}^d)\text{ with compact support in }x\\ &\qquad\Longrightarrow v_0=0\text{ and }v=0.\end{aligned} \tag{17}$$

Arguing as in Step 3 of the proof of Theorem 3.1 we obtain that the following two properties are equivalent:

1. For every $\varphi \in C^\infty([0,T]\times\mathbb{R}^d)$ with compact support in x we have

$$\int_{(0,T)\times\mathbb{R}^d} \big(\partial_t(\rho\varphi) + \operatorname{div}(b\rho\varphi)\big)v\,dtdx + \int_{\mathbb{R}^d} \rho(0,\cdot)\varphi(0,\cdot)v_0\,dx = 0.$$

2. $\rho v \in C([0,T];L^2(\mathbb{R}^d)-w)$, $\rho(0,\cdot)v_0 = \rho(0,\cdot)v(0,\cdot)$, and ρv is a weak solution of the backward dual Cauchy problem

$$\begin{cases} \partial_t(\rho v) + \operatorname{div}(b\rho v) = 0, \\ \rho(T,\cdot)v(T,\cdot) = 0. \end{cases}$$

Then we deduce that implication (17) is equivalent to statement (ii), and this concludes the proof of the theorem. □

Acknowledgment. The second author thanks the École Normale Supérieure de Paris for the kind hospitality during the preparation of this work.

REFERENCES

[1] M. Aizenman, *On vector fields as generators of flows: A counterexample to Nelson's conjecture*, Ann. Math. (2), 107 (1978), pp. 287–296.

[2] L. Ambrosio, *Transport equation and Cauchy problem for BV vector fields*, Invent. Math., 158 (2004), pp. 227–260.

[3] L. Ambrosio and G. Crippa, *Existence, Uniqueness, Stability, and Differentiability Properties of the Flow Associated to Weakly Differentiable Vector Fields*, Preprint, Scuola Normale Superiore and Department of Mathematics, Pisa, Italy, 2006. Available online at http://cvgmt.sns.it/papers/ambcri06/.

[4] L. Ambrosio, G. Crippa, and S. Maniglia, *Traces and fine properties of a BD class of vector fields and applications*, Ann. Fac. Sci. Toulouse Math. (6), 14 (2005), pp. 527–561.

[5] F. Bouchut, *Renormalized solutions to the Vlasov equation with coefficients of bounded variation*, Arch. Ration. Mech. Anal., 157 (2001), pp. 75–90.

[6] F. Bouchut, F. Golse, and M. Pulvirenti, *Kinetic Equations and Asymptotic Theory*, Gauthier-Villars, Éditions Scientifiques et Médicales Elsevier, Paris, 2000.

[7] H. Brezis, *Analyse Fonctionnelle. Théorie et Applications*, Masson, Paris, 1983.

[8] I. Capuzzo Dolcetta and B. Perthame, *On some analogy between different approaches to first order PDE's with nonsmooth coefficients*, Adv. Math. Sci. Appl., 6 (1996), pp. 689–703.

[9] F. Colombini and N. Lerner, *Sur les champs de vecteurs peu réguliers*, Séminaire: Équations aux Dérivées Partielles, Exp. No. XIV, École Polytech., Palaiseau, 2001.

[10] F. Colombini and N. Lerner, *Uniqueness of continuous solutions for BV vector fields*, Duke Math. J., 111 (2002), pp. 357–384.

[11] F. Colombini, T. Luo, and J. Rauch, *Uniqueness and nonuniqueness for nonsmooth divergence free transport*, Séminaire: Équations aux Dérivées Partielles, Exp. No. XXII, École Polytech., Palaiseau, 2003.

[12] N. Depauw, *Non-unicité du transport par un champ de vecteurs presque BV*, Séminaire: Équations aux Dérivées Partielles, Exp. No. XIX, École Polytech., Palaiseau, 2003.

[13] R. J. DiPerna and P. L. Lions, *Ordinary differential equations, transport theory, and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1329–1342

SCATTERING OF ELECTROMAGNETIC WAVES BY THIN DIELECTRIC PLANAR STRUCTURES*

HABIB AMMARI†, HYEONBAE KANG‡, AND FADIL SANTOSA§

Abstract. The problem of scattering of electromagnetic waves by a thin dielectric planar structure is considered. If the squared index of refraction of the scatterer scales as $1/h$, where h is the thickness of the structure, we show that an approximate solution to the scattering problem can be obtained by a perturbation method. The approximate solution consists of two terms, the zeroth order term and the first order corrector, both of which can be found by solving 2-D integral equations for 3-D problems. We provide error estimates for the approximation. Therefore, the method described in this work can be viewed as a computational approach which can potentially greatly simplify scattering calculations for problems involving thin scatterers.

Key words. electromagnetic scattering, Maxwell's equations, approximate solution, asymptotics, error estimates

AMS subject classifications. 35B40, 65R20, 78A45

DOI. 10.1137/040618382

1. Introduction.

To understand how light behaves in a thin-film structure with a high-refractive-index contrast, we need to solve the time-harmonic Maxwell's equation. The structure is usually surrounded by air. Thus, the domain in which Maxwell's equation must be solved will be all of $\mathbb{R}^3$. The thin-film structure is modeled by prescribing the index of refraction to a subdomain of $\mathbb{R}^3$.

The standard approach for performing the required simulation of wave propagation in such a structure is the finite-difference time-domain (FDTD) method [2], with absorbing boundary conditions. While the computation proceeds in a straightforward manner, it can be very time consuming.

In this paper, we consider a scattering problem in which an incident wave strikes a thin-film structure. Our goal is to derive an asymptotic expansion for the solution of Maxwell's equation. The approach is to exploit the fact that the scattering structure is thin, and the index of refraction is large. In this work, we assume that the squared index of refraction is $O(1/h)$, where h is the scatterer thickness. Under a regularity assumption on the index of refraction, we show that an approximate solution, consisting of a leading term and first order (in h) corrector, can be obtained by solving a 2-D integral equation for a 3-D problem. Thus, the expansion described here can be thought of as a computational approach.

This work is an extension of [5] which considered the scattering problem in the context of the scalar Helmholtz equation. In that work, a perturbation approach

*Received by the editors November 5, 2004; accepted for publication (in revised form) July 10, 2006; published electronically December 18, 2006. This research was partially supported by CNRS-KOSEF grant 14889 and F01-2003-000-00016-0 and Korea Science and Engineering Foundation grant R02-2003-000-10012-0.

http://www.siam.org/journals/sima/38-4/61838.html

†Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr).

‡School of Mathematical Sciences, Seoul National University, Seoul 151-747, Korea (hkang@math.snu.ac.kr).

§School of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church Street, SE, Minneapolis, MN 55455 (santosa@math.umn.edu). The research of this author was partially supported by the National Science Foundation.

based on the thickness h is developed to reduce the complexity of the computation by one dimension. The present paper addresses the case of Maxwell's equations, which is the correct model for the propagation phenomena. We start with the Lippmann–Schwinger formulation of the electromagnetic scattering problem which involves a 3-D (volume) integral equation. Based on delicate estimates on the kernels involved in this integral formulation, we carefully derive the asymptotic expansion of the electric field in h and provide rigorous error estimates. Using our asymptotic expansion in h, we then simplify the scattering calculations to solving a 2-D (surface) integral equation.

The paper is organized as follows. We give a description of the problem we wish to solve in the next section. The perturbation approach and its rigorous justification are presented in section 3. Our justification uses a nontrivial bound on the $\mathcal{C}^{1,\alpha}$-norm of the electric field proved in Appendix A.

The results in this paper are expected to lead to a very effective computational algorithm for solving the electromagnetic scattering problem for thin-film devices in the parameter regime, where the approximation is accurate.

2. Problem statement. Let E and H denote the electric and magnetic fields. The governing Maxwell's equations are given by

$$\begin{cases} \nabla \times E = ikH, \\ \nabla \times H = -ikn(x,z)E, \end{cases}$$

where $k > 0$ is the normalized frequency. The total field consists of the incident field and the scattered field. We write $E = E^i + E^s$, where E^i is a given incident wave. The scattered electric field E^s satisfies the Sommerfeld radiation condition. We call the function $n(x,z)$ the *squared index of refraction*.

Here $x = (x_1, x_2)$ and z is the third direction. The squared index of refraction $n(x,z)$ is defined by

$$n(x,z) := \begin{cases} 1 \quad \text{for } |z| > \dfrac{h}{2}, \\ \dfrac{n_0(x)}{h} \quad \text{for } |z| < \dfrac{h}{2}. \end{cases}$$

Therefore, the thin scattering structure has been incorporated into the definition of $n(x,z)$. Let Ω be a bounded domain in $\mathbb{R}^2$. We denote $\Omega_h := \Omega \times (-h/2, h/2)$.

We assume that n_0 is a $\mathcal{C}^2$ function, and that $(1 - n_0/h)$ is supported in Ω_d for some $d > 0$, where $\Omega_d := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > d\}$, $\Re n_0(x) \geq C > 0$, and $\Im n_0(x) \geq 0$.

The problem we wish to solve is to determine the total field E given the incident wave E^i and a scatterer $n(x,z)$. We will seek an approximate solution for this problem by exploiting the fact that h will be small. The approach we take is to write the field E as a perturbation series in h.

3. Derivation of the asymptotic expansion. For $x, x' \in \mathbb{R}^2$ and $z, z' \in \mathbb{R}$, let the fundamental solution Φ be defined by

$$\Phi(x,z,x',z') := \frac{1}{4\pi} \frac{e^{ik|(x,z)-(x',z')|}}{|(x,z)-(x',z')|}.$$

According to [3], we can rewrite the problem as the following Lippmann–Schwinger equation:

(3.1)
$$\begin{aligned} E(x,z) &= E^i(x,z) - k^2 \int_\Omega \int_{-h/2}^{h/2} \Phi(x,z,x',z') \left(1 - \frac{n_0(x')}{h}\right) E(x',z') dx' dz' \\ &\quad - \nabla_{x,z} \int_\Omega \int_{-h/2}^{h/2} \Phi(x,z,x',z') \frac{\nabla n_0(x')}{n_0(x')} \cdot E(x',z') dx' dz'. \end{aligned}$$

Consequently, it suffices to construct an approximation of the electric field E inside Ω_h and insert it into the right-hand side of (3.1) to obtain an expansion of E for all points outside the thin structure Ω_h.

After scaling $z = h\zeta$ and $z' = h\zeta'$, (3.1) can be written as

(3.2)
$$\begin{aligned} E(x,h\zeta) &= E^i(x,h\zeta) - k^2 \int_\Omega \int_{-1/2}^{1/2} \Phi(x,h\zeta,x',h\zeta')(h - n_0(x')) E(x',h\zeta') dx' d\zeta' \\ &\quad - \begin{pmatrix} h\nabla_x \\ \nabla_\zeta \end{pmatrix} \int_\Omega \int_{-1/2}^{1/2} \Phi(x,h\zeta,x',h\zeta') \frac{\nabla n_0(x')}{n_0(x')} \cdot E(x',h\zeta') dx' d\zeta'. \end{aligned}$$

For a scalar function f defined in Ω_h, define the integral operators K_1 and K_2 by

$$\begin{aligned} K_1(f)(x,\zeta) &:= \int_\Omega \int_{-1/2}^{1/2} \Phi(x,h\zeta,x',h\zeta') f(x',h\zeta') dx' d\zeta', \\ K_2(f)(x,\zeta) &:= \begin{pmatrix} h\nabla_x \\ \nabla_\zeta \end{pmatrix} \int_\Omega \int_{-1/2}^{1/2} \Phi(x,h\zeta,x',h\zeta') f(x',h\zeta') dx' d\zeta'. \end{aligned}$$

Then the integral representation (3.2) can be rewritten as

(3.3)
$$E(x,h\zeta) = E^i(x,h\zeta) + k^2 K_1(n_0 E)(x,\zeta) - hk^2 K_1(E)(x,\zeta) - K_2(n_0^{-1} \nabla n_0 \cdot E)(x,\zeta).$$

We now investigate the behavior of the integral operators K_1 and K_2 as $h \to 0$. We begin by proving the following lemma.

LEMMA 3.1. *Let $0 < \alpha < 1$. Suppose that f is $\mathcal{C}^{1,\alpha}$ in Ω_h; then*

$$\begin{aligned} K_1(n_0 f)(x,\zeta) &= \int_\Omega \Phi(x,0,x',0) n_0(x') f(x',0) dx' - \frac{h}{2}\left(\zeta^2 + \frac{1}{4}\right) n_0(x) f(x,0) \\ &\quad + O\left(h^{1+\alpha} \|f\|_{\mathcal{C}^{1,\alpha}(\Omega_h)}\right). \end{aligned} \tag{3.4}$$

Suppose that $g \in L^\infty(\Omega_h)$; then

$$K_1(g)(x,\zeta) = \int_\Omega \Phi(x,0,x',0) g(x',0) dx' + O(h \|g\|_{L^\infty(\Omega_h)}). \tag{3.5}$$

Proof. Observe that

$$\begin{aligned} &\Phi(x,h\zeta,x',h\zeta') f(x',h\zeta') - \Phi(x,0,x',0) f(x',0) \\ &= \Phi(x,h\zeta,x',h\zeta') \left(f(x',0) + h\zeta' \frac{\partial f}{\partial z}(x',0) + O(h^{1+\alpha}) \right) - \Phi(x,0,x',0) f(x',0) \\ &= \left[\Phi(x,h\zeta,x',h\zeta') - \Phi(x,0,x',0) \right] f(x',0) + \Phi(x,h\zeta,x',h\zeta') \left(h\zeta' \frac{\partial f}{\partial z}(x',0) + O(h^{1+\alpha}) \right). \end{aligned}$$

Let $S := \Omega \times (-1/2, 1/2)$. We then have

$$\begin{aligned}
&\int_S \Big[\Phi(x,h\zeta,x',h\zeta')n_0(x')f(x',h\zeta') - \Phi(x,0,x',0)n_0(x')f(x',0)\Big]dx'd\zeta' \\
&= \int_S \Big[\Phi(x,h\zeta,x',h\zeta') - \Phi(x,0,x',0)\Big]n_0(x')f(x',0)dx'd\zeta' \\
&\quad + \int_S \Phi(x,h\zeta,x',h\zeta')n_0(x')h\zeta'\frac{\partial f}{\partial z}(x',0)dx'd\zeta' + O(h^{1+\alpha})\int_S \Phi(x,h\zeta,x',h\zeta')n_0(x')dx'd\zeta' \\
&:= I_1 + I_2 + I_3.
\end{aligned}$$

Since

$$\left|\int_S \Phi(x,h\zeta,x',h\zeta')n_0(x')dx'd\zeta'\right| \le C$$

regardless of h, one can see that $I_3 = O(h^{1+\alpha})$.

To estimate I_2, we first observe that

$$\int_{-1/2}^{1/2} \zeta'\frac{\partial f}{\partial z}(x',0)d\zeta' = 0,$$

and hence

$$I_2 = h\int_S \Big[\Phi(x,h\zeta,x',h\zeta') - \Phi(x,0,x',0)\Big]n_0(x')\zeta'\frac{\partial f}{\partial z}(x',0)dx'd\zeta'.$$

Therefore, we get

$$|I_2| \le \|f\|_{C^1(\Omega_h)}h\int_S \Big|\Phi(x,h\zeta,x',h\zeta') - \Phi(x,0,x',0)\Big|dx'd\zeta'.$$

For a given $(x,\zeta) \in S$ and $h > 0$, let $S_h := \{(x',\zeta') \mid |x-x'| > 2h|\zeta-\zeta'|\}$. If $(x',\zeta') \in S_h$, then

$$(|x-x'|^2 + h^2|\zeta-\zeta'|^2)^{1/2} = |x-x'|\left(1 + O\left(\frac{h^2|\zeta-\zeta'|^2}{|x-x'|^2}\right)\right),$$

and hence

$$\begin{aligned}
\Phi(x,h\zeta,x',h\zeta') &= \frac{1}{4\pi}\frac{e^{ik(|x-x'|^2+h^2|\zeta-\zeta'|^2)^{1/2}}}{(|x-x'|^2+h^2|\zeta-\zeta'|^2)^{1/2}} \\
&= \frac{1}{4\pi}e^{ik|x-x'|}\left(1 + O\left(\frac{h^2|\zeta-\zeta'|^2}{|x-x'|}\right)\right)\left(\frac{1}{|x-x'|} + O\left(\frac{h^2|\zeta-\zeta'|^2}{|x-x'|^3}\right)\right).
\end{aligned}$$

Thus we get

(3.6)
$$\Big|\Phi(x,h\zeta,x',h\zeta') - \Phi(x,0,x',0)\Big| \le C\frac{h^2|\zeta-\zeta'|^2}{|x-x'|^2} + C\frac{h^2|\zeta-\zeta'|^2}{|x-x'|^3} + C\frac{h^4|\zeta-\zeta'|^4}{|x-x'|^4}.$$

One can easily show that

$$\int_{-1/2}^{1/2}\int_{|x-x'|>2h|\zeta-\zeta'|} \frac{h^2|\zeta-\zeta'|^2}{|x-x'|^2}dx'dz' \le Ch^2,$$

$$\int_{-1/2}^{1/2}\int_{|x-x'|>2h|\zeta-\zeta'|}\frac{h^2|\zeta-\zeta'|^2}{|x-x'|^3}dx'dz' \le Ch,$$

and

$$\int_{-1/2}^{1/2}\int_{|x-x'|>2h|\zeta-\zeta'|}\frac{h^4|\zeta-\zeta'|^4}{|x-x'|^4}dx'dz' \le Ch^2.$$

Thus we get

$$\int_{S_h}\Big|\Phi(x,h\zeta,x',h\zeta')-\Phi(x,0,x',0)\Big|dx'd\zeta' \le Ch.$$

On the other hand, one can also easily see that

$$\begin{aligned}&\int_{S\setminus S_h}\Big|\Phi(x,h\zeta,x',h\zeta')-\Phi(x,0,x',0)\Big|dx'd\zeta'\\ &\le\int_{-1/2}^{1/2}\int_{|x-x'|\le 2h|\zeta-\zeta'|}\frac{1}{|x-x'|+h|\zeta-\zeta'|}+\frac{1}{|x-x'|}dx'd\zeta' \le Ch,\end{aligned}$$

and hence we obtain

$$|I_2| \le Ch^2.$$

We now estimate I_1. Since $n_0(x')-h$ is compactly supported in Ω, we can write

$$\begin{aligned}I_1 =& \int_{-1/2}^{1/2}\int_{B_R(x)}\Big[\Phi(x,h\zeta,x',h\zeta')-\Phi(x,0,x',0)\Big](n_0(x')-h)f(x',0)dx'd\zeta'\\ &+h\int_S\Big[\Phi(x,h\zeta,x',h\zeta')-\Phi(x,0,x',0)\Big]f(x',0)dx'd\zeta',\end{aligned}$$

where $B_R(x)$ is a disk of radius R centered at x containing Ω. One can prove that the second term in the right-hand side of above identity is $O(h^2)$ as before, while the first term equals

$$\begin{aligned}&\int_{B_R(x)}\big[\Phi(x,h\zeta,x',h\zeta')-\Phi(x,0,x',0)\big]\big[(n_0(x')-h)f(x',0)-(n_0(x)-h)f(x,0)\big]dx'd\zeta'\\ &+(n_0(x)-h)f(x,0)\int_{B_R(x)}\big[\Phi(x,h\zeta,x',h\zeta')-\Phi(x,0,x',0)\big]dx'd\zeta'.\end{aligned}$$

Since $|(n(x')-h)f(x',0)-(n(x)-h)f(x,0)| \le C|x-x'|$, one can show that the first term of the above is $O(h^2)$ in a similar way to the estimate of I_2. On the other hand, it is proved in [5] that

$$\int_{B_R(x)}\big[\Phi(x,h\zeta,x',h\zeta')-\Phi(x,0,x',0)\big]dx'd\zeta' = -\frac{h}{2}\left(\zeta^2+\frac{1}{4}\right)+O(h^2) \quad \text{as } h\to 0.$$

Thus we have

$$I_1 = -\frac{h}{2}\left(\zeta^2+\frac{1}{4}\right)n_0(x)f(x,0)+O(h^2).$$

Combining all these estimates we obtain (3.4).

Estimate (3.5) can be proved in a similar but simpler way, so we omit the proof. □

We now derive the leading-order term in the asymptotic expansion of the integral operator K_2.

LEMMA 3.2. *Let $0 < \alpha < 1$. Define $\widetilde{K}_2^1(f)(x)$ by*

(3.7)
$$\widetilde{K}_2^1(f)(x) := \int_\Omega \nabla_x \Phi(x,0,x',0)[f(x',0) - f(x,0)]dx' + f(x,0)\int_\Omega \nabla_x \Phi(x,0,x',0)dx',$$

and $\widetilde{K}_2(f)(x,\zeta)$ by

$$\widetilde{K}_2(f)(x,\zeta) := \begin{pmatrix} \widetilde{K}_2^1(f)(x) \\ (\frac{\zeta}{2} - 1)f(x,0) \end{pmatrix}. \tag{3.8}$$

Then, for any $\mathcal{C}^{1,\alpha}$-function f such that $f(\cdot,z)$ is supported in Ω_d for each z,

$$K_2(f)(x,\zeta) = h\widetilde{K}_2(f)(x,\zeta) + O(h^{1+\alpha}), \tag{3.9}$$

where $O(h^{1+\alpha})$ is bounded by $Ch^{1+\alpha}\|f\|_{\mathcal{C}^{1,\alpha}(\Omega_h)}$.

Proof. Note that

$$\begin{aligned}
&\nabla_x \int_S \Phi(x,h\zeta,x',h\zeta')f(x',h\zeta')dx'd\zeta' \\
&= \int_S \nabla_x \Phi(x,h\zeta,x',h\zeta')\Big[f(x',h\zeta') - f(x,h\zeta')\Big]dx'd\zeta' \\
&\qquad + \int_S \nabla_x \Phi(x,h\zeta,x',h\zeta')f(x,h\zeta')dx'd\zeta'.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\nabla_x \int_S \Phi(x,h\zeta,x',h\zeta')f(x',h\zeta')dx'd\zeta' - \widetilde{K}_2^1(f(\cdot,0))(x) \\
&= \int_S \Big[\nabla_x \Phi(x,h\zeta,x',h\zeta') - \nabla_x \Phi(x,0,x',0)\Big]\Big[f(x',h\zeta') - f(x,h\zeta')\Big]dx'd\zeta' \\
&\quad + \int_S \nabla_x \Phi(x,0,x',0)\Big[f(x',h\zeta') - f(x,h\zeta') - f(x',0) + f(x,0)\Big]dx'd\zeta' \\
&\quad + \int_S \Big[\nabla_x \Phi(x,h\zeta,x',h\zeta') - \nabla_x \Phi(x,0,x',0)\Big]f(x,h\zeta')dx'd\zeta' \\
&\quad + \int_S \nabla_x \Phi(x,0,x',0)\Big[f(x,h\zeta') - f(x,0)\Big]dx'd\zeta' \\
&:= I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Since

$$\begin{aligned}
&\nabla_x \Phi(x,h\zeta,x',h\zeta') \\
&= \frac{1}{4\pi} e^{ik(|x-x'|^2 + h^2|\zeta-\zeta'|^2)^{1/2}}\left[ik\frac{x-x'}{|x-x'|^2 + h^2|\zeta-\zeta'|^2} - \frac{x-x'}{(|x-x'|^2 + h^2|\zeta-\zeta'|^2)^{3/2}}\right],
\end{aligned}$$

we get for $|x-x'| \geq 2h|\zeta-\zeta'|$

$$\begin{aligned}
&\nabla_x \Phi(x, h\zeta, x', h\zeta') \\
&= \frac{ik}{4\pi} e^{ik|x-x'|}\left[1 + O\left(\frac{h^2|\zeta-\zeta'|^2}{|x-x'|}\right)\right]\left[ik\frac{x-x'}{|x-x'|^2} - \frac{x-x'}{|x-x'|^3} + O\left(\frac{h^2|\zeta-\zeta'|^2}{|x-x'|^4}\right)\right] \\
&= \nabla_x \Phi(x, 0, x', 0) + O\left(\frac{h^2|\zeta-\zeta'|^2}{|x-x'|^4}\right).
\end{aligned}$$

It then follows that

$$\begin{aligned}
&\int_{S_h}\left[\nabla_x \Phi(x, h\zeta, x', h\zeta') - \nabla_x \Phi(x, 0, x', 0)\right]\Big[f(x', h\zeta') - f(x, h\zeta')\Big]dx'd\zeta' \\
&\leq C\|f\|_{\mathcal{C}^1(\Omega_h)} \int_{-1/2}^{1/2}\int_{|x-x'|\geq 2h|\zeta-\zeta'|} \frac{h^2|\zeta-\zeta'|^2}{|x-x'|^4}|x-x'|dx'd\zeta' \\
&\leq Ch.
\end{aligned}$$

On the other hand, we get

$$\begin{aligned}
&\int_{S\setminus S_h}\left[\nabla_x \Phi(x, h\zeta, x', h\zeta') - \nabla_x \Phi(x, 0, x', 0)\right]\Big[f(x', h\zeta') - f(x, h\zeta')\Big]dx'd\zeta' \\
&\leq C\|f\|_{\mathcal{C}^1(\Omega_h)} \int_{-1/2}^{1/2}\int_{|x-x'|\geq 2h|\zeta-\zeta'|} \frac{1}{|x-x'|^2}|x-x'|dx'd\zeta' \\
&\leq Ch.
\end{aligned}$$

Therefore, we obtain

$$|I_1| \leq Ch.$$

Observe that

$$\begin{aligned}
\Big|f(x', h\zeta') - f(x, h\zeta') - f(x', 0) + f(x, 0)\Big| &= \left|\int_0^{h\zeta'}\left[\frac{\partial}{\partial z}f(x', z) - \frac{\partial}{\partial z}f(x, z)\right]\right| \\
&\leq C\|f\|_{\mathcal{C}^{1,\alpha}(\Omega_h)} h|x-x'|^\alpha.
\end{aligned}$$

It then follows that

$$|I_2| \leq C\|f\|_{\mathcal{C}^{1,\alpha}(\Omega_h)} h\int_{-1/2}^{1/2}\int_\Omega \frac{1}{|x-x'|^2}|x-x'|^\alpha dx'd\zeta' \leq C\|f\|_{\mathcal{C}^{1,\alpha}(\Omega_h)} h.$$

If $x \in \operatorname{supp}(f)$, then $B_d(x) \subset \Omega$. Moreover,

$$\int_{B_d(x)} \nabla_x \Phi(x, h\zeta, x', h\zeta')dx' = \int_{B_d(x)} \nabla_x \Phi(x, 0, x', 0)dx' = 0 \quad \text{for all } \zeta \text{ and } \zeta'.$$

Therefore we get

$$\begin{aligned}
|I_3| &= \int_{-1/2}^{1/2}\int_{\Omega\setminus B_d(x)}\Big[\nabla_x \Phi(x, h\zeta, x', h\zeta') - \nabla_x \Phi(x, 0, x', 0)\Big]f(x, h\zeta')dx'd\zeta' \\
&\leq C\|f\|_{L^\infty(\Omega_h)} h^2 \int_{-1/2}^{1/2}|\zeta-\zeta'|^2 d\zeta' \leq C\|f\|_{L^\infty(\Omega_h)} h^2.
\end{aligned}$$

Finally, we arrive at

$$|I_4| = C\|f\|_{\mathcal{C}^1(\Omega_h)} h \int_{\Omega\setminus B_d(x)} \frac{1}{|x-x'|^2} dx' \le C\|f\|_{\mathcal{C}^1(\Omega_h)} h.$$

So far we proved that

$$\nabla_x \int_S \Phi(x, h\zeta, x', h\zeta') f(x', h\zeta') dx' d\zeta' = \widetilde{K}_2^1(f(\cdot,0))(x) + O(h^\alpha \|f\|_{\mathcal{C}^{1,\alpha}(\Omega_h)}) \quad \text{as } h \to 0.$$

We now investigate the behavior of

$$\frac{\partial}{\partial z} \int_S \Phi(x, h\zeta, x', h\zeta') f(x', h\zeta') dx' d\zeta'$$

as $h \to 0$.

Elementary computations show that

$$\begin{aligned}
&\frac{\partial}{\partial z}\Phi(x, h\zeta, x', h\zeta') \\
&= \frac{1}{4\pi} e^{ik(|x-x'|^2+h^2|\zeta-\zeta'|^2)^{1/2}} \left[ik \frac{h^2(\zeta-\zeta')}{|x-x'|^2+h^2|\zeta-\zeta'|^2} - \frac{h^2(\zeta-\zeta')}{(|x-x'|^2+h^2|\zeta-\zeta'|^2)^{3/2}} \right] \\
&= -\frac{1}{4\pi} \frac{h^2(\zeta-\zeta')}{(|x-x'|^2+h^2|\zeta-\zeta'|^2)^{3/2}} + O\left(\frac{h^2|\zeta-\zeta'|}{|x-x'|^2+h^2|\zeta-\zeta'|^2} \right).
\end{aligned}$$

Since $-\frac{1}{4\pi}\frac{t}{(|x-x'|^2+t^2)^{3/2}}$ is the Poisson kernel for the half space and $f(\cdot, h\zeta')$ has a compact support in Ω, we get

$$-\frac{1}{4\pi} \int_\Omega \frac{h^2(\zeta-\zeta')}{(|x-x'|^2+h^2|\zeta-\zeta'|^2)^{3/2}} f(x', h\zeta') dx' = -\frac{h}{2} f(x,0) \frac{\zeta-\zeta'}{|\zeta-\zeta'|} + O(h^2),$$

where the $O(h^2)$-term is bounded by $h^2\|f\|_{\mathcal{C}^1(\Omega_h)}$. On the other hand,

$$\int_\Omega \frac{h^2|\zeta-\zeta'|}{|x-x'|^2+h^2|\zeta-\zeta'|^2} |f(x', h\zeta')| dx' \le C\|f\|_{L^\infty(\Omega_h)} h^2|\zeta-\zeta'|\,|\ln(h|\zeta-\zeta'|)|.$$

It then follows that

$$\begin{aligned}
&\frac{\partial}{\partial z} \int_S \Phi(x, h\zeta, x', h\zeta') f(x', h\zeta') dx' d\zeta' \\
&= -\frac{h}{2} f(x,0) \int_{-1/2}^{1/2} \frac{\zeta-\zeta'}{|\zeta-\zeta'|} d\zeta' + O(h^2|\ln h|) \\
&= h\left(\frac{\zeta}{2} - 1\right) f(x,0) + O(h^2|\ln h|).
\end{aligned}$$

This completes the proof. □

REMARK 3.3. *It is worth noticing that the above lemma applies if f is $\mathcal{C}^{1,\alpha}$ in the z variable and only $\mathcal{C}^1$ in the x variable.*

The following lemma is also of use to us.

LEMMA 3.4. *The operator*

$$T : f \in \mathcal{C}^0(\overline{\Omega}) \mapsto f(x) - k^2 \int_\Omega \Phi(x, 0, x', 0) n_0(x') f(x')\, dx' \in \mathcal{C}^0(\overline{\Omega})$$

is invertible.

Proof. As proved in [5], the operator

$$f \in L^2(\Omega) \mapsto \int_\Omega \Phi(x,0,x',0) n_0(x') f(x')\, dx' \in L^2(\Omega)$$

is compact. Since $n_0(x)$ is compactly supported in Ω, we can actually prove that

$$f \in \mathcal{C}^0(\overline{\Omega}) \mapsto \int_\Omega \Phi(x,0,x',0) n_0(x') f(x')\, dx' \in \mathcal{C}^0(\overline{\Omega})$$

is compact. Thus, by the Fredholm alternative, it suffices to show that T is injective on $\mathcal{C}^0(\overline{\Omega})$. Let $f \in \mathcal{C}^0(\overline{\Omega})$ be a solution to

$$f(x) - k^2 \int_\Omega \Phi(x,0,x',0) n_0(x') f(x')\, dx' = 0.$$

Define the function $u(x,z)$ in $\mathbb{R}^3 \setminus \overline{\Omega}$ by

$$u(x,z) = \int_\Omega \Phi(x,z,x',0) n_0(x') f(x')\, dx'.$$

It follows that

$$\begin{cases} (\Delta + k^2) u(x) = 0 \quad \text{in } \mathbb{R}^3 \setminus \overline{\Omega}, \\ u|_+ - u|_- = 0 \quad \text{on } \Omega, \\ \dfrac{\partial u}{\partial z}\Big|_+ (x) - \dfrac{\partial u}{\partial z}\Big|_- (x) = n_0(x) f(x) \quad \text{on } \Omega, \\ u \text{ satisfies the Sommerfeld radiation condition.} \end{cases}$$

Let B_R be a sphere of center the origin and of radius R which is such that $\Omega \subset B_R$. Integrating by parts we obtain that

$$\int_{B_R \setminus \overline{\Omega}} |\nabla u|^2 - k^2 \int_{B_R} |u|^2 - \int_{\partial B_R} \frac{\partial u}{\partial \nu} \overline{u} - \int_\Omega \left[\frac{\partial u}{\partial z}\right] \overline{u} = 0,$$

where $[\partial u/\partial z] = \partial u/\partial z|_+ - \partial u/\partial z|_-$ denotes the jump of $\partial u/\partial z$ across Ω. Since $u = (1/k^2) f$ and $[\partial u/\partial z] = n_0 f$ on Ω we get

$$\int_\Omega \left[\frac{\partial u}{\partial z}\right] \overline{u} = \frac{1}{k^2} \int_\Omega n_0(x) |f(x)|^2\, dx.$$

Thus,

$$\Im \int_{\partial B_R} \frac{\partial u}{\partial \nu} \overline{u} = -\Im \frac{1}{k^2} \int_\Omega n_0(x) |f(x)|^2\, dx \le 0,$$

since $\Im n_0(x) \ge 0$. By the Rellich lemma (see [3]), the above inequality implies that $u \equiv 0$ in $\mathbb{R}^3 \setminus \overline{\Omega}$, and hence $n_0 f = [\partial u/\partial z] = 0$ which gives $f = 0$ on Ω since $\Re n_0(x) \ge C > 0$. ∎

We finally establish the following bound on the $\mathcal{C}^{1,\alpha}$-norm of E.

PROPOSITION 3.5. *Let E be the solution of the Lippmann–Schwinger equation (3.1). Then for any $0 < \alpha < 1$, there exists a constant C independent of h such that*

$$\|E\|_{\mathcal{C}^{1,\alpha}(\overline{\Omega_h})} \leq C|\log h|^2. \tag{3.10}$$

We will give a proof of Proposition 3.5 in Appendix A.

At this point we have all the necessary ingredients to state and prove our main result in this paper. Let

$$\widetilde{K}_1(f)(x) = \int_\Omega \Phi(x,0,x',0)f(x')\,dx', \quad x \in \Omega, \tag{3.11}$$

and $\widetilde{K}_2(f)$ be as in (3.8). Equation (3.3), together with Lemmas 3.1 and 3.2, yields the pointwise approximation

(3.12)

$$\begin{aligned} E(x,h\zeta) &= E^i(x,h\zeta) + k^2\widetilde{K}_1(n_0E_0)(x,0) \\ &\quad - h\left[k^2\widetilde{K}_1(E_0)(x) + \frac{k^2}{2}\left(\zeta^2+\frac{1}{4}\right)n_0(x)E_0(x) + \widetilde{K}_2(n_0^{-1}\nabla n_0\cdot E_0)(x,\zeta)\right] \\ &\quad + O(h^{1+\alpha}), \end{aligned}$$

where $E_0(x) = E(x,0)$ and $O(h^{1+\alpha}) \leq Ch^{1+\alpha}||E||_{\mathcal{C}^{1,\alpha}(\Omega_h)}$ for some positive constant C. Using Lemma 3.4, we define $E^{(0)}$ and $E^{(1)}$ as the unique solutions to the integral equations posed on Ω:

$$T(E^{(0)}) := E^{(0)} - k^2\widetilde{K}_1(n_0E^{(0)})(x) = E^i(x,0)$$

and

$$\begin{aligned} T(E^{(1)}) &= -\left[k^2\widetilde{K}_1(E^{(0)})(x) + \frac{k^2}{2}\left(\zeta^2+\frac{1}{4}\right)n_0(x)E^{(0)}(x) + \widetilde{K}_2(n_0^{-1}\nabla n_0\cdot E^{(0)})(x,\zeta)\right] \\ &\quad + \zeta\partial_z E^i(x,0). \end{aligned}$$

REMARK 3.6. *Note that $E^{(0)}$ is independent of ζ while $E^{(1)}(x,\zeta)$ is a second order polynomial in the stretched variable ζ.*

Let

(3.13)

$$\begin{aligned} \widetilde{E}(x,h\zeta) &= E^i(x,h\zeta) + k^2\widetilde{K}_1(n_0(E^{(0)}+hE^{(1)})(x,0) \\ &\quad - h\left[k^2\widetilde{K}_1(E^{(0)})(x) + \frac{k^2}{2}\left(\zeta^2+\frac{1}{4}\right)n_0(x)E^{(0)}(x) + \widetilde{K}_2(n_0^{-1}\nabla n_0\cdot E^{(0)})(x,\zeta)\right], \end{aligned}$$

for $x \in \Omega$ and $\zeta \in (-1/2, 1/2)$. The following error estimate is our main result in this paper.

THEOREM 3.7. *Let $\widetilde{E}$ be defined by (3.13) in $\Omega\times(-1/2,1/2)$. For any $0 < \nu < 1$ we have*

$$||E(x,h\zeta) - \widetilde{E}(x,h\zeta)||_{\mathcal{C}^0(\overline{\Omega}\times[-1/2,1/2])} \leq Ch^{1+\nu},$$

for some positive constant C independent of h and E.

Proof. Let $0 < \nu < 1$. From (3.12) we immediately obtain that

$$T\Big(E_0 - (E^{(0)} + hE^{(1)})\Big) = O(h^{1+\alpha}),$$

where $\alpha = (1+\nu)/2$, which shows that

$$||E_0 - (E^{(0)} + hE^{(1)})||_{\mathcal{C}^0(\overline{\Omega})} \leq Ch^{1+\alpha}||E||_{\mathcal{C}^{1,\alpha}(\Omega_h)}.$$

Inserting this approximation of E_0 into the right-hand side of (3.12) and using the fact from Lemma 3.4 that T has a bounded inverse together with the definition (3.13) of $\widetilde{E}$, we arrive at the desired estimate since, by Proposition 3.5, the quantity $h^{\frac{1-\alpha}{2}}\,||E||_{\mathcal{C}^{1,\alpha}(\Omega_h)}$ is uniformly bounded in h. ∎

To conclude the paper we make the following remarks.

REMARK 3.8. *To obtain an approximate solution to the scattering problem, we find $E^{(0)}$ and $E^{(1)}$ by solving integral equations in $\Omega \subset\subset \mathbb{R}^2$. The electric field within the thin structure is easily obtained using* (3.13). *Should we wish to evaluate the field outside the scatterer, we can use the Lippman–Schwinger equation. Thus, the analysis described in this work can be viewed as a computational method. The method is efficient since only* 2*-D integral equations need to be solved to obtain the desired approximate solution. We note that in* [5] *numerical calculations were carried out to assess the accuracy of the method for the scalar problem in two dimensions. See also* [6].

REMARK 3.9. *If n_0 is constant, then the leading-order term $E^{(0)}$ satisfies the integral equation*

$$E^{(0)}(x,0) - k^2 n_0 \int_\Omega \Phi(x,0,x',0)E^{(0)}(x',0)\,dx' = E^i(x,0), \quad x \in \Omega,$$

and therefore, the tangential component of $E^{(0)}$ on $\Omega \times \{0\}$ is continuous while $e_z \times (\nabla \times E^{(0)})$ on $\Omega \times \{0\}$ has a jump given by $-k^2 n_0 e_z \times (e_z \times E^{(0)})$, where e_z is the basis vector in the z-direction. These jump conditions are exactly those derived by Bouchitté in [1].

REMARK 3.10. *We should emphasize the fact that the asymptotic expansion derived in this paper is not valid in the case of piecewise constant media such as holes in the slab. The regularity assumption on n_0 is essential to first write the Lippmann–Schwinger equation on the electric field E and then to establish the asymptotic expansion* (3.12) *using Lemmas* 3.1 *and* 3.2. *In fact, the estimate of the remainder in* (3.12) *may not be valid if the regularity assumption on n_0 does not hold. This important case in practice requires further delicate analysis and will be the subject of a forthcoming work.*

REMARK 3.11. *We believe that the approach presented in this paper can be easily generalized to deal with a dielectric layer having a curved mean surface.*

Appendix A. Proof of Proposition 3.5. Here we prove Proposition 3.5. Define

$$Tf(x,z) := \int_\Omega \int_{-h/2}^{h/2} \Phi(x,z,x',z')f(x',z')dx'dz', \quad (x,z) \in \Omega_h.$$

We show the following regularity properties of the operator T.

LEMMA A.1. *Let $m(x)$ be a $\mathcal{C}^1$-function compactly supported in Ω. For any $\alpha \in (0,1)$, there exists a constant C independent of h and f such that*

(i) $\|T(mf)\|_{\mathcal{C}^{2,\alpha}(\overline{\Omega_h})} \le C\|m\|_{\mathcal{C}^1(\Omega)}\|f\|_{\mathcal{C}^\alpha(\overline{\Omega_h})}$,
(ii) $\|T(mf)\|_{\mathcal{C}^1(\overline{\Omega_h})} \le Ch|\log h|\,\|m\|_{L^\infty(\Omega)}\|f\|_{L^\infty(\Omega_h)}$,
(iii) $\|T(mf)\|_{\mathcal{C}^2(\overline{\Omega_h})} \le Ch|\log h|\,\|m\|_{\mathcal{C}^1(\Omega)}\|f\|_{\mathcal{C}^1(\overline{\Omega_h})}$.

Before proving Lemma A.1, let us first show that Proposition 3.5 follows from Lemma A.1. Equation (3.1) can be written as

$$E = E^i - k^2 T\left(\left(1 - \frac{n_0}{h}\right)E\right) - \nabla_{x,z} T\left(\frac{\nabla n_0}{n_0}\cdot E\right). \tag{A.1}$$

Since the solution E to (A.1) is uniformly bounded in $L^\infty(\Omega_h)$, it follows from Lemma A.1(ii) that $\|E\|_{\mathcal{C}^1(\overline{\Omega_h})} \le C|\log h|$, and hence from (A.1)(i), (iii) we obtain that $\|E\|_{\mathcal{C}^{1,\alpha}(\overline{\Omega_h})} \le C|\log h|^2$ for some constant C independent of h.

Proof of Lemma A.1. We first note that

$$|D^s\Phi(x,z,x',z')| \le \frac{C}{|(x,z)-(x',z')|^{1+s}}, \tag{A.2}$$

for $s = 0, 1, \ldots$, where D^s denotes any sth order derivative.

Let $(x,z) = y$, $(x',z') = y'$, and D_{ij} denote the second order partial derivative with respect to y_i and y_j. Then it is known that for $i,j = 1,2,3$,

$$\begin{aligned} D_{ij}T(mf)(y) &= \int_{\Omega_h} D_{ij}\Phi(y-y')m(x')(f(y')-f(y))dy \\ &\quad - f(y)\int_{\partial\Omega_h} D_i\Phi(y-y')m(x')\nu_j(y')d\sigma(y') \\ &:= I_1(y) - f(y)I_2(y), \end{aligned} \tag{A.3}$$

where $\nu = (\nu_1,\nu_2,\nu_3)$ is the outward unit normal to $\partial\Omega_h$. See Lemma 4.2 in [4, p. 55]. By a standard Hölder estimate (see, for example, the proof of Lemma 4.4 in [4, p. 57]), we can show that

$$|I_1(y) - I_1(\bar y)| \le C\|m\|_{L^\infty(\Omega)}\|f\|_{\mathcal{C}^\alpha(\overline{\Omega_h})}|y-\bar y|^\alpha, \quad y,\bar y \in \Omega_h.$$

Observe that $\partial\Omega_h$ consists of two parts: $\Omega\times\{h/2,-h/2\}$ and $\partial\Omega\times[-h/2,h/2]$. Since m has a compact support in Ω, we have

$$I_2(y) = \int_{\Omega\times\{h/2,-h/2\}} D_i\Phi(y-y')m(x')\nu_j(y')dx'.$$

Moreover, $\nu_j = 0$ on $\partial\Omega\times\{h/2,-h/2\}$ if $j \ne 3$, and then we have only to consider the quantity I_2 when $j = 3$. If $j = 3$, then

$$I_2(y) = \int_\Omega D_i\Phi(x,z,x',h/2)m(x')dx' - \int_\Omega D_i\Phi(x,z,x',-h/2)m(x')dx'.$$

Suppose that $i = 1$ or 2. Then, since m has a compact support in Ω and $D_i\Phi(x,z,x',z') = -D_i'\Phi(x,z,x',z')$ where D' denotes the derivative with respect to y' variables, we have

$$I_2(y) = \int_\Omega \Phi(x,z,x',h/2)D_i m(x')dx' - \int_\Omega \Phi(x,z,x',-h/2)D_i m(x')dx'. \tag{A.4}$$

Thus, we have

$$|I_2(y) - I_2(\bar{y})| \le C\|m\|_{\mathcal{C}^1(\Omega)}|y - \bar{y}|^\alpha, \quad y, \bar{y} \in \Omega_h,$$

for any $0 < \alpha < 1$.

Suppose now that $i = 3$. In this case

$$I_2(y) = -\int_{\partial\Omega} \frac{\partial}{\partial z'}\Phi\left(x, z, x', \frac{h}{2}\right) m(x')dx' + \int_{\partial\Omega} \frac{\partial}{\partial z'}\Phi\left(x, z, x', -\frac{h}{2}\right) m(x')dx'.$$

Using once again the fact that m has a compact support and $\Omega_h = \Omega \times (-h/2, h/2)$, we can write $I_2(y)$ as

$$I_2(y) = -\int_{\partial\Omega_h} \nabla_{y'}\Phi(y - y') \cdot \nu(y) m(x')d\sigma(y').$$

Since $(\Delta + k^2)\Phi(y - y') = \delta(y - y')$, we obtain from the divergence theorem that

$$I_2(y) = -m(x) + k^2 \int_{\Omega_h} \Phi(y - y')m(x')dy' - \int_{\Omega_h} \nabla_{y'}\Phi(y - y') \cdot \nabla m(x')dy'. \tag{A.5}$$

Thus it follows from (A.2) and standard arguments that

$$|I_2(y) - I_2(\bar{y})| \le C\|m\|_{\mathcal{C}^1(\Omega)}|y - \bar{y}|^\alpha,$$

and hence

$$|f(y)I_2(y) - f(\bar{y})I_2(\bar{y})| \le C\|m\|_{\mathcal{C}^1(\Omega)}\|f\|_{\mathcal{C}^\alpha(\overline{\Omega_h})}|y - \bar{y}|^\alpha.$$

This completes the proof of (i).

To prove (ii), we first compute

$$\begin{aligned}
\int_{\Omega_h} |D_j\Phi(x, z, x', z')|dx'dz' &\le C \int_{-h/2}^{h/2} \int_\Omega \frac{1}{|x - x'|^2 + |z - z'|^2} dx'dz' \\
&\le C \int_{-h/2}^{h/2} (1 + |\log|z - z'||)dz' \\
&\le Ch|\log h|,
\end{aligned}$$

for h small enough. Therefore, we get

$$|D_j T(mf)(x, z)| \le Ch|\log h|\|mf\|_{L^\infty(\Omega_h)}.$$

To prove (iii) we use (A.3), (A.4), and (A.5). Then the same estimates lead us to (iii). For example, from (A.4) we get

$$\begin{aligned}
|I_2(y)| &\le C \int_{-h/2}^{h/2} \int_\Omega \left|\frac{\partial}{\partial z'}\Phi(x, z, x', z')\right| |\nabla m(x')|dx'dz' \\
&\le Ch|\log h|\|\nabla m\|_{L^\infty(\Omega)}.
\end{aligned}$$

This completes the proof. ∎

REFERENCES

[1] G. Bouchitté, *Analyse limite de la diffraction d'ondes électromagnétiques par une structure mince*, C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 311 (1990), pp. 51–56.
[2] G. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations*, Springer-Verlag, Berlin, 2002.
[3] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Appl. Math. Sci. 93, Springer-Verlag, New York, 1992.
[4] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.
[5] S. Moskow, F. Santosa, and J. Zhang, *An approximate method for scattering by thin structures*, SIAM J. Appl. Math., 66 (2005), pp. 187–205.
[6] J. H. Richmond, *Scattering by thin dielectric strips*, IEEE Trans. Antennas and Propagation, 33 (1985), pp. 64–68.

SIAM J. MATH. ANAL.
Vol. 38, No. 4, pp. 1343–1370

ON TRANSONIC SHOCKS IN TWO-DIMENSIONAL VARIABLE-AREA DUCTS FOR STEADY EULER SYSTEM*

HAIRONG YUAN†

Abstract. This paper concerns transonic shocks in compressible inviscid flow passing a two-dimensional variable-area duct for the complete steady Euler system. The flow is supersonic at the entrance of the duct, whose boundaries are slightly curved. The condition of impenetrability is posed on the boundaries. After crossing a nearly flat shock front, which passes through a fixed point on the boundary of the duct, the flow becomes subsonic. We show that to ensure the stability of such shocks, pressure should not be completely given at the exit: it only should be given with freedom one, that is, containing an unknown constant to be determined by the upstream flow and the profile of the duct. Careful analysis shows that this is due to the requirement of conservation of mass in the duct. We used Lagrangian transformation and characteristic decomposition to write the Euler system as a 2×2 system, which is valid for general smooth flows. Due to such a simplification, we can employ the theory of boundary value problems for elliptic equations to discuss well-posedness or ill-posedness of transonic shock problems in variable-area duct for various conditions given at the exit.

Key words. Euler system, transonic shocks, free boundary problem, hyperbolic-elliptic composite system, ill-posed problem

AMS subject classifications. 35L65, 35L67, 76N10

DOI. 10.1137/050642447

1. Introduction.

Using nozzles and pipes to transport and control fluid flows has numerous applications. For instance, one of the important phenomena in gas dynamics is that by appropriate design of a nozzle there may generate one or several shock waves to adjust the supersonic gas flow at the entrance of the nozzle to a certain subsonic state at the exit, which is required, for example, for the jet engines of some types of supersonic airplanes to work. (See, for instance, [31, section 7.13] for detailed discussions.) Other examples are various wind tunnels [25]. Since compressible flows in nozzles exhibit abundant phenomena; such as choking, local supersonic bubble, formation of shock waves and their interactions with the boundaries of nozzles, etc. (see [12, Chapter 5] and [24, section 6.2.3]), rigorous and thorough mathematical analysis of flows in nozzles is a formidable task.

Nevertheless, progress has been made for several model problems. For unsteady quasi–one-dimensional gas flow in a duct of variable area (see [27, section 8.1]), Liu showed in a series of papers [22], [23], [17] that supersonic and subsonic flows are stable, and for transonic flows, the shock waves tend to decelerate along an expanding duct and accelerate along a contracting duct. See also [18]. For inviscid isentropic irrotational gas flows, using the full velocity potential equation, Chen and Feldman studied the existence and stability of multidimensional transonic shocks through an infinite nozzle and determined the state of the flows at infinity by the data of the coming flows and the geometry of the nozzle [4], [5], [6]. In [28] Xin and Yin considered a similar problem for finite nozzle with a class of conditions involving potential giving

*Received by the editors October 11, 2005; accepted for publication (in revised form) July 10, 2006; published electronically December 18, 2006. This research was supported in part by NSFC Project 10531020.

http://www.siam.org/journals/sima/38-4/64244.html

†Department of Mathematics, East China Normal University, Shanghai 200062, China (hairongyuan0110@gmail.com, hryuan@math.ecnu.edu.cn).

at the exit. It is also remarkable that Kuz'min [19] studied subsonic-supersonic smooth flows in nozzles by using the Chaplygin equation (which is equivalent to the full velocity potential equation) and the simplified von Kármán equation. See also, for example, [9], [10], [13] for other interesting and important works concerning systems of conservation laws, multidimensional shock waves, and flow patterns in nature.

In this paper we study a class of transonic flows with shocks in a two-dimensional variable-area duct for the steady full Euler system, which is a more precise description of compressible inviscid flows. The flow is supersonic at the entrance of the duct, whose boundaries are suitable perturbations of straight lines, while the flow becomes subsonic across a nearly flat shock front. The condition of impenetrability is posed on the boundaries. We will show that for given pressure at the exit, in general such flow patterns may not exist except when the pressure at the exit satisfies an additional restriction: Its mean value has already been determined by the state of the flow at the entrance and the geometry of the duct. Precisely, to ensure the stability of such transonic shocks, the pressure at the exit can be given only apart from a constant difference, that is, it should contain an unknown constant to be solved simultaneously with the flow fields in the duct. The proof reveals that the requirement of conservation of mass in the duct is closely connected to this phenomenon (see Remark 8.1 in section 8.1).

On the other hand, given pressure at the exit is a physically well accepted condition for flows in nozzles [12]. So our result indicates that the transonic shock we investigated here is unstable and not likely to be observed in practice. However, since the flow fields of the transonic shock we studied here are relatively simple, it may help us gain some insight into understanding those more complicated transonic shocks appearing, for example, in de Laval nozzles. Note that all the works cited above on transonic shocks and [7], [8], [29] are devoted to the study of the class of transonic shocks we investigate here.

In this paper we also discuss, from the mathematical point of view, the well-posedness or ill-posedness of a transonic shock problem in variable-area ducts if other conditions are given at the exit. It is shown that for given density, entropy, Mach number, or the velocity component parallel to the axis of the duct, the problem is in general ill-posed; however, it is well-posed for the given velocity component, which is perpendicular to the axis of the duct. For a list of such results, see section 11.

We remark that in [8] Chen has discussed the special case when the boundaries of the duct are straight lines and the flow has certain symmetric properties, while the upstream supersonic flow is perturbed. The author [29] has also investigated the case for flows in a cylinder with cylindrical symmetry by a different method from [8], and the radial velocity vanishing condition was posed at the exit. In [7] Chen and Yuan developed the methods initiated in [29] and solved the transonic shock problem for the three-dimensional steady full Euler system under the periodic conditions on the lateral boundary of the duct, and hence obtained the solution of the transonic shock problem in a three-dimensional duct with a constant square section under certain assumptions on the symmetry of the coming flow. The ill-posedness for given pressure at the exit is also demonstrated in detail there. However, due to the special structure of the two-dimensional stationary Euler system, we developed a different and more powerful approach here and obtained more results.

Now we comment on several difficulties which lie in the transonic shock problem we investigate presently. First is the treatment of the shock front, which is a free boundary and should be determined with the solutions (the subsonic states of the

gas flows) simultaneously. Fruitful techniques have been developed in [1], [2], [3], [8], [20] to deal with the type of free boundaries we meet here. In a rough way, those techniques allow us, by suitable reformulation of the Rankine–Hugoniot jump conditions, to construct a boundary modifying mapping, whose fixed point is the desired free boundary. The definition of the boundary modifying mapping involves solving a series of nonlinear fixed boundary problems.

Second, the steady full Euler system is a hyperbolic-elliptic composite system for subsonic flow. For such a system, classical techniques such as energy estimates, maximum principle, and estimates of fundamental solutions, are not valid in a straightforward way. One needs to separate the "elliptic part" and "hyperbolic part" appropriately to use the classical theory of elliptic and hyperbolic differential equations. To cope with the curved boundary, we also write the system in Lagrangian coordinates by virtue of the law of conservation of mass, and then decompose it into a 2×2 system (which is elliptic for subsonic flow, hyperbolic for supersonic flow, and of mixed type for transonic flow) and two algebraic equations (one is Bernoulli's law and the other is the invariance of entropy along streamlines for C^1 flows). This simplifies greatly the Euler system and enables us to study the transonic shock problem comprehensively. For example, one of the merits of this approach is that it avoids loss of derivatives. We remark that this formulation may be used to study other smooth flow patterns in ducts. Such a technic has already been used by Chen to study a flat Mach configuration in [11] and by Fang to study the transonic shocks attached to a curved wedge [14].

Third, later on we will find out that we need to solve an elliptic system in a rectangular domain. It is well known that the corners may cause singularities in the solutions (even the well-posedness; see an example in [30]), which in turn affect the regularity of the shock front, and then the smoothness of the boundary itself, and then may cause new trouble in the regularity of solutions. Another feature is that the "hyperbolic part" may transport the singularity at the corners produced by the "elliptic part" to other points in the domain. We are lucky that we can use weighted Hölder spaces and the results established in [16] by Gilbarg, Hörmander and in [21] by Lieberman to overcome this difficulty.

Fourth, as mentioned above, it turns out we are in fact dealing with an ill-posed problem if we give directly the pressure at the exit. We will show its relation to the Neumann boundary problems for Poisson equations and determine appropriate boundary conditions to make such a problem well-posed.

We will use the following well-known Banach contraction mapping principle twice to solve the transonic shock problem:

> *Any contractive mapping on a complete metric space has one and only one fixed point.*

To find the transonic shock front, we will show that the boundary modifying mapping is contractive (see section 10). However, to define the boundary modifying mapping, we need again the Banach contraction mapping principle to show that a series of nonlinear fixed boundary problems are uniquely solvable under some hypothesis (see section 9). Due to our great efforts contributed to simplify the original problem (section 3–7), obtaining the necessary estimates is straightforward and not hard work.

The paper is organized as follows. In section 2 we rigorously formulate the problem of transonic shocks in variable-area ducts (denoted as problem (**A**)) and state our main results, i.e., Theorem 2.6. In section 3 we write the Euler system in Lagrangian coordinates, which also transform the curved boundaries of ducts into straight lines.

In section 4 we decompose the resulted system into the "elliptic part" and "hyperbolic part" and in section 5 we show the existence of supersonic flow in the ducts. In section 6 we formulate a free boundary problem (denoted as problem (**B**)) and then reduce it to a set of fixed boundary problems (denoted as problem (**C**)) and a boundary modifying problem. In section 7, by rewriting the Rankine–Hugoniot jump conditions we express problem (**C**) in an equivalent but more transparent form (denoted as problem (**D**)). Section 8 is devoted to the linearized version of problem (**D**). In section 9 we solve problem (**C**) by the Banach contraction mapping principle. In section 10 we construct the boundary modifying mapping and show that it has a fixed point by using the Banach contraction mapping principle once again, thus finishing the proof of our main results, Theorem 2.6. The last section of this paper, section 11, is devoted to well-posedness or ill-posedness of the transonic shock problem (**A**) for various conditions given at the exits of the ducts. The detailed proofs of these results are omitted since they can be done in the same spirit as the proof of Theorem 2.6, but we have sketched out the main points.

2. Formulation of the transonic shock problem and main results.

2.1. Problem (A) and background solution. The Euler system, which models two-dimensional inviscid steady gas flow, is of the form

$$\begin{cases} \nabla \cdot \mathbf{m} = 0, \\ \nabla \cdot \left(\dfrac{\mathbf{m} \otimes \mathbf{m}}{\rho} \right) + \nabla p = 0, \end{cases} \tag{2.1}$$

with Bernoulli's law

$$\frac{1}{2}\mathbf{u}^2 + i = \text{const}, \tag{2.2}$$

where ρ, p, i are the density, pressure, and enthalpy of the fluid, while $\mathbf{u} = (u, v)$ and $\mathbf{m} = \rho\mathbf{u}$ are the velocity and the momentum density vector, respectively. The first equation in (2.1) is the conservation of mass, the second is the conservation of momentum, and the Bernoulli's law corresponds to the conservation of energy. Note that the const in (2.2) depends on streamlines but is invariant on the same streamline even across a shock [12].

In the case of polytropic gas $p = A(S)\rho^\gamma, \gamma \in (1, \infty)$, with S the entropy, (2.2) takes the form

$$\frac{1}{2}(u^2 + v^2) + \frac{a^2}{\gamma - 1} = \text{const}, \tag{2.3}$$

where $a = \sqrt{\gamma A(S)\rho^{\gamma-1}}$ is the local speed of sound. For C^1 flow, (2.1) can also be written as a symmetric system:

$$\begin{pmatrix} \rho u & 0 & 1 \\ 0 & \rho u & 0 \\ 1 & 0 & \frac{u}{\rho a^2} \end{pmatrix} \partial_x U + \begin{pmatrix} \rho v & 0 & 0 \\ 0 & \rho v & 1 \\ 0 & 1 & \frac{v}{\rho a^2} \end{pmatrix} \partial_y U = 0. \tag{2.4}$$

In this form, the first two equations are the conservation of momentum, and the last is the conservation of mass.

Remark 2.1. In (2.4) we have set

$$U = \begin{pmatrix} u & v & p \end{pmatrix}^t \tag{2.5}$$

as the state of the gas. Since $a^2 = \gamma p/\rho$, we may obtain ρ from (2.3) once the const in it is known. Sometimes we will also set

$$U = \begin{pmatrix} u & v & p & \rho \end{pmatrix}^t. \tag{2.6}$$

Later we will introduce

$$w = \frac{v}{u}. \tag{2.7}$$

To simplify the notation, for u nonzero we will also set U as

$$\begin{pmatrix} u & w & p \end{pmatrix}^t. \tag{2.8}$$

There will be no confusion in using U to express these vectors later.

Without loss of generality, we set $\mathbb{P} := \{(x,y) \in \mathbb{R}^2 : x \in [-1,1], 0 \le y \le \Gamma(x)\}$ to be a two-dimensional duct with variable sections, and denote the upper wall $\{y = \Gamma(x) : x \in [-1,1]\}$ as Γ_+ with $\Gamma(x)$ a positive function, and $\Gamma_- = \{y = 0 : x \in [-1,1]\}$ the lower wall. We also set $\Gamma_s = \{x = s : y \in [0, \Gamma(s)]\}$ for $s \in [-1,1]$.

We are interested in the following boundary value problem (**A**):

$$(\mathbf{A}) : \begin{cases} (2.1), (2.2) & \text{in} \quad \mathbb{P}, \\ v = 0 & \text{on} \quad \Gamma_-, \\ v - u\Gamma'(x) = 0 & \text{on} \quad \Gamma_+, \\ U = U_b^- & \text{on} \quad \Gamma_{-1}, \\ p = p_1 & \text{on} \quad \Gamma_1. \end{cases} \tag{2.9}$$

The conditions on $\Gamma_\pm$ mean that the boundary is impermeable and it is natural for ducts without holes on its boundaries. We suppose $U_b^- := \begin{pmatrix} u_b^- & 0 & p_b^- & \rho_b^- \end{pmatrix}^t$ is a constant supersonic state with $\rho_b^- > 0$ on Γ_{-1}, which represents supersonic flow entering the duct when $u_b^- > 0$. Hence the const in Bernoulli's law (2.3) is

$$c_0 = (u_b^-)^2/2 + (a_b^-)^2/(\gamma - 1)$$

and independent of streamlines. It is necessary to control the pressure at Γ_1 to obtain transonic flows in $\mathbb{P}$; otherwise the flow may be purely supersonic in $\mathbb{P}$, which is why we need the last condition in (2.9). However, the following simple but fundamental result indicates we may have trouble if we give p_1 in an arbitrary way. (Giving other conditions on Γ_1 instead of p will be discussed in section 11. In the following sections 2–10 we concentrate only on the typical case, i.e., problem (**A**).)

PROPOSITION 2.1. *For the special case $\Gamma(x) \equiv 1$, suppose the solution U to (2.9) depends only on x; then for given supersonic state U_b^-, there exists a unique constant $p_1 = p_b^+$ determined by U_b^- such that*

$$S_b : \{x = 0\} \tag{2.10}$$

is a shock front with uniform supersonic state U_b^- ahead of it (i.e., in $\{x < 0\}$) and uniform subsonic state $U_b^+ = \begin{pmatrix} u_b^+ & 0 & p_b^+ & \rho_b^+ \end{pmatrix}^t$ behind it (i.e., in $\{x > 0\}$). S_b and (U_b^-, U_b^+) make up a piecewise smooth weak entropy solution to (2.9) containing shocks. Here "uniform" means that the states $U_b^\pm$ are constant vectors.

Proof. From the one-dimensional steady Euler system it is obvious that any solution without a jump must be uniform. So $U \equiv U_b^-$ for $x \in [-1, 0]$. The Rankine–Hugoniot jump conditions (see [4, 13]) now takes the form

$$\rho u = \rho_b^- u_b^-, \quad \rho u^2 + p = \rho_b^- (u_b^-)^2 + p_b^-, \quad Eu = E_b^- u_b^-,$$

where $E = \frac{\gamma}{\gamma-1} p + \frac{1}{2}\rho |\mathbf{u}|^2$. By supersonic condition $u_b^- > a_b^-$ it is easy to get

$$\begin{aligned} u &= \frac{c_*}{u_b^-}, \qquad c_* := 2c_0 \frac{\gamma - 1}{\gamma + 1}, \\ \rho &= \frac{\rho_b^- (u_b^-)^2}{c_*}, \\ p &= \rho_b^- (u_b^-)^2 + p_b^- - c_* \rho_b^-. \end{aligned}$$

Direct calculation demonstrates that $u < a$ (subsonic) and the Lax entropy condition (see [13])

$$p > p_b^-. \tag{2.11}$$

So $U_b^+ \equiv U = (u, 0, p, S)^t$ for $x \in [0, 1]$, and thus $p_1 = p$ is uniquely determined by U_b^- as well as $U_b^\pm$. □

Remark 2.2. In the rest of this paper we call the above obtained $(U_b^\pm)$ and $S_b : \{x_0 = 0\}$ a *background solution* and denote it as $U_b = (U_b^-, U_b^+; S_b)$. Equation (2.10) is used to fix the position of the shock front since we may set $x = c$ for any $c \in (-1, 1)$ as the shock front and obtain the same $U_b^\pm$. One may also observe from the above result that the pressure $p = p_1$ at Γ_1 is necessary, though it cannot be given arbitrarily. The main result of this paper, Theorem 2.6 below, shows that for two-dimensional flow this observation is also valid.

2.2. Function spaces. Although the background solution provides us with some useful information, however, when $\Gamma(x)$ is slightly curved, rigorously solving problem (**A**) still involves several difficulties, as mentioned in the introduction. For subsonic flow, the steady Euler system is of hyperbolic-elliptic composite type: it has a real (generalized) eigenvalue of multiplicity 1 and a pair of conjugate complex eigenvalues. Due to conservation of mass we can introduce the Lagrangian transformation to reduce the original equations to two algebraic equations (Bernoulli's law and constancy of entropy along streamlines for C^1 flows) and a 2×2 system of partial differential equations, which is hyperbolic for supersonic flow, elliptic for subsonic flow, and of mixed type for transonic flow. Thus to obtain the subsonic flow behind the shock front S, we have to confront elliptic boundary value problems on rectangular domains

$$\Omega = \Big\{(x, y) : 0 \le y \le \Gamma(x), f(y) \le x \le 1\Big\}, \tag{2.12}$$

where $x = f(y)$ is the equation of the shock front S which satisfies $f(0) = 0$. Suppose S and Γ_+ intersect at the point $\Sigma_4 = (x_*, y_*)$. It is well known that the corners

$$\Sigma_1 = (0, 0), \quad \Sigma_2 = (1, 0), \quad \Sigma_3 = (1, \Gamma(1)), \quad \Sigma_4 = (x_*, y_*) \tag{2.13}$$

in general will cause singularities to the solutions, and the popular Schauder theory for $C^{2,\alpha} (\alpha \in (0, 1))$ domains (see [15]) may be invalid. The loss of regularity at the corners will also influence the regularity of the shock front S itself, which in turn has

an effect on the smoothness of the domain Ω, and thus may cause new trouble when solving the elliptic problems. The hyperbolic part may also transport the singularities at corners to other points in the duct. Fortunately, Gilbarg and Hörmander [16] and Lieberman [21] have established the *intermediate Schauder estimates* to attack elliptic problems on nonsmooth domain (see also the Notes of Chapter 6 in [15]), and their theory is powerful enough to handle our dilemma. Following their ideas, we introduce the function spaces $H_a^{(b)}(\Omega)$, $H_a^{'b}[0,y_*]$ to describe precisely the regularity of our desired subsonic flow and the transonic shock front, respectively. Our definition is a little different from theirs, but due to the boundary regularity estimates (Lemmas 6.18 and 6.29 in [15]), there is no problem later in using their theorems. Note that in the following we always suppose that $0 \le k < a = k+\alpha \le k+1, a+b > 0$, with k an integer and $\alpha \in (0,1]$ for such spaces.

The Banach spaces $H_a^{'(b)}[0,y_*]$ is defined as follows. A function f on $[0,y_*]$ is in $H_a^{'(b)}[0,y_*]$ if and only if

$$\|f\|_{a;[0,y_*]}^{'(b)} := \sup_{\delta>0} \delta^{a+b} \|f\|_{C^a[\delta,y_*-\delta]} < \infty, \tag{2.14}$$

with C^a here the usual Hölder space $C^{k,\alpha}$. We define $\|\cdot\|_{a;[0,y_*]}^{'(b)}$ as the norm of $H_a^{'(b)}[0,y_*]$.

The Banach space $H_a^{(b)}(\Omega)$ is by definition the set of functions ϕ defined on Ω with the property that

$$\|\phi\|_{a;\Omega}^{(b)} := \sup_{\delta>0} \delta^{a+b} \|\phi\|_{C^a(\Omega_\delta)} < \infty, \tag{2.15}$$

where

$$\Omega_\delta := \Big\{ P = (\bar\xi,\bar\eta) \in \Omega : \text{distance}(P,\Gamma^{\pm}) > \delta \Big\}. \tag{2.16}$$

Here Ω as in (2.12) with $f \in H_a^{'(b)}[0,y_*]$, while $\Gamma^+ := \{(x,\Gamma(x)) : x_* \le x \le 1\}, \Gamma^- := \{(x,0) : 0 \le x \le 1\}$ are, respectively, the upper and lower boundary of Ω.

Notice that the Lagrangian transformation also has the advantage that it straightens the curved boundary since it straightens the streamlines. So later by introducing certain homeomorphisms $\Phi : \Omega \to [0,1;0,1]$ which are of class $H_a^{(b)}(\Omega)$ with $b < -1$, we will actually solve elliptic boundary problems on the square $[0,1;0,1]$. For simplicity, we write $H_a^{(b)}([0,1;0,1])$ as $H_a^{(b)}$, and $H_a^{'(b)}[0,1]$ as $H_a^{'(b)}$. The corresponding norms are simply denoted as $\|\cdot\|_a^{(b)}, \|\cdot\|_a^{'(b)}$, respectively. By direct calculations one can verify the following.

PROPOSITION 2.2. *Suppose $b < -1$, Ω as before, $u \in H_a^{(b)}$, and $\Phi : \Omega \to [0,1;0,1]$ satisfy $\Phi \in H_a^{(b)}(\bar\Omega;\mathbb{R}^2)$. Then $u \circ \Phi \in H_a^{(b)}(\bar\Omega)$ and*

$$\| u \circ \Phi \|_{H_a^{(b)}(\bar\Omega)} \le C \| u \|_{H_a^{(b)}}, \tag{2.17}$$

where $C = C\big(n, \| \Phi \|_{H_a^{(b)}(\bar\Omega;\mathbb{R}^n)}\big)$.

A similar result also holds for $H_a^{'(b)}$. This means that the homeomorphisms introduced later (including the Lagrangian transformation) will not influence our resultant estimates.

We list here several useful properties of spaces $H_a^{(b)}$ (they also hold for $H_a^{'(b)}$) which are used later to obtain estimates of certain nonlinear terms. The proof and more information about this class of weighted Hölder spaces can be found in [16].

PROPOSITION 2.3.

$$\|\nabla\phi\|_{a-1}^{(1+b)} \le C\,\|\phi\|_a^{(b)} \qquad \text{if} \quad a>1, \tag{2.18}$$

$$\|\phi\|_a^{(b)} \le C\,\|\phi\|_a^{(b')} \qquad \text{if} \quad b'\le b. \tag{2.19}$$

PROPOSITION 2.4. *If $0\le a'\le a, a'+b\geqslant 0$, and b is not an integer ≤ 0, then*

$$\|\phi\|_{a'}^{(b)} \le C\,\|\phi\|_a^{(b)}. \tag{2.20}$$

PROPOSITION 2.5. *If $0\le c_j\le a+b, a\geqslant 0$, then*

$$\|\phi\psi\|_a^{(b)} \le C(\|\phi\|_a^{(b-c_1)}\,\|\psi\|_0^{(c_1)} + \|\phi\|_0^{(c_2)}\,\|\psi\|_a^{(b-c_2)}). \tag{2.21}$$

2.3. Main results. There are two main results in this paper. Now we can state precisely the first one as the following theorem. Another is indicated in Remark 2.6. For details, see section 11.

THEOREM 2.6. *There exists a $\varepsilon_0>0$ such that if*

$$\|\Gamma(x)-1\|_{C^5[-1,1]} \le \varepsilon < \varepsilon_0, \tag{2.22}$$

$$\frac{\mathrm{d}^k(\Gamma(x)-1)}{\mathrm{d}x^k}\Big|_{x=-1} = 0, \quad k=0,1,2,3,4,5, \tag{2.23}$$

then there is a unique $e\in\mathbb{R}$ with

$$|e| < C_0\varepsilon \tag{2.24}$$

such that (2.9) *with*

$$p_1 = p_b^+ + e \tag{2.25}$$

has a unique weak entropy solution $(U_-,U_+;S)$ with the following properties:

(i) *U_- is supersonic, U_+ is subsonic, and S is the shock front separating U_- and U_+ with entropy condition.*

(ii) *$S: x=f(y)$, $y\in[0,y_*]$, with y_* satisfying $y_*=\Gamma(f(y_*))$ and*

$$f(0)=0. \tag{2.26}$$

(iii) *For some $\alpha\in(0,1)$, the following estimates hold:*

$$\left\|U_- - U_b^-\right\|_{C^{3,\alpha}(\mathbb{P}_-)} < C_0\varepsilon, \tag{2.27}$$

$$\left\|U_+ - U_b^+\right\|_{2+\alpha;\Omega}^{(-\alpha)} < C_0\varepsilon, \tag{2.28}$$

$$\|f\|_{3+\alpha;[0,y_*]}^{'(-1-\alpha)} < C_0\varepsilon. \tag{2.29}$$

Here ε_0, C_0 are constants depending only on U_b, and we have

$$\mathbb{P}_- := \{(x,y)\in\mathbb{P} : x < f(y)\},$$

with Ω the same as in (2.12).

Remark 2.3. We suppose (2.26) holds to fix the position of the shock front. This is necessary, as indicated by the translation invariance along the x axis for the background solution: for the same p_1 in Proposition 2.1 at the exit, the position of the shock cannot be uniquely determined (see also Remark 2.2 and [4], [28]). We note that this phenomena is different from those transonic shocks observed in de Laval nozzles.

Remark 2.4. Equation (2.25) may be replaced by

$$p_1 = p_b^+ + g(y) + e \tag{2.30}$$

with e a constant to be determined simultaneously with U for any $g \in C^{2,\alpha}[0, \Gamma(1)]$ with small norm. There is no additional difficulty in the proof. Giving pressure at the exit in this way implies that the value of the pressure at the exit can be given only apart from a constant difference. We note that e in most cases does not vanish, as was shown by Proposition 2.1 if g is a nonzero number: it has already been determined by the coming upstream flow and the shape of the duct and g. This implies that for given pressure at the exit of the duct the transonic shock problem is ill-posed. As can be seen from the background solution, the ill-posedness is not related to the fact that we fixed the position of the shock.

Remark 2.5. Our proof can be modified to treat the case when the upstream supersonic flow at the entrance of the duct is also perturbed slightly if certain orders of compatibility conditions hold at the entrance, and a similar result can be proved. The major difference is that the constant in Bernoulli's law may be different on different streamlines. In [7] we have studied this case for the three-dimensional Euler system.

Remark 2.6. We note that the method developed in this paper provides us with more information than just presented in Theorem 2.6. It indicates clearly the well-posedness of giving v, w at Γ_1; ill-posedness of giving u, ρ, S, or the Mach number $M = |\mathbf{u}|/a$ as well as p there will be discussed in detail in section 11.

Remark 2.7. We emphasize here that the uniqueness of transonic shock in Theorem 2.6 is proved only in the class of functions satisfying properties (i)–(iii) listed there. The "global uniqueness" is an interesting open problem. Noting the nonuniqueness of transonic shocks claimed by Smith in [26] and symmetry breaking phenomena discussed by Kuz'min [19] and references therein, rigorous analysis of uniqueness or nonuniqueness of certain problems in aerodynamics is very important to understand some widely used models in practice and numerical simulations.

3. Euler equations in Lagrangian coordinates. The Euler equations (2.1), (2.2) are difficult to handle directly. In this section we use conservation of mass to write them in Lagrangian coordinates, which simplifies the geometry of the domain, as well as the "hyperbolic" part of the Euler system, as will be shown in the next section.

Set

$$w = \frac{v}{u}, \tag{3.1}$$

and denote the integral curves of

$$\begin{cases} \dfrac{\mathrm{d}\hat{y}(x,h)}{\mathrm{d}x} = w(x, \hat{y}(x,h)), \\ \hat{y}(0,h) = h \end{cases} \tag{3.2}$$

by

$$\begin{cases} x = \xi, \\ \hat{y} = \hat{y}(\xi, h). \end{cases} \tag{3.3}$$

(These are exactly streamlines.) Let

$$\eta = \eta(x, h) = \int_{\hat{y}(x,0)}^{\hat{y}(x,h)} \rho u(x, y) \mathrm{d}y \tag{3.4}$$

be the flux of mass between two such curves. Then by using the first equation in (2.1) we have

$$\begin{aligned} \partial_x \eta = & \rho u(x, \hat{y}(x, h)) \frac{\partial \hat{y}(x, h)}{\partial x} \\ & - \rho u(x, \hat{y}(x, 0)) \frac{\partial \hat{y}(x, 0)}{\partial x} - \int_{\hat{y}(x,0)}^{\hat{y}(x,h)} \partial_y (\rho v(x, y)) \mathrm{d}y \\ = & \, 0. \end{aligned} \tag{3.5}$$

Hence $\eta = \eta(-1, h)$ and $\eta(-1, 0) = 0$. Thus if

$$\frac{\partial \eta(0, h)}{\partial h} = \rho u(0, \hat{y}(0, h)) \frac{\partial \hat{y}(0, h)}{\partial h} = \rho u(0, \hat{y}(0, h)) \neq 0, \tag{3.6}$$

we may obtain the inverse function $h = h(\eta)$ of $\eta = \eta(-1, h)$ and $h(0) = 0$. Set

$$y(x, \eta) = \hat{y}(x, h(\eta)); \tag{3.7}$$

then (3.4) becomes

$$\eta = \int_{y(x,0)}^{y(x,\eta)} \rho u(x, s) \mathrm{d}s, \tag{3.8}$$

and by differentiating it with η one has

$$\frac{\partial y}{\partial \eta} = \frac{1}{\rho u}. \tag{3.9}$$

Now we introduce the following Lagrangian transformation $(x, y) \mapsto (\xi, \eta)$:

$$\begin{cases} x = \xi, \\ y = y(\xi, \eta). \end{cases} \tag{3.10}$$

Then by

$$\frac{\partial (x, y)}{\partial (\xi, \eta)} = \begin{pmatrix} 1 & 0 \\ w & \frac{1}{\rho u} \end{pmatrix}, \tag{3.11}$$

we have

$$\frac{\partial (\xi, \eta)}{\partial (x, y)} = \begin{pmatrix} 1 & 0 \\ -\rho v & \rho u \end{pmatrix}; \tag{3.12}$$

thus

$$\begin{cases} \partial_x = \partial_\xi - \rho v \partial_\eta, \\ \partial_y = \rho u \partial_\eta. \end{cases} \tag{3.13}$$

So a little computation shows that (2.1) or (2.4) may be written as conservation laws

$$\begin{cases} \partial_\xi(\frac{1}{\rho u}) - \partial_\eta w = 0 & \text{(conservation of mass)}, \\ \partial_\xi(u + \frac{p}{\rho u}) - \partial_\eta(pw) = 0 & \text{(conservation of momentum along } \xi), \\ \partial_\xi v + \partial_\eta p = 0 & \text{(conservation of momentum along } \eta), \end{cases} \tag{3.14}$$

or as symmetric system

$$A\partial_\xi U + B\partial_\eta U = 0, \tag{3.15}$$

with

$$A = \begin{pmatrix} u & 0 & \frac{1}{\rho} \\ 0 & u & 0 \\ \frac{1}{\rho} & 0 & \frac{u}{\rho^2 a^2} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & -v \\ 0 & 0 & u \\ -v & u & 0 \end{pmatrix} \tag{3.16}$$

and

$$U = \begin{pmatrix} u & v & p \end{pmatrix}^t. \tag{3.17}$$

The above transformation is valid if and only if

$$\rho u \neq 0. \tag{3.18}$$

Note that in (3.16) the last equation corresponds to conservation of mass.

Next we consider boundary conditions on $\Gamma_\pm$. By (2.9) and (3.2) (and uniqueness of solutions to Cauchy problems of ODEs) $\Gamma_\pm$ are streamlines. Thus Γ_- in the (ξ, η) coordinates is

$$\tilde{\Gamma}_- : \ \eta = 0, \quad \xi \in [-1, 1], \tag{3.19}$$

while Γ_+ is

$$\tilde{\Gamma}_+ : \ \eta = \eta_0 = \int_0^1 \rho u(-1, s) \mathrm{d}s, \quad \xi \in [-1, 1]. \tag{3.20}$$

In the latter, without loss of generality we always suppose that $\eta_0 \equiv 1$ by suitable normalization of the unit of the coming flow. The corresponding boundary conditions are

$$\begin{cases} w = 0 & \text{on} \quad \tilde{\Gamma}_-, \\ w = \Gamma'(\xi) & \text{on} \quad \tilde{\Gamma}_+. \end{cases} \tag{3.21}$$

Since the shock front may be curved, with the equation

$$\xi = \psi(\eta), \quad \eta \in [0, 1], \tag{3.22}$$

we introduce further the following transformation $\mathbf{\Phi}_\psi : (\xi, \eta) \mapsto (\bar{\xi}, \bar{\eta})$ to straighten the shock front:

$$\begin{cases} \bar{\xi} = \dfrac{\xi - \psi(\eta)}{1 - \psi(\eta)}, \\ \bar{\eta} = \eta, \end{cases} \quad \text{or} \quad \begin{cases} \xi = (1 - \psi(\bar{\eta}))\bar{\xi} + \psi(\bar{\eta}), \\ \eta = \bar{\eta}. \end{cases} \tag{3.23}$$

So

$$(\bar{\xi}, \bar{\eta}) \in [0, 1; 0, 1]$$

and

$$\begin{cases} \dfrac{\partial}{\partial \xi} = \dfrac{1}{1 - \psi(\bar{\eta})} \dfrac{\partial}{\partial \bar{\xi}}, \\ \dfrac{\partial}{\partial \eta} = \dfrac{\partial}{\partial \bar{\eta}} + \dfrac{(\bar{\xi} - 1)\psi'(\bar{\eta})}{1 - \psi(\bar{\eta})} \dfrac{\partial}{\partial \bar{\xi}}. \end{cases} \tag{3.24}$$

Thus (3.15) becomes

$$\bar{A}\partial_{\bar{\xi}} U + \bar{B}\partial_{\bar{\eta}} U = 0, \tag{3.25}$$

with

$$\bar{A} = A + (\bar{\xi} - 1)\psi'(\bar{\eta})B, \quad \bar{B} = (1 - \psi(\bar{\eta}))B. \tag{3.26}$$

Finally, note that after passing the shock front, (3.21) in the $(\bar{\xi}, \bar{\eta})$ coordinates is

$$\begin{cases} w = 0 & \text{on} \quad \bar{\Gamma}_- := \{(\bar{\xi}, 0) : \bar{\xi} \in [0, 1]\}, \\ w = \Gamma'((1 - \psi(1))\bar{\xi} + \psi(1)) & \text{on} \quad \bar{\Gamma}_+ := \{(\bar{\xi}, 1) : \bar{\xi} \in [0, 1]\}. \end{cases} \tag{3.27}$$

4. Decomposition of elliptic-hyperbolic composite system. The idea involved in this section to write (3.25) as separate elliptic and hyperbolic equations is rudimentary. Let λ be a generalized eigenvalue of $\bar{B}$ with respect to $\bar{A}$:

$$\det(\lambda \bar{A} - \bar{B}) = 0, \tag{4.1}$$

and let the corresponding generalized left (row) eigenvector be l, i.e.,

$$l\bar{B} = \lambda l \bar{A}, \tag{4.2}$$

then multiply (3.25) from the left by l to get

$$l\bar{A}(\partial_{\bar{\xi}} + \lambda \partial_{\bar{\eta}})U = 0. \tag{4.3}$$

Now suppose λ (and thus l) is complex:

$$\lambda = \lambda_R + i\lambda_I, \quad l = l_R + i l_I, \quad i = \sqrt{-1}; \tag{4.4}$$

then (4.3) is equivalent to

$$\begin{cases} l_R \bar{A}\partial_I U + l_I \bar{A}\partial_R U = 0, \\ l_R \bar{A}\partial_R U - l_I \bar{A}\partial_I U = 0, \end{cases} \tag{4.5}$$

with

$$\partial_R = \partial_{\bar\xi} + \lambda_R \partial_{\bar\eta}, \qquad \partial_I = \lambda_I \partial_{\bar\eta}. \tag{4.6}$$

So roughly speaking, the real eigenvalue λ corresponds to the hyperbolic equation (4.3) while the complex λ and its conjugation $\bar\lambda$ correspond to the elliptic system (4.5).

System (3.25) has a real eigenvalue λ_0 and a pair of complex eigenvalues $\lambda = \lambda_R \pm i\lambda_I$ if the Mach number $M = |\mathbf{u}|/a < 1$ (the computation is straightforward and we omit it):

$$\lambda_0 = 0, \tag{4.7}$$

$$\lambda_R = -\frac{(1-\psi)\big((\bar\xi-1)\psi'(u^2+v^2) - \frac{v}{\rho}\big)}{\frac{1}{\rho^2}\left(\frac{u^2}{a^2}-1\right) - (\bar\xi-1)^2\psi'^2(u^2+v^2) + \frac{2}{\rho}(\bar\xi-1)\psi' v}, \tag{4.8}$$

$$\lambda_I = \frac{(1-\psi)\frac{u}{\rho}\sqrt{1-M^2}}{\frac{1}{\rho^2}\left(\frac{u^2}{a^2}-1\right) - (\bar\xi-1)^2\psi'^2(u^2+v^2) + \frac{2}{\rho}(\bar\xi-1)\psi' v}. \tag{4.9}$$

Here and in the following we write $\psi = \psi(\bar\eta)$, $\psi' = \psi'(\bar\eta)$. The corresponding left eigenvectors are

$$l_0 = \begin{pmatrix} u & v & 0 \end{pmatrix}, \tag{4.10}$$

$$l_R = \begin{pmatrix} \frac{\lambda_R}{\rho} - \lambda_R(\bar\xi-1)v\psi' + (1-\psi)v & \lambda_R(\bar\xi-1)\psi' u - (1-\psi)u & -\lambda_R u \end{pmatrix}, \tag{4.11}$$

$$l_I = \begin{pmatrix} \frac{\lambda_I}{\rho} - \lambda_I(\bar\xi-1)v\psi' & \lambda_I(\bar\xi-1)\psi' u & -\lambda_I u \end{pmatrix}. \tag{4.12}$$

Thus

$$l_0\bar A = \begin{pmatrix} u^2 & uv & \frac{u}{\rho} \end{pmatrix}, \tag{4.13}$$

$$l_R\bar A = \begin{pmatrix} (1-\psi)uv & -(1-\psi)u^2 & 0 \end{pmatrix}, \tag{4.14}$$

$$l_I\bar A = \begin{pmatrix} 0 & 0 & -(1-\psi)u\frac{\sqrt{1-M^2}}{\rho} \end{pmatrix}. \tag{4.15}$$

By (4.3), (4.7), and (4.13) we get the hyperbolic equation

$$\frac{1}{2}\partial_{\bar\xi}(u^2+v^2) + \frac{1}{\rho}\partial_{\bar\xi} p = 0 \tag{4.16}$$

if $u \neq 0$. On the other hand, from Bernoulli's law (2.3) one gets

$$\frac{1}{2}\partial_{\bar\xi}(u^2+v^2) + \frac{1}{\gamma-1}\partial_{\bar\xi} a^2 = 0;$$

hence

$$\frac{1}{\rho}\partial_{\bar\xi} p - \frac{1}{\gamma-1}\partial_{\bar\xi} a^2 = 0.$$

By $p = A(S)\rho^\gamma, a^2 = \gamma A(S)\rho^{\gamma-1}$ the above equation is actually the constancy of entropy along streamlines for C^1 solutions

$$\partial_{\bar\xi}\left(\frac{p}{\rho^\gamma}\right) = 0. \tag{4.17}$$

Similarly, we can write (4.5) as

$$\begin{cases} \partial_{\bar\xi} p + \lambda_R \partial_{\bar\eta} p - \beta_1 \partial_{\bar\eta} w = 0, \\ \partial_{\bar\xi} w + \beta_2 \partial_{\bar\eta} p + \lambda_R \partial_{\bar\eta} w = 0, \end{cases} \tag{4.18}$$

with

$$\begin{cases} \beta_1 = -\dfrac{(1-\psi)u^3}{\frac{1}{\rho^2}\left(\frac{u^2}{a^2}-1\right) - (\bar\xi-1)^2\psi'^2(u^2+v^2) + \frac{2}{\rho}(\bar\xi-1)\psi' v}, \\ \beta_2 = -\dfrac{\frac{1}{u\rho^2}(1-\psi)(1-M^2)}{\frac{1}{\rho^2}\left(\frac{u^2}{a^2}-1\right) - (\bar\xi-1)^2\psi'^2(u^2+v^2) + \frac{2}{\rho}(\bar\xi-1)\psi' v}. \end{cases} \tag{4.19}$$

When $M < 1$, (4.18) is an elliptic system. However, for $M > 1$, i.e., supersonic flow, we can also carry out similar calculations to obtain (4.18), which is a hyperbolic system. Thus if

$$\det \begin{pmatrix} l_0 \\ l_R \\ l_I \end{pmatrix} = (1-\psi)\lambda_I u(u^2+v^2) \tag{4.20}$$

is nonzero, (4.17) and (4.18) are equivalent to (3.25) for C^1 solutions. This is true if

$$\rho u \neq 0, \quad M \neq 1, \quad u \neq a, \qquad \|\psi\|_{C^1} \ll 1. \tag{4.21}$$

Remark 4.1. The first equation in (4.18) is in essence the conservation of mass. In fact, it is obtained by multiplying (3.25) from the left by l_I. Notice that the third argument in l_I is nonzero and the last equation in (3.25) is the conservation of mass.

5. Existence of supersonic flow. In this section we always set $\psi \equiv 0$, and we will use (4.18) to show existence and uniqueness of supersonic flow in the duct $\mathbb{P}$ when its boundary is slightly curved, without considering the conditions at the exit. Now (4.18) is

$$\partial_\xi \begin{pmatrix} w \\ p \end{pmatrix} + \begin{pmatrix} \varpi & \beta \\ \kappa & \varpi \end{pmatrix} \partial_\eta \begin{pmatrix} w \\ p \end{pmatrix} = 0, \tag{5.1}$$

with

$$\varpi = \frac{\rho v a^2}{u^2 - a^2}, \quad \beta = \frac{a^2(M^2-1)}{u(u^2-a^2)}, \quad \kappa = \frac{\rho^2 a^2 u^3}{u^2-a^2}.$$

Consider the following mixed initial-boundary value problem:

$$\begin{cases} (5.1) & \text{in} \quad \tilde{\mathbb{P}} = [-1,1]\times[0,1], \\ p = p_b^- & \text{on} \quad \tilde\Gamma_{-1} : \xi = -1, \\ w = 0 & \text{on} \quad \tilde\Gamma_{-1} : \xi = -1, \\ w = 0 & \text{on} \quad \tilde\Gamma_- : \eta = 0, \\ w = \Gamma'(\xi) & \text{on} \quad \tilde\Gamma_+ : \eta = 1, \end{cases} \tag{5.2}$$

where

$$p = d_0 \rho^\gamma, \quad \text{or} \quad \rho = \left(\frac{p}{d_0}\right)^{\frac{1}{\gamma}}, \tag{5.3}$$

$$u = \left\{ \frac{2}{1+w^2} \left(c_0 - \frac{\gamma d_0}{\gamma - 1} \left(\frac{p}{d_0} \right)^{\frac{\gamma-1}{\gamma}} \right) \right\}^{\frac{1}{2}}, \tag{5.4}$$

with

$$d_0 = p_b^- / (\rho_b^-)^\gamma. \tag{5.5}$$

Equation (5.3) originates from (4.17), and (5.4) comes from Bernoulli's law (2.3) and (5.3).

Remark 5.1. The system (5.1) is genuinely nonlinear in a neighborhood of U_b^-. Direct computation shows that the eigenvalues are

$$\lambda_\pm = \varpi \pm \sqrt{\beta\kappa} = \frac{\rho a^2}{u^2 - a^2}(v \pm u\sqrt{M^2 - 1}), \tag{5.6}$$

and the corresponding left (resp., right) eigenvectors $l_\pm$ ($r_\pm$) are

$$l_\pm = \begin{pmatrix} \pm\sqrt{\kappa} & \sqrt{\beta} \end{pmatrix}, \tag{5.7}$$

$$r_\pm = \begin{pmatrix} \pm\sqrt{\beta} & \sqrt{\kappa} \end{pmatrix}^t. \tag{5.8}$$

Thus

$$\nabla\lambda_\pm \cdot r_\pm(U_b^-) = \frac{1}{2}(1+\gamma) \frac{\rho u^4 \sqrt{u}}{(u^2 - a^2)^2} \bigg|_{U = U_b^-} \neq 0. \tag{5.9}$$

Set $W = (w, p)^t$. The characteristic form of (5.1) is

$$(\sqrt{\kappa}, \sqrt{\beta})(\partial_\xi + (\varpi + \sqrt{\beta\kappa})\partial_\eta)W = 0, \tag{5.10}$$

$$(-\sqrt{\kappa}, \sqrt{\beta})(\partial_\xi + (\varpi - \sqrt{\beta\kappa})\partial_\eta)W = 0. \tag{5.11}$$

THEOREM 5.1. *There exists a positive ε_0 such that if* (2.22) *and* (2.23) *hold, then problem* (5.2) *has a unique solution* (w, p) *satisfying*

$$\|w\|_{C^{3,\alpha}(\tilde{\mathbb{P}})} + \left\|p - p_b^-\right\|_{C^{3,\alpha}(\tilde{\mathbb{P}})} \leq C_0\varepsilon. \tag{5.12}$$

The constants ε_0, C_0 depend solely on U_b^-, and $\alpha \in (0,1)$ may be arbitrary.

COROLLARY 5.2. *Under the same assumptions of Theorem* 5.1*, by* (3.1)*,* (5.3)*,* (5.4)*, and* (5.12) *we also have*

$$\left\|u - u_b^-\right\|_{C^{3,\alpha}(\mathbb{P})} + \|v\|_{C^{3,\alpha}(\mathbb{P})} + \left\|\rho - \rho_b^-\right\|_{C^{3,\alpha}(\mathbb{P})} + \left\|p - p_b^-\right\|_{C^{3,\alpha}(\mathbb{P})} \leq C_0\varepsilon. \tag{5.13}$$

Remark 5.2. Hereafter we denote the u, v, p, ρ, w, S obtained in Theorem 5.1 and Corollary 5.2 as $u_-, v_-, p_-, \rho_-, w_-, S_-$ in accordance with the notation in Theorem 2.6.

The proof of Theorem 5.1 is standard; it just needs a little modification of the proof of Theorem 3.3 in Chapter 4 of [20] (p. 180). (We may get local existence directly by this theorem.) In fact, if the boundary condition is a small perturbation of zero, then the existence can be semiglobal. It means that the life span of the smooth solution depends on the smallness of the perturbation of boundary data. In other words, the life span can be larger than any given number, provided the perturbation is small enough.

Remark 5.3. For $\Gamma'(\xi) < 0$ and the case when perturbation is not small, Chen has proved in [10] that the solution may blow up in finite distance from the entrance and shocks will appear.

6. Free boundary problem (B) and fixed boundary problem (C).

6.1. Problem (B). Knowing the supersonic flow, now we are in the position to determine the shock front and the subsonic state behind it simultaneously, satisfying the restrictions of pressure at the exit. We formulate it as the free boundary problem (**B**).

Let

$$S : \xi = \psi(\eta), \qquad \eta \in [0, 1], \tag{6.1}$$

be the shock front. By (3.14) the following Rankine–Hugoniot jump conditions [12] should hold across S:

$$-\left[\frac{1}{\rho u}\right] = [w]\ \psi'(\eta), \tag{6.2}$$

$$-\left[u + \frac{p}{\rho u}\right] = [pw]\ \psi'(\eta), \tag{6.3}$$

$$[v] = [p]\ \psi'(\eta). \tag{6.4}$$

By (6.4) we have

$$\psi'(\eta) = \frac{[v]}{[p]}. \tag{6.5}$$

Due to Remark 2.3 concerning (2.26), we set

$$\psi(0) = 0. \tag{6.6}$$

Substituting (6.5) in (6.2) and (6.3), we have

$$G_1(U, U_-) := [w][uw] + \left[\frac{1}{\rho u}\right][p] = 0, \tag{6.7}$$

$$G_2(U, U_-) := [pw][uw] + \left[u + \frac{p}{\rho u}\right][p] = 0. \tag{6.8}$$

Here $U := \begin{pmatrix} u & w & p \end{pmatrix}^t$, and $U_- := \begin{pmatrix} u_- & w_- & p_- \end{pmatrix}^t$ with $U_- = U_-(\psi(\eta), \eta)$. Note that (6.5), (6.7), and (6.8) are equivalent to (6.2)–(6.4) provided $[p] \neq 0$, which is guaranteed by (2.11) if the perturbations are small.

Problem (**B**) can be stated now as the following:

Find $U, \psi(\eta)$ and a real number e such that
(i) $\psi(\eta)$ satisfies (6.1), (6.5), (6.6);
(ii) (6.7), (6.8) hold on S;
(iii) $w = 0$ on $\tilde{\Gamma}_-$;
(iv) $w = \Gamma'(\xi)$ on $\tilde{\Gamma}_+$;
(v) $p = p_b^+ + e$ on $\tilde{\Gamma}_1$.
(vi) Set $\Omega_\psi := \{(\xi, \eta) : \eta \in [0, 1], \psi(\eta) \leq \xi \leq 1\}$; then (4.17), (4.18) should hold in Ω_ψ as well as Bernoulli's law (2.3).

6.2. Problem (C). The idea of dealing with problem (**B**) is, roughly speaking, by iteration: first we fix the boundary and solve a fixed boundary problem, use (6.5), (6.6) to update the boundary, and then solve another fixed boundary problem, etc.

Set

$$\mathcal{S}_\sigma = \Big\{\psi(\eta) \in H^{'(-1-\alpha)}_{3+\alpha}[0,1] : \|\psi\|^{'(-1-\alpha)}_{3+\alpha;[0,1]} \le \sigma, \psi(0) = 0\Big\} \tag{6.9}$$

with

$$\sigma \le \sigma_0 < \frac{1}{2}. \tag{6.10}$$

For any $\psi \in \mathcal{S}_\sigma$, we may use transformation (3.23), which is of class $H^{(-1-\alpha)}_{3+\alpha;\Omega_\psi}$, to state the fixed boundary problem $(\mathbf{C}_\psi)$ as follows:

Find U and $e \in \mathbb{R}$ such that
(i) (4.17), (4.18), and (2.3) hold in $\Omega := [0,1] \times [0,1]$;
(ii) (3.27) holds on $\bar{\Gamma}_\pm$;
(iii) (6.7), (6.8) hold on $\bar{\xi} = 0$;
(iv) $p = p_b^+ + e$ holds on $\bar{\xi} = 1$.

Now if problem $(\mathbf{C}_\psi)$ is uniquely solvable, with the solution U_ψ, then by the Cauchy problem of the ODE (note that $\eta = \bar{\eta}$)

$$\begin{cases} \tilde{\psi}'(\eta) = \frac{[v_\psi]}{[p_\psi]}, \\ \tilde{\psi}(0) = 0, \end{cases} \tag{6.11}$$

later (section 10) we will construct a mapping $\Psi : \mathcal{S}_\sigma \to \mathcal{S}_\sigma$ given by $\Psi(\psi) = \tilde{\psi}$ if ε_0 in Theorem 2.6 is small. Clearly the fixed point $\bar{\psi}$ of Ψ corresponds to the desired shock front in problem $(\mathbf{B})$, and the solution $U_{\bar{\psi}}$ obtained by problem $(\mathbf{C}_{\bar{\psi}})$ is the subsonic state we are looking for. We call Ψ the *boundary modifying mapping*.

7. Problem (D): An equivalent form of problem (C). This section is devoted to writing problem $(\mathbf{C}_\psi)$ in an equivalent, but more transparent and tractable, form called problem $(\mathbf{D}_\psi)$. This is a nonlinear boundary problem for nonlinear systems.

We first deal with the boundary conditions. Since $G_i(U_b^+, U_b^-) = 0$ for $i = 1, 2$ holds, we may write (6.7), (6.8) as

$$\begin{aligned} &\nabla_+ G_i(U_b^+, U_b^-) \cdot (U - U_b^+) \\ &\quad = \nabla_+ G_i(U_b^+, U_b^-) \cdot (U - U_b^+) - (G_i(U, U_b^-) - G_i(U_b^+, U_b^-)) \\ &+ (G_i(U, U_b^-) - G_i(U, U_-)) \\ &\quad := g_i(U, U_-), \end{aligned} \tag{7.1}$$

where $\nabla_+ G_i(U, U_-)$ is the gradient of $G_i(U, U_-)$ with respect to the variables U. By direct calculations (note that here by Bernoulli's law we consider ρ as a function of p, w, u),

$$\begin{aligned} \nabla_+ G_i(U_b^+, U_b^-) &= \left.\frac{\partial(G_1, G_2)(U, U_-)}{\partial(u, w, p)}\right|_{(U,U_-)=(U_b^+, U_b^-)} \\ &= \begin{pmatrix} a_1 & 0 & b_1 \\ a_2 & 0 & b_2 \end{pmatrix} \end{aligned} \tag{7.2}$$

with

$$\begin{aligned}
a_1 &= -\frac{2c_0+u^2}{2c_0-u^2}\cdot\frac{[p]}{\rho u^2}\Big|_{(U_b^+,U_b^-)}\\
&= -\frac{2c_0+{u_b^+}^2}{2c_0-{u_b^+}^2}\cdot\frac{p_b^+-p_b^-}{\rho_b^+{u_b^+}^2},\\
b_1 &= -\frac{[p]}{\rho up}\Big|_{(U_b^+,U_b^-)},\\
a_2 &= [p]\left(\frac{1}{\gamma}-\frac{p}{\rho u^2}\right)\Big|_{(U_b^+,U_b^-)},\\
b_2 &= \left[u+\frac{p}{\rho u}\right]\Big|_{(U_b^+,U_b^-)}\\
&= 0.
\end{aligned}$$

Thus

$$d_1 := \det\begin{pmatrix} a_1 & b_1\\ a_2 & b_2\end{pmatrix} = \frac{[p]^2}{\rho up}\left(\frac{1}{\gamma}-\frac{p}{\rho u^2}\right)\Big|_{(U_b^+,U_b^-)} = -\frac{[p]^2}{\gamma\rho up}\left(\frac{a^2}{u^2}-1\right)\Big|_{(U_b^+,U_b^-)} \neq 0,$$

and we may rewrite (7.1) as

$$p-p_b^+ = \frac{1}{d_1}(a_1g_2-a_2g_1) := h_1(U,U_-), \tag{7.3}$$

$$u-u_b^+ = \frac{1}{d_1}(b_2g_1-b_1g_2) := h_2(U,U_-). \tag{7.4}$$

They should hold on $\bar{\xi}=0$.

Next we manipulate the equations. In the following we denote the value of U on $\bar{\xi}=0$ as U_0. For example, by (7.3), (7.4), $p_0 = h_1 + p_b^+, u_0 = h_2 + u_b^+$.

Now by (4.17) and Bernoulli's law (2.3) we get

$$\rho_0 = \frac{\gamma}{\gamma-1}\cdot\frac{p_0}{c_0-\frac{1}{2}u_0^2(1+w_0^2)}, \tag{7.5}$$

$$\rho = \rho_0\left(\frac{p}{p_0}\right)^{\frac{1}{\gamma}}, \tag{7.6}$$

$$u = \left\{\frac{2}{1+w^2}\cdot\left(c_0-\frac{\gamma}{\gamma-1}\cdot\frac{p}{\rho}\right)\right\}^{\frac{1}{2}}, \tag{7.7}$$

while w, p may be solved from (4.18).

Let

$$\lambda_i := \beta_i|_{U=U_b^+,\psi=0} \qquad \text{for} \quad i=1,2, \tag{7.8}$$

with β_i defined as in (4.19), and let

$$f_1(U,\psi) = -\lambda_R\partial_{\bar{\eta}}p + (\beta_1-\lambda_1)\partial_{\bar{\eta}}w, \tag{7.9}$$

$$f_2(U,\psi) = -\lambda_R\partial_{\bar{\eta}}w + (\lambda_2-\beta_2)\partial_{\bar{\eta}}p; \tag{7.10}$$

then λ_i is positive and we may write (4.18) as

$$\begin{cases} \partial_{\bar\xi} p - \lambda_1 \partial_{\bar\eta} w = f_1(U,\psi), \\ \partial_{\bar\xi} w + \lambda_2 \partial_{\bar\eta} p = f_2(U,\psi). \end{cases} \tag{7.11}$$

For subsonic flow this is a first order nonlinear elliptic system.

So far problem (**C**) may be expressed in the following equivalent way if (4.21) holds.

Problem (**D1**)—boundary value problem for a first order elliptic system:

$$\begin{cases} \partial_{\bar\xi} p - \lambda_1 \partial_{\bar\eta} w = f_1(U,\psi) & \text{in} \quad \Omega, \\ \partial_{\bar\xi} w + \lambda_2 \partial_{\bar\eta} p = f_2(U,\psi) & \text{in} \quad \Omega, \\ p = p_b^+ + h_1(U,U_-) & \text{on} \quad \bar\xi = 0, \\ p = p_b^+ + e & \text{on} \quad \bar\xi = 1, \\ w = 0 & \text{on} \quad \bar\eta = 0, \\ w = \Gamma'((1-\psi(1))\bar\xi + \psi(1)) & \text{on} \quad \bar\eta = 1. \end{cases} \tag{7.12}$$

Problem (**D2**)—algebraic equations (recall that $U_0 = U|_{\bar\xi=0}$):

$$\begin{cases} u_0 = u_b^+ + h_2(U,U_-), \\ p_0 = p_b^+ + h_1(U,U_-), \\ \rho_0 = \dfrac{\gamma}{\gamma-1} \cdot \dfrac{p_0}{c_0 - \frac{1}{2}u_0^2(1+w_0^2)}, \\ \rho = \rho_0 \left(\frac{p}{p_0}\right)^{\frac{1}{\gamma}}, \\ u = \left\{ \frac{2}{1+w^2} \cdot \left(c_0 - \frac{\gamma}{\gamma-1} \cdot \frac{p}{\rho} \right) \right\}^{\frac{1}{2}}. \end{cases} \tag{7.13}$$

We call the above two coupled problems problem ($\mathbf{D}_\psi$) (or problem (**D**) for simplicity). The equivalence for smooth solutions is obvious from the deductions in the above sections.

8. Solving linearized problem (D). It is nature and standard to use iteration methods, such as the Banach contraction mapping principle, to solve problem (**D**). Thus in this section we concentrate on the related "linearized" problems.

8.1. Linearized problem (D1). Problem **D1** solves $\bar p, \bar w$, and $e \in \mathbb{R}$ satisfying

$$\begin{cases} \partial_{\bar\xi} \bar p - \lambda_1 \partial_{\bar\eta} \bar w = f_1 & \text{in} \quad \Omega, \\ \partial_{\bar\xi} \bar w + \lambda_2 \partial_{\bar\eta} \bar p = f_2 & \text{in} \quad \Omega, \\ \bar p = p_b^+ + h_1 & \text{on} \quad \bar\xi = 0, \\ \bar p = p_b^+ + e & \text{on} \quad \bar\xi = 1, \\ \bar w = 0 & \text{on} \quad \bar\eta = 0, \\ \bar w = g(\bar\xi) & \text{on} \quad \bar\eta = 1, \end{cases} \tag{8.1}$$

where f_1, f_2, h_1, g are suitable nonhomogeneous terms. This is a boundary value problem on a domain with a piecewise smooth boundary, so we need a generalized version of the usual Schauder theory for elliptic equations.

We may separate problem (8.1) into the following two problems:

$$\begin{cases} \partial_{\bar\xi}\bar p_1 - \lambda_1\partial_{\bar\eta}\bar w_1 = f_1 & \text{in} \quad \Omega, \\ \partial_{\bar\xi}\bar w_1 + \lambda_2\partial_{\bar\eta}\bar p_1 = 0 & \text{in} \quad \Omega, \\ \bar p_1 = p_b^+ + h_1 & \text{on} \quad \bar\xi = 0, \\ \bar p_1 = p_b^+ + e & \text{on} \quad \bar\xi = 1, \\ \bar w_1 = 0 & \text{on} \quad \bar\eta = 0, \\ \bar w_1 = g(\bar\xi) & \text{on} \quad \bar\eta = 1; \end{cases} \tag{8.2}$$

$$\begin{cases} \partial_{\bar\xi}\bar p_2 - \lambda_1\partial_{\bar\eta}\bar w_2 = 0 & \text{in} \quad \Omega, \\ \partial_{\bar\xi}\bar w_2 + \lambda_2\partial_{\bar\eta}\bar p_2 = f_2 & \text{in} \quad \Omega, \\ \bar p_2 = 0 & \text{on} \quad \bar\xi = 0, \\ \bar p_2 = 0 & \text{on} \quad \bar\xi = 1, \\ \bar w_2 = 0 & \text{on} \quad \bar\eta = 0, \\ \bar w_2 = 0 & \text{on} \quad \bar\eta = 1. \end{cases} \tag{8.3}$$

Then

$$\bar p = \bar p_1 + \bar p_2, \qquad \bar w = \bar w_1 + \bar w_2 \tag{8.4}$$

is the solution of problem (8.1). Since Ω is simply connected, we may introduce potentials $\phi_1(\bar\xi,\bar\eta), \phi_2(\bar\xi,\bar\eta)$ such that

$$\partial_{\bar\xi}\phi_1 = -\lambda_2(\bar p_1 - p_b^+), \qquad \partial_{\bar\eta}\phi_1 = \bar w_1, \tag{8.5}$$

$$\partial_{\bar\xi}\phi_2 = \lambda_1\bar w_2, \qquad \partial_{\bar\eta}\phi_2 = \bar p_2, \tag{8.6}$$

and write (8.2), (8.3) as

$$\begin{cases} \frac{1}{\lambda_2}\partial_{\bar\xi}^2\phi_1 + \lambda_1\partial_{\bar\eta}^2\phi_1 = -f_1 & \text{in} \quad \Omega, \\ \partial_{\bar\xi}\phi_1 = -\lambda_2 h_1 & \text{on} \quad \bar\xi = 0, \\ \partial_{\bar\xi}\phi_1 = -\lambda_2 e & \text{on} \quad \bar\xi = 1, \\ \partial_{\bar\eta}\phi_1 = 0 & \text{on} \quad \bar\eta = 0, \\ \partial_{\bar\eta}\phi_1 = g(\bar\xi) & \text{on} \quad \bar\eta = 1; \end{cases} \tag{8.7}$$

$$\begin{cases} \frac{1}{\lambda_1}\partial_{\bar\xi}^2\phi_2 + \lambda_2\partial_{\bar\eta}^2\phi_2 = f_2 & \text{in} \quad \Omega, \\ \phi_2 = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{8.8}$$

Equation (8.8) is the Dirichlet problem for Poisson equations. By Theorem 7.2 and Remark (2) following it in [16], and Theorem 1.4 in [21], we know there is a unique solution ϕ_2 and

$$\|\phi_2\|_{3+\alpha}^{(-1-\alpha)} \le C\,\|f_2\|_{1+\alpha}^{(1-\alpha)}, \tag{8.9}$$

where C depends only on U_b and σ_0, and $\alpha \in (0,\alpha_0)$ may be arbitrary for a fixed α_0.

Next we consider (8.7), which is actually the Neumann problem for Poisson equations:

$$\begin{cases} \triangle\phi = f & \text{in} \quad \Omega, \\ \frac{\partial\phi}{\partial\nu} = g & \text{on} \quad \partial\Omega, \end{cases}$$

where ν is the unit outward normal of $\partial\Omega$. It is well known that such a problem is solvable if and only if

$$\int_\Omega f \ \mathrm{d}\bar{\xi}\mathrm{d}\bar{\eta} = \int_{\partial\Omega} g \ \mathrm{d}s. \tag{8.10}$$

This is in essence the reason why we have to introduce the number e in the pressure giving at the exit of the duct.

Applying (8.10) to (8.7), we get

$$\int_\Omega f_1 \ \mathrm{d}\bar{\xi}\mathrm{d}\bar{\eta} = \int_0^1 (e - h_1) \ \mathrm{d}\bar{\eta} - \lambda_1 \int_0^1 g(\bar{\xi}) \ \mathrm{d}\bar{\xi};$$

thus if we take

$$e = \int_\Omega f_1 \ \mathrm{d}\bar{\xi}\mathrm{d}\bar{\eta} + \lambda_1 \int_0^1 g(\bar{\xi}) \ \mathrm{d}\bar{\xi} + \int_0^1 h_1 \ \mathrm{d}\bar{\eta}, \tag{8.11}$$

(8.7) is solvable, and any two solutions differ only from a constant.

Now if (8.11) holds, then by Theorem 1.4 in [21] there exists a unique solution ϕ_1 to (8.7) with $\phi_1(0,0) = 0$, and furthermore the following estimate holds for $\alpha \in (0, \alpha_1)$ with a fixed $\alpha_1 \in (0,1)$:

$$\|\phi_1\|_{3+\alpha}^{(-1-\alpha)} \le C\Big(\|f_1\|_{1+\alpha}^{(1-\alpha)} + \|g\|_{C^{2+\alpha}[0,1]} + \|h_1\|_{2+\alpha}^{\prime(-\alpha)} \Big). \tag{8.12}$$

So finally by (8.4)–(8.6), (8.9), and (8.12) we get

$$\begin{aligned} &\big\|\bar{p} - p_b^+\big\|_{2+\alpha}^{(-\alpha)} + \|\bar{w}\|_{2+\alpha}^{(-\alpha)} + |e| \\ &\le C\Big(\|f_1\|_{1+\alpha}^{(1-\alpha)} + \|f_2\|_{1+\alpha}^{(1-\alpha)} + \|g\|_{C^{2+\alpha}[0,1]} + \|h_1\|_{2+\alpha}^{\prime(-\alpha)} \Big). \end{aligned} \tag{8.13}$$

Remark 8.1. An important observation is that the first equation in (8.1), which is responsible for the well-posedness or ill-posedness of the transonic shock problem under pressure giving on the exit, is in essence the equation of conservation of mass, as can be checked by tracing its origin. This fact can be seen more clearly by using another completely different method which was developed in [7].

8.2. "Linearized" problem (D2). This problem solves (recall that $U_0 = U|_{\bar{\xi}=0}$):

$$\begin{cases} \bar{u}_0 = u_b^+ + h_2(U, U_-), \\ \bar{p}_0 = p_b^+ + h_1(U, U_-), \\ \bar{\rho}_0 = \dfrac{\gamma}{\gamma - 1} \cdot \dfrac{\bar{p}_0}{c_0 - \frac{1}{2}\bar{u}_0^2(1 + \bar{w}_0^2)}, \\ \bar{\rho} = \bar{\rho}_0 \left(\frac{\bar{p}}{\bar{p}_0}\right)^{\frac{1}{\gamma}}, \\ \bar{u} = \left\{ \frac{2}{1+\bar{w}^2} \cdot \left(c_0 - \frac{\gamma}{\gamma - 1} \cdot \frac{\bar{p}}{\bar{\rho}} \right) \right\}^{\frac{1}{2}}. \end{cases} \tag{8.14}$$

9. Solution of problem (C). With the above preparations, we solve in this section problem (**D**) (i.e., problem (**C**)) by the Banach contraction mapping principle.

Set

$$\mathcal{O}_\delta := \Big\{U = (u, w, p, \rho)^t : \big\|u - u_b^+\big\|_{2+\alpha}^{(-\alpha)} + \|w\|_{2+\alpha}^{(-\alpha)} + \big\|p - p_b^+\big\|_{2+\alpha}^{(-\alpha)} + \big\|\rho - \rho_b^+\big\|_{2+\alpha}^{(-\alpha)} \le \delta\Big\}, \tag{9.1}$$

with

$$\delta < \delta_0 \tag{9.2}$$

and δ_0 a constant depending only on U_b such that our preparations in the preceding sections are valid. $\mathcal{O}_\delta$ is a closed subset of Banach space $(H_{2+\alpha}^{(-\alpha)})^4$. By Proposition 2.2 the norm here is in fact equivalent to that used in Theorem 2.6. We will construct a mapping $\mathcal{T}$ from $\mathcal{O}_\delta$ to $\mathcal{O}_\delta$ by problem (**D**) and show this mapping contracts when ε_0 (the perturbation of the wall of the duct) is small.

For any $U = (u, w, p, \rho)^t \in \mathcal{O}_\delta$, substitute

$$f_1 = f_1(U, \psi), \tag{9.3}$$

$$f_2 = f_2(U, \psi), \tag{9.4}$$

$$h_1 = h_1(U, U_-), \tag{9.5}$$

$$g(\bar\xi) = \Gamma'((1 - \psi(1))\bar\xi + \psi(1)) \tag{9.6}$$

in problem (**D1**), with $f_i(U, \psi)(i = 1, 2)$ as in (7.9), (7.10) and $h_1(U, U_-)$ defined by (7.3). By (2.18)–(2.21) and (7.1), (7.9), (7.10) we have the following estimates for $i = 1, 2$:

$$\|f_i(U, \psi)\|_{1+\alpha}^{(1-\alpha)} \le C(\delta^2 + \delta\sigma), \tag{9.7}$$

$$\|h_i(U, U_-)\|_{2+\alpha}^{'(-\alpha)} \le C(\delta^2 + \varepsilon). \tag{9.8}$$

$$\|g\|_{C^{2,\alpha}[0,1]} \le C\varepsilon. \tag{9.9}$$

We may get unique $\bar p, \bar w, e$ from problem (**D1**), and by (8.13) we have

$$\big\|\bar p - p_b^+\big\|_{2+\alpha}^{(-\alpha)} + \|\bar w\|_{2+\alpha}^{(-\alpha)} + |e| \le C(\delta^2 + \delta\sigma + \varepsilon). \tag{9.10}$$

Now consider problem (**D2**). By (8.14) and analyticity of each expression, we easily obtain that

$$\big\|\bar u - u_b^+\big\|_{2+\alpha}^{(-\alpha)} \le C(\delta^2 + \delta\sigma + \varepsilon), \tag{9.11}$$

$$\big\|\bar\rho - \rho_b^+\big\|_{2+\alpha}^{(-\alpha)} \le C(\delta^2 + \delta\sigma + \varepsilon). \tag{9.12}$$

So far we obtained the unique $\bar U := (\bar p, \bar u, \bar w, \bar\rho)^t$ from $U = (p, u, w, \rho)^t \in \mathcal{O}_\delta$ and have the estimate

$$\big\|\bar p - p_b^+\big\|_{2+\alpha}^{(-\alpha)} + \|\bar w\|_{2+\alpha}^{(-\alpha)} + \big\|\bar u - u_b^+\big\|_{2+\alpha}^{(-\alpha)} + \big\|\bar\rho - \rho_b^+\big\|_{2+\alpha}^{(-\alpha)} + |e| \le C(\delta^2 + \delta\sigma + \varepsilon). \tag{9.13}$$

Now choosing ε_0, σ_0 such that

$$C\delta_0 \le \frac{1}{4}, \tag{9.14}$$

$$C\sigma_0 \le \frac{1}{4}, \tag{9.15}$$

and

$$\delta = 2C\varepsilon, \tag{9.16}$$

we get

$$\|\bar{p} - p_b^+\|_{2+\alpha}^{(-\alpha)} + \|\bar{w}\|_{2+\alpha}^{(-\alpha)} + \|\bar{u} - u_b^+\|_{2+\alpha}^{(-\alpha)} + \|\bar{\rho} - \rho_b^+\|_{2+\alpha}^{(-\alpha)} + |e| \le \delta. \tag{9.17}$$

Thus the mapping $\mathcal{T} : U = (u, w, p, \rho)^t \mapsto \bar{U} = (\bar{u}, \bar{w}, \bar{p}, \bar{\rho})^t$ maps $\mathcal{O}_\delta$ into $\mathcal{O}_\delta$.

What is left is to show the contraction of $\mathcal{T}$. For $i = 1, 2$, suppose

$$\psi^{(i)} \in \mathcal{S}_\sigma,$$

$$U^{(i)} = (u^{(i)}, w^{(i)}, p^{(i)}, \rho^{(i)})^t \in \mathcal{O}_\delta,$$

and denote

$$\bar{U}^{(i)} = (\bar{u}^{(i)}, \bar{w}^{(i)}, \bar{p}^{(i)}, \bar{\rho}^{(i)})^t = \mathcal{T}((u^{(i)}, w^{(i)}, p^{(i)}, \rho^{(i)})^t),$$

$$U_-^{(i)} = U_-(\psi^{(i)}(\bar{\eta}), \bar{\eta}), \tag{9.18}$$

$$g^{(i)} = \Gamma'((1 - \psi^{(i)}(1))\bar{\xi} + \psi^{(i)}(1)). \tag{9.19}$$

Recall that $\|U_- - U_b^-\|_{C^{3,\alpha}(\mathbb{P})} \le C_0\varepsilon$; we get

$$\left\|U_-^{(1)} - U_-^{(2)}\right\|_{2+\alpha}^{'(-\alpha)} \le C\varepsilon \left\|\psi^{(1)} - \psi^{(2)}\right\|_{2+\alpha}^{'(-\alpha)} \le C\varepsilon \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)}.$$

Then direct calculation shows that for $j = 1, 2$,

$$\begin{aligned} &\left\|f_j(U^{(1)}, \psi^{(1)}) - f_j(U^{(2)}, \psi^{(2)})\right\|_{1+\alpha}^{(1-\alpha)} \\ &\le C(\delta + \sigma)\Big(\left\|U^{(1)} - U^{(2)}\right\|_{2+\alpha}^{(-\alpha)} + \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)} \Big), \end{aligned} \tag{9.20}$$

$$\begin{aligned} &\left\|h_j(U^{(1)}, U_-^{(1)}) - h_j(U^{(2)}, U_-^{(2)})\right\|_{2+\alpha}^{'(-\alpha)} \\ &\le C(\delta + \varepsilon)\Big(\left\|U^{(1)} - U^{(2)}\right\|_{2+\alpha}^{(-\alpha)} + \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)} \Big). \end{aligned} \tag{9.21}$$

$$\left\|g^{(1)} - g^{(2)}\right\|_{C^{2,\alpha}[0,1]} \le C\varepsilon \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)}. \tag{9.22}$$

Thus by solving corresponding problem (**D**) we get

$$\begin{aligned} &\left\|\bar{U}^{(1)} - \bar{U}^{(2)}\right\|_{2+\alpha}^{(-\alpha)} \\ &\le C(\delta + \sigma + \varepsilon)\Big(\left\|U^{(1)} - U^{(2)}\right\|_{2+\alpha}^{(-\alpha)} + \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)} \Big). \end{aligned} \tag{9.23}$$

For $\psi^{(1)} = \psi^{(2)} = \psi$, we obtain contraction of $\mathcal{T}$ by choosing $\delta_0, \varepsilon_0, \sigma_0$ such that

$$C(\varepsilon_0 + \delta_0 + \sigma_0) < \frac{1}{2}. \tag{9.24}$$

This solves problem ($\mathbf{C}_\psi$) for any $\psi \in \mathcal{S}_\sigma$ with σ_0 small.

10. Solution of problem (B): Determination of shock front. For any $\psi \in \mathcal{S}_\sigma$, we have solved problem $(\mathbf{C}_\psi)$ to obtain the unique solution U_ψ. Now by solving the Cauchy problem

$$\begin{cases} \tilde{\psi}'(\eta) = \frac{[v_\psi]}{[p_\psi]}, \\ \tilde{\psi}(0) = 0, \end{cases} \tag{10.1}$$

we constructed a mapping by $\Psi(\psi) = \tilde{\psi}$.

Obviously we have

$$\left\|\tilde{\psi}\right\|_{3+\alpha}^{'(-1-\alpha)} \le C(\delta+\varepsilon) \le C\varepsilon \tag{10.2}$$

by using (9.16). Taking

$$\sigma = C\varepsilon, \tag{10.3}$$

we then have $\Psi : \mathcal{S}_\sigma \mapsto \mathcal{S}_\sigma$ provided ε_0 is sufficiently small.

For $\psi^{(i)} \in \mathcal{S}_\sigma$, $i = 1, 2$, denote $U^{(i)}$ as $U_{\psi^{(i)}}$; then by (9.23) (recall now that $\bar{U}^{(i)} = U^{(i)}$) one gets

$$\left\|U^{(1)} - U^{(2)}\right\|_{2+\alpha}^{(-\alpha)} \le C(\delta+\sigma+\varepsilon) \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)}; \tag{10.4}$$

thus

$$\begin{aligned} &\left\|\tilde{\psi}^{(1)} - \tilde{\psi}^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)} \\ &\le C \left\|U^{(1)} - U^{(2)}\right\|_{2+\alpha}^{(-\alpha)} + C\varepsilon \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)} \\ &\le C(\delta+\sigma+\varepsilon) \left\|\psi^{(1)} - \psi^{(2)}\right\|_{3+\alpha}^{'(-1-\alpha)}. \end{aligned} \tag{10.5}$$

If (9.24) holds, Ψ contracts on $\mathcal{S}_\sigma$. By the Banach contraction mapping principle we know the free boundary problem $(\mathbf{B})$ is uniquely solvable. Combining this result and Theorem 5.1, we proved Theorem 2.6.

Remark 10.1. We explain here further why the uniqueness claimed in Theorem 2.6 holds. By (10.5) we see the fixed point, i.e., the transonic shock front $\bar{\psi}$ is unique. Uniqueness for nonlinear problem $(\mathbf{D}_{\bar{\psi}})$ in section 9 (thus problem $(\mathbf{C}_{\bar{\psi}})$) follows from the contraction argument of the mapping $\mathcal{T}$ defined on the line after estimate (9.17) (consult (9.23) with the case $\psi^{(1)} = \psi^{(2)} = \bar{\psi}$). Moreover, uniqueness of constant e for nonlinear problem $(\mathbf{D}_{\bar{\psi}})$ follows by writing the nonlinear problem as a linear problem (8.1) with right-hand sides defined by (9.3)–(9.6), and then using uniqueness of e for linear problem (8.1). Thus the subsonic flow is also unique.

11. Discussion on well-posedness or ill-posedness for other conditions given at the exit of duct. In this section we discuss the well-posedness or ill-posedness of transonic shock problem $(\mathbf{A})$ if other conditions are given at the exit of the duct. Our tool is the well-posedness or ill-posedness of the corresponding boundary value problem for elliptic systems (8.1). We remark that all the results listed below can be proved rigorously in the same fashion as we have done for giving

pressure at the exit before. We just sketch out the key points that are needed for such proofs.

1. As have been shown above (in section 8), given p at Γ_1 corresponds to the Neumann problem for Poisson equations. So in general for given pressure at the exit, problem (**A**) is ill-posed.

2. However, if we give w at the exit, then we need to solve a mixed boundary value problem for Poisson equations with Dirichlet data nonempty. It is well known that such a problem is well-posed; i.e., we can give w at the exit in an arbitrary way and also obtain a unique solution (although we require that $w|_{\Gamma_1}$ should be small since we are dealing with a small perturbation problem). So for w given at the exit, problem (**A**) is well-posed.

3. For arbitrarily given $v = g$ at the exit, Problem (**A**) is still well-posed. In fact, this corresponds to given

$$w = \frac{g}{u}$$

for any $U = (u, v, p, \rho)^t \in \mathcal{O}_\delta$ in the linearized problem (**D1**). The linearized problem (**D1**) is then uniquely solvable, and the contraction by

$$w^{(1)} - w^{(2)} = g\frac{u^{(1)} - u^{(2)}}{u^{(1)}u^{(2)}}$$

is still available, since g is small.

4. For given entropy $S = g$ at the exit, problem (**A**) is ill-posed. In fact, by constancy of entropy along streamlines behind the shock front, we get

$$\frac{p_0}{\rho_0^\gamma} = g.$$

Now by the third equation in (7.13), one has

$$w_0 = \left\{ 2\frac{c_0 - p_0\frac{\gamma}{\gamma-1}\left(\frac{g}{p_0}\right)^{\frac{1}{\gamma}}}{u_0^2} \right\}^{\frac{1}{2}}.$$

Since p_0, u_0 are known, we obtain both p_0 and w_0 on $\bar{\xi} = 0$. This means we encounter an initial-boundary value problem for the elliptic system (i.e., the first two equations in (8.1) or (7.12)). It is well known that such problems are ill-posed.

5. For giving Mach number $M = g$ at the exit, Problem (**A**) is ill-posed. Indeed, by Bernoulli's law (2.3) we have

$$\gamma\frac{p}{\rho} = a^2 = \frac{c_0}{\frac{1}{\gamma-1} + \frac{1}{2}g^2}.$$

From constancy of entropy we get

$$\frac{p}{\rho^\gamma} = \frac{p_0}{\rho_0^\gamma},$$

where ρ_0 can be expressed by the third equation in (7.13). Thus we can solve the above two equations to obtain

$$p = p_0 h,$$

where

$$h = \left\{ \frac{c_0}{(c_0 - \frac{1}{2}u_0^2(1+w_0^2))(1+\frac{1}{2}(\gamma-1)g^2)} \right\}^{\frac{\gamma}{\gamma-1}}.$$

Thus we get

$$p - p_b^+ = (p_0 - p_b^+)h + p_b^+(h-1).$$

Notice that

$$h(U_b^+) = 1$$

and

$$\|h-1\| \le C(\|u_0 - u_b^+\| + \|g - M_b^+\| + \|w_0\|^2).$$

(Note here that w_0 is itself small and that fortunately w_0^2 appears in the expression of h. This observation is of crucial importance!) With such an estimate in hand, the remaining work is the same as giving pressure at the exit. Thus problem (**A**) is not well-posed for the given arbitrary Mach number at the exit of the duct.

6. For given arbitrary density $\rho = g$ at the exit, problem (**A**) is not well-posed. From constancy of entropy we have

$$p = \frac{p_0 g^\gamma}{\rho_0^\gamma}.$$

Due to the third equation in (7.13), we get

$$p = \left(\frac{\gamma}{\gamma-1}\right)^\gamma p_0^{1-\gamma} g^\gamma \left(c_0 - \frac{1}{2}u_0^2(1+w_0^2)\right)^\gamma.$$

The left analysis is the same as in **5**.

7. For given $u = g$ at the exit, problem (**A**) is still ill-posed. By Bernoulli's law we get

$$\frac{p}{\rho} = \frac{\gamma-1}{\gamma}\Big(c_0 - \frac{1}{2}g^2(1+w_1^2)\Big);$$

here w_1 is the value of w restricted on $\bar{\xi} = 1$. Using constancy of entropy we also have

$$\frac{p}{\rho^\gamma} = \frac{p_0}{\rho_0^\gamma}.$$

By the expressions of ρ_0 (the third equation in (7.13)) we can solve from the above two equations that

$$p = p_0 \left(\frac{c_0 - \frac{1}{2}g^2(1+w_1^2)}{c_0 - \frac{1}{2}u_0^2(1+w_0^2)}\right)^{\frac{\gamma}{\gamma-1}}.$$

Note w_0, w_1 are small quantities and appeared as squares in the above expressions; u_0, p_0 are known and are second order terms as shown in section 9. Thus we can use similar methods as above to show the ill-posedness of problem (**A**).

Remark 11.1. All the above results may be surprising at first glance. However, a deep consideration of the background solution suggests these results are natural, since for the one-dimensional case, S, p, ρ, M, u have already been completely determined.

Acknowledgments. The author would like to thank Shuxing Chen, Yongqian Zhang, and Beixiang Fang for their generous help in doing this work.

REFERENCES

[1] S. Canic, B. Kerfitz, and G. Lieberman, *A proof of existence of perturbed steady transonic shocks via a free boundary problem*, Comm. Pure Appl. Math., 53 (2000), pp. 484–511.
[2] S. Canic, B. Kerfitz, and E. H. Kim, *A free boundary problems for unsteady transonic small disturbance equation: Transonic regular reflection*, Methods Appl. Anal., 7 (2000), pp. 313–336.
[3] S. Canic, B. Kerfitz, and E. H. Kim, *Free boundary problems for a quasilinear degenerate elliptic equation: Transonic regular reflection*, Comm. Pure Appl. Math., 55 (2002), pp. 71–92.
[4] G.-Q. Chen and M. Feldman, *Multidimensional transonic shocks and free boundary problems for nonlinear equations of mixed type*, J. Amer. Math. Soc., 16 (2003), pp. 461–494.
[5] G.-Q. Chen and M. Feldman, *Steady Transonic Shocks and Free Boundary Problems in Infinite Cylinders for the Euler Equations*, Comm. Pure Appl. Math., 57 (2004), pp. 310–356.
[6] G.-Q. Chen and M. Feldman, *Existence and stability of multidimensional transonic flows through an infinite nozzle of arbitrary cross-sections*, Arch. Ration. Mech. Anal., to appear.
[7] S. Chen and H. Yuan, *Transonic shocks in compressible flow passing a duct for three-dimensional Euler systems*, Arch. Ration. Mech. Anal., submitted.
[8] S. Chen, *Stability of transonic shock fronts in two-dimensional Euler systems*, Trans. Amer. Math. Soc., 357 (2005), pp. 287–308.
[9] S. Chen, *A free boundary problem of elliptic equation arising in supersonic flow past a conical body*, Z. Angew. Math. Phys., 54 (2003), pp. 1–23.
[10] S. Chen, *How does a shock in supersonic flow grow out of smooth data?*, J. Math. Phys., 42 (2001), pp. 1154–1172.
[11] S. Chen, *Stability of Mach Configuration*, Comm. Pure Appl. Math., 59 (2006), pp. 1–35.
[12] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience Publishers Inc., New York, 1948.
[13] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, Heiderberg, New York, 2000.
[14] B. Fang, *Stability of transonic shocks for the full Euler equations in supersonic flow past a wedge*, Math. Methods Appl. Sci., 29 (2006), pp. 1–26.
[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 224, Springer, Berlin, New York, 1983.
[16] D. Gilbarg and L. Hörmander, *Intermediate Schauder Estimate*, Arch. Rational Mech. Anal., 74 (1980), pp. 297–318.
[17] H. M. Glaz and T.-P. Liu, *The asymptotic analysis of wave interactions and numerical calculations of transonic nozzle flow*, Adv. Appl. Math., 5 (1984), pp. 111–146.
[18] J. Glimm, G. Marshall, and B. Plohr, *A generalized Riemann problem for quasi-one-dimensioanl gas flows*, Adv. Appl. Math., 5 (1984), pp. 1–30.
[19] A.G. Kuz'min, *Boundary-Value Problems for Transonic Flow*, John Wiley & Sons, New York, 2002.
[20] T.-T. Li and W.-C. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke Univ. Math. Ser. 5, Duke University, Mathematics Department, Durham, NC, 1985.
[21] G. Lieberman, *Oblique derivative problems in Lipschitz domains* II*: Discontinuous boundary data*, J. Reine Angew. Math., 389 (1988), pp. 1–21.
[22] T.-P. Liu, *Transonic gas flows in a variable area duct*, Arch. Rational Mech. Anal., 80 (1982), pp. 1–18.
[23] T.-P. Liu, *Nonlinear stability and instability of transonic gas flow through a nozzle*, Comm. Math. Phys., 83 (1982), pp. 243–260.
[24] H. Ockendon and J. R. Ockendon, *Waves and Compressible Flows*, Springer-Verlag, New York, 2004.
[25] A. Pope and K. L. Goin, *High speed wind tunnel testing*, John Wiley & Sons, New York, 1965.
[26] D. H. Smith, *Non-uniqueness and multi-shock solutions for transonic nozzle flows*, IMA J. Appl. Math., 71 (2006), pp. 120–132.
[27] B. Whitham, *Linear and nonlinear waves*, John Wiley & Sons, New York, 1974.
[28] Z. Xin and H. Yin, *Transonic shock in a nozzle* I: *Two dimensional case*, Comm. Pure. Appl. Math., 58 (2005), pp. 999–1050.

[29] H. Yuan, *Transonic shocks for steady Euler flows with cylindrical symmetry*, Nonlinear Anal., to appear.
[30] H. Yuan, *A note on mixed boundary value problem for Laplace equation*, Acta Math. Sin. Ser. A, 46 (2003), pp. 1091–1096 (in Chinese).
[31] M. J. Zucrow and J. D. Hoffman, *Gas Dynamics*, Vol. 1, John Wiley & Sons, New York, 1976.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1371–1388

CHEMICAL KINETICS ON SURFACES: A SINGULAR LIMIT OF A REACTION-DIFFUSION SYSTEM*

G. FIBICH†, I. GANNOT‡, A. HAMMER§, AND S. SCHOCHET†

Abstract. We show that chemical kinetics relations can be used to describe processes that involve binding and dissociation reactions that take place on surfaces. From a mathematical perspective, the problem we study is a singular limit of a reaction-diffusion system in which one of the variables concentrates on a lower-dimensional set in the limit, while the other continues to diffuse in a fixed domain.

Key words. chemical kinetics, surface, binding, dissociation, singular limit, invariant region

AMS subject classifications. 92C45, 92C50, 35K57, 35B25

DOI. 10.1137/050633767

1. Introduction.

Numerous biological processes involve binding and dissociation reactions that take place on surfaces. For example, in *antibody-antigen* interactions, antibodies immobilize and agglutinate infectious agents by binding to specific receptors located on the surface of antigens [1, 19, 22]. Additional examples include the binding of proteins to cell membranes either to initiate transduction of external signals into the cell (*signal transduction*) or to open the ion channels of the membrane (see, e.g., [18]); the binding of microbiological cultures to attachment sites on the inner walls of flow reactors [12]; and the phenomenon of surface plasmon resonance, which involves interactions of biopolymers with various ligands [13].

A natural way to model surface reactions is to adapt the standard chemical-kinetics approach used for reactions occurring in volumes. This means that the binding rate for surface reactions is assumed to be proportional to the product of the volumetric concentration of the reactant at the surface and the surface concentration of the binding sites [18, 20]. There is a methodological problem with this approach, however, since chemical-kinetics relations are usually derived under the assumption that reactions take place in a volume, in which the two reactants are well mixed.

Our goal here is to justify the use of chemical-kinetics relations for reactions that take place on surfaces. To do so, we will first construct a volumetric model in which the binding sites, and hence also the binding and dissociation reactions, take place in a narrow volumetric layer around the surface. We will then show that as the width of the binding sites layer shrinks to zero, the volumetric model reduces to a surface model, in which binding sites are located on the surface, and for which the reactions are still described by chemical-kinetics relations.

From a mathematical perspective, the problem we study is a singular limit of a

*Received by the editors June 16, 2005; accepted for publication (in revised form) July 5, 2006; published electronically December 26, 2006.
http://www.siam.org/journals/sima/38-5/63376.html

†School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (fibich@math.tau.ac.il, schochet@post.tau.ac.il).

‡Department of Biomedical Engineering, Faculty of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel (gannot@eng.tau.ac.il) and Program of Biomedical Engineering, Department of Electrical and Computer Engineering, School of Engineering and Applied Sciences, George Washington University, Washington, DC 20052 (gannoti@mail.nih.gov).

§Department of Biomedical Engineering, Faculty of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel (amit_hammer@hotmail.com).

reaction-diffusion system. The problem here differs from the widely studied case of thin domains (see, e.g., [10]) in that here only one of the variables concentrates on a lower-dimensional set in the limit, while the other continues to diffuse in a fixed domain.

In this study we analyze the problem of chemical kinetics on surfaces in the context of a mathematical model for a novel in-vivo imaging technique for identifying and locating cancerous tumors, which are sphere-like, isolated, three-dimensional objects, with a smooth boundary.[1] This model consists of a diffusion equation in the volume surrounding the tumor, and binding and dissociation reactions that take place on the tumor surface. We note, however, that the methodology developed in this study, namely, the study of "surface models" as the limit of "volumetric models," can be applied to other models that involve binding and dissociation reactions on surfaces. Moreover, our approach may also be relevant to problems in combustion (e.g., burning of coal) in which the reactions take place between one or more species (e.g., coal) that are confined to narrow regions near the reactor boundaries and other species that are free to diffuse over larger domains (e.g., oxygen) [2].

The paper is organized as follows. In section 2 we construct two mathematical models, a surface model and a volumetric model. In section 3 we present a heuristic derivation of the surface model from the volumetric model. This is done by taking the limit as the width of the volumetric layer tends to zero while assuming that the limits of the concentrations exist in appropriate senses. For further clarity, we consider only the radially symmetric case. In section 4 we rigorously prove that solutions of the volumetric model converge to those of the surface model, without assuming radial symmetry. Finally, in section 5 we comment briefly on the possibility of extending our results to other models.

2. Mathematical models.

2.1. Fluorophore-antibody imaging. Our interest in this problem originated from the need to model a novel in-vivo imaging technique for identifying and locating cancerous tumors [11]. This method is based on one of the immune system responses to tumors, which is the concentration of white blood cells, known as T cells, around the tumor. These T cells have receptors which are specific to some antibodies of the immune system. The imaging technique involves selecting an antibody with high specificity to T cells [3], artificially conjugating it with a fluorescent marker compound, and injecting the fluorescenated antibodies into the suspected tumor area [5, 6, 7, 8, 9]. After some time, the fluorescenated antibodies, hereinafter denoted *markers*, will diffuse away from the tumor area, except for those that are bound to the T cell receptors around the tumor. Hence, when an external laser excitation is applied, the fluorescence of the markers indicates the location of the tumor.

The mathematical model that we use to describe the method of fluorophore-antibody imaging involves diffusion of markers in the tissue, binding of markers to T cells receptors (binding sites), and dissociation of markers that are already bound to sites. The methodology developed in this study can, however, be extended to more complex models that allow for diffusion of markers into the tissue area, advection effects, etc.

To simplify the presentation, in this section we assume that the tumor is the radially symmetric ball $0 \le r < r_{tumor}$, where r is the radial distance from the tumor

[1] This corresponds to the common solid tumors such as breast, lung, and sarcoma, at the early stages of the tumor (i.e., before it develops a nonsmooth surface).

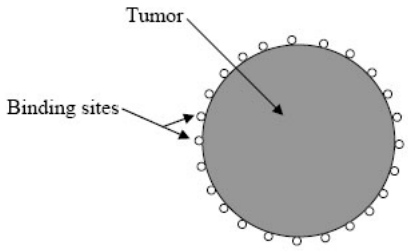


FIG. 1. *The surface model. Binding sites are located on the boundary of the tumor.*

center. The assumption of radial symmetry is reasonable for young, small-size tumors. Our results, however,are also valid for more advanced tumors, which are half-spheres or isolated, three-dimensional lumps, so long that their boundary remains smooth (see section 4). For simplicity we also assume that the initial markers distribution is radially symmetric.

2.2. The surface model. We first develop a surface model in which the T cell binding sites are located on the tumor surface (see Figure 1). In the tissue area, the motion of the markers is governed by the diffusion equation

$$\frac{\partial M(r,t)}{\partial t} = D\Delta M(r,t), \qquad 0 < t, \quad r_{tumor} < r < \infty, \tag{1}$$

where $M(r,t)$ is the volumetric concentration of the free (i.e., unbound) markers, and D is the diffusion coefficient of markers in the tissue. The initial condition for (1) is

$$M(r,0) = M_0(r), \tag{2}$$

where $M_0(r)$ is the initial concentration of markers.

Let us assume that chemical-kinetics relations can be used to model the reactions that take place on a surface. Then the free sites concentration at the tumor surface ($r = r_{tumor}$) is governed by the equation

$$\frac{\partial S(t)}{\partial t} = k_d^{sur}[S_{tot} - S(t)] - k_b^{sur} M(r_{tumor}, t) S(t), \tag{3}$$

where $S(t)$ denotes the free sites concentration, k_b^{sur} and k_d^{sur} are the binding and dissociation rate constants, respectively, and S_{tot} is the total concentration of sites, both free and occupied. Equation (3) shows that the free sites concentration increases as a result of dissociation of marker-site complexes and decreases as a result of binding of free markers to binding sites. The dissociation rate is linearly proportional to the bound sites concentration, $[S_{tot} - S(t)]$. Under the assumption of chemical kinetics, the binding rate is linearly proportional to the concentration of free binding sites, $S(t)$, and also linearly proportional to the concentration of free markers on the tumor boundary, $M(r_{tumor}, t)$.

We assume that before the injection of markers (i.e., at $t = 0$), all binding sites are unoccupied, so the initial condition for the sites equation is given by the total sites concentration S_{tot}, i.e.,

$$S(0) = S_{tot}. \tag{4}$$

We now derive the boundary conditions for the markers at the tumor boundary. Markers may be either free (i.e., in the tissue area) or bound (to T cell binding sites). If we assume that a single marker binds to a single site, then the concentration of bound markers on the surface equals $S_{tot} - S(t)$. In view of the fact that the total area of the tumor surface is $4\pi r_{tumor}^2$, the total number of bound markers is then $4\pi r_{tumor}^2 [S_{tot} - S(t)]$. Thus, the global conservation of markers is given by

$$\underset{\text{bound markers}}{4\pi r_{tumor}^2 [S_{tot} - S(t)]} + \underset{\text{free markers}}{4\pi \int_{r_{tumor}}^{\infty} M(r,t)\, r^2 dr} = \underset{\text{total markers}}{4\pi \int_{r_{tumor}}^{\infty} M_0(r)\, r^2 dr}.$$

Differentiating this equation with respect to t, and using (1) and the formula for Δ in polar coordinates, shows that

$$\begin{aligned} r_{tumor}^2 \frac{\partial S(t)}{\partial t} &= \int_{r_{tumor}}^{\infty} \frac{\partial M(r,t)}{\partial t}\, r^2 dr \\ &= D \int_{r_{tumor}}^{\infty} \Delta M(r,t)\, r^2 dr = D \int_{r_{tumor}}^{\infty} \left[\frac{\partial^2 M}{\partial r^2} + \frac{2}{r} \frac{\partial M}{\partial r} \right] r^2 dr \\ &= D \int_{r_{tumor}}^{\infty} \frac{\partial \left[r^2 \frac{\partial M}{\partial r} \right]}{\partial r} dr = -D r_{tumor}^2 \frac{\partial M}{\partial r}(r_{tumor}, t), \end{aligned} \tag{5}$$

provided that the concentration of markers decays sufficiently rapidly at large distances so that they have no flux at infinity. Upon substituting in the value of $\frac{\partial S(t)}{\partial t}$ from (3) we find that the boundary condition at the tumor surface is given by

$$\frac{\partial M}{\partial r}(r_{tumor}, t) = \frac{1}{D} \{ k_b^{sur} M(r_{tumor}, t) S(t) - k_d^{sur} [S_{tot} - S(t)] \}. \tag{6}$$

2.3. The volumetric model. As mentioned in the introduction, there is a methodological problem with the surface model since we used chemical-kinetics relations to model surface reactions; see (3). In order to avoid this problem, we now adopt a different approach and assume that the T cell binding sites are located in a narrow volumetric layer around the tumor (see Figure 2). This approach also has a physiological justification. Indeed, data collected in histological staining experiments show that T cells are not located strictly on the tumor surface but rather in a thin volumetric layer around the tumor (see Figure 3). The existence of a layer of T-lymphocytes (CD3 positive cells) around the tumor was also reported in, e.g., [4, 25].

Let ε denote the width of the volumetric layer in which binding sites are located. Then the volume density $S_{\text{tot}}^{\varepsilon}(r)$ of total binding sites vanishes identically for $r > r_{tumor} + \varepsilon$, i.e.,

$$S_{\text{tot}}^{\varepsilon}(r) \equiv 0 \qquad \text{for } r > r_{tumor} + \varepsilon. \tag{7}$$

Of course this implies that the density $S^{\varepsilon}(t,r)$ of free binding sites also vanishes for $r > r_{tumor} + \varepsilon$.

The equation of evolution for the concentration $M^{\varepsilon}(t,r)$ of free markers now takes the form

$$\begin{aligned} \frac{\partial M^{\varepsilon}(r,t)}{\partial t} &= D \Delta M^{\varepsilon}(r,t) + k_d^{vol} [S_{\text{tot}}^{\varepsilon}(r) - S^{\varepsilon}(r,t)] - k_b^{vol} M^{\varepsilon}(r,t) S^{\varepsilon}(r,t), \\ & 0 < t, \quad r_{tumor} < r < \infty, \end{aligned} \tag{8}$$

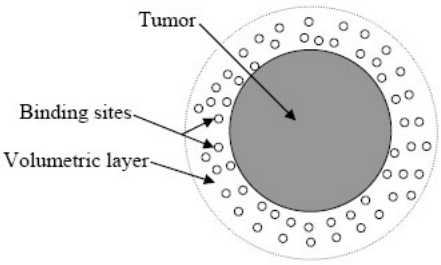


FIG. 2. *The volumetric model. Binding sites are located in a thin volumetric layer around the tumor.*

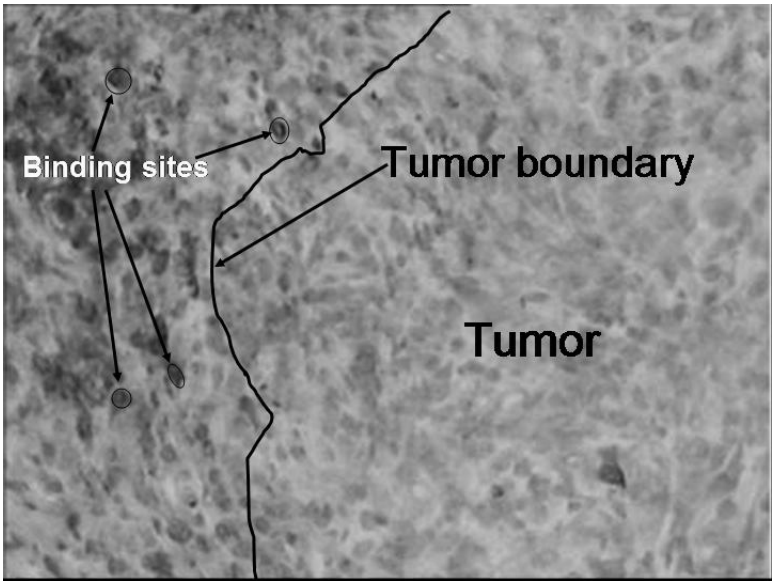


FIG. 3. *Histological staining of a 5-day old tumor squamous cell carcinoma in the oral cavity (mag. X200). Binding sites (stained in black) can be seen to the left of the solid line. Figure supplied by Dr. Gallya Gannot, National Institutes of Health, Bethesda, MD.*

where k_b^{vol} and k_d^{vol} are the volumetric model binding and dissociation rate constants, respectively. The first term on the right is the diffusion term, the second term describes the dissociation of marker-site complexes, and the third term describes the creation of these complexes. Note that in contrast to (3), there is no problem with using chemical-kinetics relations.

The rate of change of the free sites concentration is derived using chemical-kinetics relations, as was done for (3) of the surface model, yielding

$$\frac{\partial S^{\varepsilon}(r,t)}{\partial t} = k_d^{vol}\left[S_{\text{tot}}^{\varepsilon}(r) - S^{\varepsilon}(r,t)\right] - k_b^{vol} M^{\varepsilon}(r,t) S^{\varepsilon}(r,t). \tag{9}$$

The only difference is that now these reactions take place in a volumetric layer rather than on a surface as in the previous model.

As in the surface model, we assume that at time $t = 0$ all the binding sites are unoccupied. Therefore, the initial condition for the free sites concentration is given

by

$$S^\varepsilon(r,0) = S^\varepsilon_{\text{tot}}(r). \tag{10}$$

The initial distribution of markers will be assumed to be the same as for the surface model:

$$M^\varepsilon(r,0) = M_0(r), \tag{11}$$

independently of ε. We also assume that markers cannot diffuse into the tumor. Therefore, at the tumor surface we impose the no-flux boundary condition

$$\frac{\partial M^\varepsilon(r,t)}{\partial r}(r_{tumor},t) = 0. \tag{12}$$

3. Heuristic justification of the surface model from the volumetric model. We now show, under some scientifically natural assumptions, that as the width ε of the binding sites layer goes to zero the volumetric model reduces to the surface model.

As already mentioned, we assume that both models have the same initial distribution of markers, and that in both models all the sites are unoccupied at time $t = 0$. We also assume that both models have the same total number of sites, i.e.,

$$\underbrace{4\pi \int_{r_{tumor}}^{\infty} S^\varepsilon_{\text{tot}}(r)\, r^2\, dr}_{\substack{\text{total number of sites} \\ \text{for the volumetric model}}} = \underbrace{4\pi r^2_{tumor} S_{tot}}_{\substack{\text{total number of sites} \\ \text{for the surface model}}}. \tag{13}$$

Next, we assume that the solution $M^\varepsilon(r,t)$, $S^\varepsilon(r,t)$ of the volumetric model (7)–(12) and the solution $M(r,t)$, $S(t)$ of the surface model (1)–(4), (6) exist for all positive time and are unique.

Finally, we assume that as ε tends to zero both the concentration $M^\varepsilon(r,t)$ of markers in the volumetric model and the radial integral of the concentration $S^\varepsilon(r,t)$ of free sites in that model tend to definite values. In other words, we assume that

$$M^0(r,t) := \lim_{\varepsilon\to 0} M^\varepsilon(r,t) \quad \text{and} \quad \mathcal{S}^0(t) := \frac{1}{r^2_{tumor}} \lim_{\varepsilon\to 0} \int_{r_{tumor}}^{\infty} S^\varepsilon(r,t)\, r^2\, dr \tag{14}$$

exist. Furthermore, we assume that in the region $r > r_{tumor}$ those equations may be differentiated and integrated as often as needed, and that the order of the resulting derivatives, integrals, and limits may be freely interchanged.

PROPOSITION 3.1. *Under the above assumptions, the limits* (14) *of the volumetric model are the solution of the surface model with the same binding and dissociation constants. In other words,*

$$M^0(r,t) \equiv M(r,t), \tag{15}$$

and

$$\mathcal{S}^0(t) \equiv S(t), \tag{16}$$

the latter of which may also be expressed as

$$\lim_{\varepsilon\to 0} S^\varepsilon(r,t) \equiv S(t)\cdot\delta(r - r_{tumor}),$$

where δ is the Dirac δ function.

Proof. Since $S^{\varepsilon}_{\text{tot}}(r)$ is nonnegative, (7) plus (13) implies that

$$\lim_{\varepsilon\to 0} S^{\varepsilon}_{\text{tot}}(r) = S_{tot}\cdot\delta(r - r_{tumor}). \tag{17}$$

Similarly, since

$$0 \le S^{\varepsilon}(r,t\,) \le S^{\varepsilon}_{\text{tot}}(r), \tag{18}$$

it follows from (7) plus (14) that

$$\lim_{\varepsilon\to 0} S^{\varepsilon}(r,t) = \mathcal{S}^0(t)\cdot\delta(r - r_{tumor}), \tag{19}$$

and (17)–(19) imply further that

$$0 \le \mathcal{S}^0(t) \le S_{tot}.$$

We now show that for $r > r_{tumor}$ and $t \ge 0$, (15)–(16) hold, by showing that M^0 and $\mathcal{S}^0$ satisfy the surface model equations (1)–(4), (6).

Integrating both sides of (9) and taking the limit as $\varepsilon \to 0$ gives

$$\begin{aligned}
&\frac{1}{r^2_{tumor}}\lim_{\varepsilon\to 0}\int_{r_{tumor}}^{\infty}\frac{\partial S^{\varepsilon}(r,t)}{\partial t}\; r^2\,dr\\
&= \frac{1}{r^2_{tumor}}\lim_{\varepsilon\to 0}\int_{r_{tumor}}^{\infty} k_d^{vol}\left[S^{\varepsilon}_{\text{tot}}(r) - S^{\varepsilon}(r,t)\right]\; r^2\,dr\\
&\quad -\frac{1}{r^2_{tumor}}\lim_{\varepsilon\to 0}\int_{r_{tumor}}^{\infty} k_b^{vol} M^{\varepsilon}(r,t)S^{\varepsilon}(r,t)\; r^2\,dr.
\end{aligned}$$

Combining the above with (14), (17), and (19) gives

$$\mathcal{S}^0_t = k_d^{vol}\left[S_{tot} - \mathcal{S}^0\right] - k_b^{vol} M^0(r_{tumor},t)\mathcal{S}^0, \tag{20}$$

which corresponds to (3).

Similarly, in light of (14), (17), and (19) the limit $\varepsilon \to 0$ of the volumetric markers equation (8) yields

$$\frac{\partial M^0(r,t)}{\partial t} = D\Delta M^0(r,t) \qquad \text{for } r > r_{tumor}, \tag{21}$$

which is the analogue of (1).

From (10), (13), and (14) it follows that

$$\mathcal{S}^0(0) = S_{tot}.$$

More simply, (11) plus (14) implies

$$M^0(r,0) = M_0(r).$$

These two equations correspond to (4) and (2).

Finally, we derive the boundary condition for M^0 at $r = r_{tumor}$. In contrast to the assumed differentiability of the limits (14) for $r > r_{tumor}$, on the boundary we must expect that

$$\frac{\partial M^0}{\partial r}\Big|_{r_{tumor}} \ne \lim_{\varepsilon\to 0}\frac{\partial M^{\varepsilon}}{\partial r}\Big|_{r_{tumor}} = 0.$$

Indeed, although we assume, as in the surface model, that the markers cannot diffuse into the tumor, the existence of binding sites on the boundary effectively results in "disappearance" of markers at $r = r_{tumor}$. Hence, in the surface model the boundary $r = r_{tumor}$ is absorbing.

To derive the correct boundary condition there, we first note that the conservation law for the total number of markers may be written as

$$\frac{d}{dt}\frac{1}{r_{tumor}^2}\int_{r_{tumor}}^{\infty}\{M^\varepsilon(r,t)+[S^\varepsilon_{\rm tot}(r)-S^\varepsilon(r,t)]\}\ r^2\,dr=0,$$

since $M^\varepsilon(r,t)$ and $[S^\varepsilon_{\rm tot}(r)-S^\varepsilon(r,t)]$ are the concentrations of free and bound markers, respectively. Taking the limit as $\varepsilon\to 0$ and using (14), (17), and (19) gives

$$\frac{d}{dt}\left\{\frac{1}{r_{tumor}^2}\int_{r_{tumor}}^{\infty}M^0(r,t)\ r^2\,dr+\left[S_{tot}-\mathcal{S}^0\right]\right\}=0.$$

Therefore, by (14), (21), (20), and a calculation similar to (5) we get

$$\begin{aligned}\frac{\partial M^0(r_{tumor},t)}{\partial r}&=-\frac{1}{D}\frac{\partial \mathcal{S}^0}{\partial t}\\&=-\frac{1}{D}k_d^{vol}\left[S_{tot}-S^0(r,t)\right]+\frac{1}{D}k_b^{vol}M^0(r_{tumor},t)S^0(r,t),\end{aligned}$$

in accordance with (6). ∎

4. Rigorous justification of the surface model from the volumetric model. We now present a rigorous derivation of the surface model as the limit of the volumetric model. Unlike in sections 2 and 3, we do not make the assumption of radial symmetry.

4.1. Equations and results. When radial symmetry is not assumed we may write the equations governing the concentrations of unbound markers M^ε and unoccupied sites S^ε in the volumetric model in the form

$$M^\varepsilon_t(t,x)=D\Delta M^\varepsilon+k_d\,(S^\varepsilon_{\rm tot}(x)-S^\varepsilon)-k_bS^\varepsilon M^\varepsilon, \tag{22}$$

$$S^\varepsilon_t(t,x)=k_d\,(S^\varepsilon_{\rm tot}(x)-S^\varepsilon)-k_bS^\varepsilon M^\varepsilon. \tag{23}$$

Here x is a vector in $\mathbb{R}^d$ for some $d>1$ and $S^\varepsilon_{\rm tot}(x)$ is the total concentration of sites, which depends only on x since the sites are still assumed to remain stationary. Equations (22)–(23) are to hold in a smooth domain Ω in $\mathbb{R}^d$ whose inner surface $\partial_i\Omega$ is the boundary of the region occupied by the tumor. In order to reduce the technical complications we will assume that Ω is bounded, and its outer boundary, far from the tumor, will be denoted $\partial_o\Omega$. See Figure 4. However, the case when Ω is the entire exterior of the tumor surface $\partial_i\Omega$ could also be treated; in particular, the theorem on invariant regions (Theorem 3 below) is still valid in that case.

The total concentration of sites is clearly nonnegative, i.e.,

$$S^\varepsilon_{\rm tot}(x)\ge 0. \tag{24}$$

Also, since the sites are located near the surface of the tumor,

$$S^\varepsilon_{\rm tot}(x)=0\quad \text{for } d(x,\partial_i\Omega)>\varepsilon, \tag{25}$$

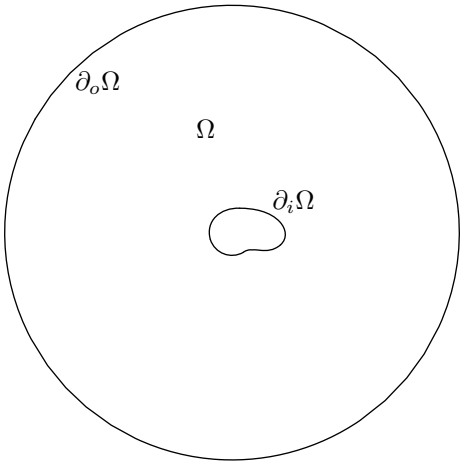


FIG. 4. *The domain Ω and its inner and outer boundaries $\partial_i\Omega$ and $\partial_o\Omega$.*

where $d(x, \partial_i\Omega)$ is the distance from a point x to the interior region Ω_i occupied by the tumor. We shall also assume that

$$S^{\varepsilon,0}_{\text{tot}}(y) := \int_0^\varepsilon S^\varepsilon_{\text{tot}}(y + \tau\nu)\, d\tau \text{ converges uniformly to some } S^0_{\text{tot}}(y) \text{ as } \varepsilon \to 0, \tag{26}$$

where ν denotes the unit normal on $\partial_i\Omega$ pointing into Ω. Note that conditions (24)–(26) are satisfied when $S^\varepsilon_{\text{tot}}(y + \tau\nu) = \frac{1}{\varepsilon}\phi(y)\psi(\frac{\tau}{\varepsilon})$, where ϕ and ψ are nonnegative continuous functions and $\psi(s)$ vanishes for $s > 1$.

The boundary condition for M^ε is

$$\nu \cdot \nabla M^\varepsilon = 0 \qquad \text{on } \partial\Omega, \tag{27}$$

which means that markers do not leave the region Ω. The initial conditions are

$$S^\varepsilon(0, x) = S^\varepsilon_{\text{tot}}(x), \tag{28}$$

which means that all sites are originally unoccupied, and

$$M^\varepsilon(0, x) = M_0(x) \geq 0. \tag{29}$$

Furthermore, we will assume that

$$S^\varepsilon_{\text{tot}}(x) M_0(x) \equiv 0, \tag{30}$$

i.e., that initial locations of the markers and sites do not overlap. Physically this is another expression of the assumption that time zero occurs before the sites start to become occupied, since if the marker and site locations overlapped at time zero, then some sites would have become occupied before then. Mathematically it avoids having an initial layer in which the site-occupation reaction would quickly reduce the size of the reaction term down to order one.

The local-in-time existence of a unique solution to (22)–(23), (27)–(29) can be obtained via the method of [23], by substituting the Green's function for the Neumann boundary value problem for $u_t = D\Delta u$ in place of the whole-space Green's function

in [23, eq. (2.7)]. Furthermore, the uniform bounds on the solution obtained in the next subsection imply that solution exists for all positive time.

The equations for the surface model are

$$M_t = D\Delta M \tag{31}$$

in Ω,

$$\nu \cdot \nabla M = \frac{k_d \left(S^0_{\rm tot}(x) - S\right) - k_b S M}{D} \tag{32}$$

on the inner boundary $\partial_i\Omega$ bounding the tumor region, (27) on the outer boundary $\partial_o\Omega$, and

$$S_t = k_d \left(S^0_{\rm tot}(y) - S\right) - k_b S M \tag{33}$$

on the inner boundary $\partial_i\Omega$. Although the existence of solutions to the surface model could be obtained by an appropriate adaptation of the method used for the volumetric model, we will obtain existence here as a by-product of our convergence result.

Our main result is that solutions of the volumetric model converge to those of the boundary model as the parameter ε tends to zero.

THEOREM 1. *Assume that S^ε_{tot} satisfies* (24)–(26). *Let $(M^\varepsilon, S^\varepsilon)$ be the solution of* (22)–(23) *and* (27) *having initial data of the form* (28)–(29) *belonging to $C^2(\overline{\Omega})$ and satisfying* (30). *Then as $\varepsilon \to 0$, M^ε and*

$$\mathcal{S}^\varepsilon(y,t) := \int_0^\varepsilon S^\varepsilon(y + \tau\nu, t)\, d\tau \tag{34}$$

converge to the unique solution (M, S) of (31) *and* (33) *satisfying* (32) *on $\partial_i\Omega$ and* (27) *on $\partial_o\Omega$ and having initial data* (29) *and $S(0,y) = S^0_{tot}(y)$.*

In the next subsection we will prove some uniform bounds that will be used in the subsequent subsection to take the limit as $\varepsilon \to 0$. Those uniform bounds also imply an upper bound for the fraction of sites that are occupied at any time.

THEOREM 2. *The ratio*

$$\frac{\int_\Omega \left[S^\varepsilon_{tot}(x) - S^\varepsilon(x,t)\right]\, dx}{\int_\Omega S^\varepsilon_{tot}(x)\, dx}$$

of occupied sites to total sites is never more than

$$\frac{k_b \max_{x\in\Omega} M_0(x)}{k_d + k_b \max_{x\in\Omega} M_0(x)}. \tag{35}$$

4.2. Uniform estimates.

4.2.1. Invariant regions. In order to be able to take the limit of the solutions as $\varepsilon \to 0$, we need certain uniform bounds on those solutions. Some of the required bounds follow from the theory of invariant domains. See [24, Chap. 14] for an introduction and references. The version we will apply, in which the hypotheses have been weakened somewhat, is the following special case of [21, Thm. 3].

THEOREM 3. *Assume that the following hold:*

1. *Domain: Ω is either a smooth bounded domain or the exterior of a smooth bounded domain.*

2. *Smoothness:* $u := (u_1, \ldots, u_p)$ *is continuous on* $[0,T] \times \overline{\Omega}$, *and* u_t, u_{x_j}, *and* $u_{x_j x_k}$ *are continuous on* $(0,T) \times \Omega$.
3. *PDE:* u *satisfies the system*

$$\partial_t u_j - d_j \Delta u_j = f_j(t,x,u) \tag{36}$$

for $0 < t < T \leq \infty$ *and* $x \in \Omega$, *where* $d_j \geq 0$ *and* $f_j \in C^1$.
4. *Region: The region* $\mathcal{R} := \{u \mid a_j \leq u_j \leq b_j\}$ *is invariant for the system of ODEs obtained by setting every* d_j *in* (36) *to zero, i.e.,* $f_j \leq 0$ *when* $u_j = b_j$ *and* $a_k \leq u \leq b_k$ *for each* $k \neq j$, *and* $f_j \geq 0$ *when* $u_j = a_j$ *and* $a_k \leq u \leq b_k$ *for each* $k \neq j$.
5. *Initial condition:* $u(0,x) = u_0(x)$ *for* $x \in \Omega$, *where* $u_0(x) \in \mathcal{R}$ *for all such* x.
6. *Boundary condition:* $\partial_\nu u = 0$ *for* $0 < t < T$ *and* $x \in \partial\Omega$, *where* ν *denotes the exterior normal on the boundary* $\partial\Omega$.
7. *Behavior at infinity: If* Ω *is unbounded, then* $u(t,x) = o(|x|^2)$ *as* $x \to \infty$.

Then the solution $u(t,x)$ *remains in* $\mathcal{R}$ *for* $0 \leq t \leq T$ *and* $x \in \overline{\Omega}$.

The region that we wish to show to be invariant for (22)–(23) depends on x, as will be apparent below, so Theorem 3 does not apply directly. That theorem could be extended to the case of x-dependent invariant regions either by using the fact that the S-component does not diffuse or via the approach of [15]. However, it will in any case be convenient to transform our system via

$$N^\varepsilon = 1 + \frac{k_b M^\varepsilon}{k_d}, \qquad R^\varepsilon = \frac{S^\varepsilon}{S^\varepsilon_{\rm tot}(x)}, \tag{37}$$

which yields equations whose invariant sets will not depend on x. Although $S^\varepsilon_{\rm tot}(x)$ vanishes in much of the domain Ω, which makes $R^\varepsilon(0,x)$ undefined, the initial condition (28) implies that $R^\varepsilon(0,x) \equiv 1$ where $S^\varepsilon_{\rm tot}(x)$ is nonzero, so the initial data for R^ε extend naturally to

$$R^\varepsilon(0,x) = 1. \tag{38}$$

The initial data

$$N^\varepsilon(0,x) = 1 + \frac{k_b M_0(x)}{k_d} \tag{39}$$

for N^ε are obtained directly from those of M via the transformation (37).

In terms of the new variables N^ε and R^ε, system (22)–(23) becomes

$$N^\varepsilon_t = D\Delta N^\varepsilon + k_b S^\varepsilon_{\rm tot}(x)\left[1 - N^\varepsilon R^\varepsilon\right], \tag{40}$$

$$R^\varepsilon_t = k_d\left[1 - N^\varepsilon R^\varepsilon\right]. \tag{41}$$

The boundary condition (27) becomes

$$\nu \cdot \nabla N^\varepsilon = 0 \qquad \text{on } \partial\Omega. \tag{42}$$

LEMMA 4. *Suppose that* N^ε *and* R^ε *satisfy the system* (40)–(41) *plus the boundary condition* (42) *and have initial data* (38)–(39). *Define* $N_{\max} := \max N^\varepsilon(0,x)$. *Then*

$$1 \leq N^\varepsilon \leq N_{\max}, \qquad \tfrac{1}{N_{\max}} \leq R^\varepsilon \leq 1. \tag{43}$$

Proof. The vector field $(k_b S^\varepsilon_{\rm tot}(x)(1 - N^\varepsilon R^\varepsilon), k_d(1 - N^\varepsilon R^\varepsilon))$ points left and down at points above the curve $R^\varepsilon = \frac{1}{N^\varepsilon}$, and up and to the right below that curve. For

any k, that curve intersects the rectangle $1 \le N^\varepsilon \le k$, $\frac{1}{k} \le R^\varepsilon \le 1$ at the upper left and lower right corners, and so points inwards everywhere on the boundary of that rectangle, except at those corners, where it vanishes. Since $M_0(x) \ge 0$, the initial data (39) for N^ε satisfy $N^\varepsilon(0,x) \ge 1$, while the initial data for R^ε are identically one. Thus, the initial data lie in the rectangle (43). Now apply Theorem 3. ∎

Translating the bounds (43) back into the original variables $(M^\varepsilon, S^\varepsilon)$ yields Theorem 2.

In order to obtain convergence of solutions as $\varepsilon \to 0$ it is necessary to obtain estimates for derivatives as well. However, since $S^\varepsilon_{\rm tot}(x)$ and its derivatives are not uniformly bounded, it is not convenient to estimate the evolution of spatial derivatives of N^ε. But $S^\varepsilon_{\rm tot}(x)$ is independent of time, so we can obtain estimates for time derivatives; a spatial estimate will then be obtained by using the theory of elliptic PDEs.

LEMMA 5. *Let N^ε and R^ε satisfy the conditions of Lemma* 4. *Suppose in addition that condition* (30) *holds and that for some $\tilde b_\pm$,*

$$\tilde b_- \le D\Delta N_0(x) \le \tilde b_+. \tag{44}$$

Then for some b_-, b_+, B_-, and B_+

$$b_- \le N^\varepsilon_t \le b_+, \qquad B_- \le R^\varepsilon_t \le B_+. \tag{45}$$

Proof. Since the equation for R^ε is an ODE, a uniform bound on its time derivative follows from the bounds for N^ε and R^ε; i.e., the second half of (45) holds. Taking the time derivative of (40) for N^ε and substituting for the time derivative of R^ε from (41) yields

$$(N^\varepsilon_t)_t = D\Delta(N^\varepsilon_t) - k_b R^\varepsilon S^\varepsilon_{\rm tot}(x)\left[(N^\varepsilon_t) + \frac{k_d N^\varepsilon(1-N^\varepsilon R^\varepsilon)}{R^\varepsilon}\right]. \tag{46}$$

The bounds (43) imply both a lower bound $\hat b_-$ and an upper bound $\hat b_+$ for the expression $\frac{k_d N^\varepsilon(1-N^\varepsilon R^\varepsilon)}{R^\varepsilon}$ appearing in (46). Condition (30) implies that $N^\varepsilon_t(0,x) = D\Delta N^\varepsilon(0,x)$, so the bounds (44) imply the same bounds for $N^\varepsilon_t(0,x)$. Define $b_- := \min\{\hat b_-, \tilde b_-\}$ and $b_+ := \max\{\hat b_+, \tilde b_+\}$. Since differentiating (42) with respect to time shows that $\nu\cdot\nabla(N^\varepsilon_t)$ also vanishes on the boundary, another application of Theorem 3 shows that the first half of (45) holds. ∎

4.2.2. Elliptic estimate. Solving (40) for ΔN^ε yields

$$\Delta N^\varepsilon = \frac{N^\varepsilon_t - k_b S^\varepsilon_{\rm tot}(x)\,[1 - N^\varepsilon R^\varepsilon]}{D}. \tag{47}$$

Although the right side of (47) is not known to be uniformly bounded, it is a uniformly bounded function times the known expression $S^\varepsilon_{\rm tot}(x)$ plus a bounded function. This will allow us to obtain uniform estimates for N^ε and R^ε in an appropriate Hölder space, and also to determine the behavior of $\nu\cdot\nabla N^\varepsilon$ near the boundary.

DEFINITION 6. *Let Ω be a domain in $\mathbb{R}^d$, let $B_r(x_0)$ denote the ball of radius r centered at x_0, and suppose that $1 \le p \le \infty$. A measurable function f belongs to the Morrey space $\mathcal{M}^p(\Omega)$ if*

$$\|f\|_{\mathcal{M}^p} := \sup_{x_0\in\Omega}\sup_{r>0} \frac{\int_{B_r(x_0)\cap\Omega} |f(x)|\,dx}{r^{d(1-1/p)}} < \infty. \tag{48}$$

An easy calculation shows that $L^p(\Omega) \subset \mathcal{M}^p(\Omega)$ [9, sect. 7.9]. However, the reverse is not true. In particular, although $S^\varepsilon_{\rm tot}$ is uniformly bounded in L^p only for $p = 1$, its structural properties (24)–(26) ensure that $S^\varepsilon_{\rm tot}$, and hence also $g(x)S^\varepsilon_{\rm tot}$ for any bounded g, belong to $\mathcal{M}^d(\Omega)$.

LEMMA 7. *Suppose that S^ε_{tot} satisfies* (24)–(26) *and that $g^\varepsilon(x)$ is uniformly bounded. Then for some fixed constant c,*

$$\|g^\varepsilon(x)S^\varepsilon_{\rm tot}(x)\|_{\mathcal{M}^d} \le c. \tag{49}$$

Proof. Since g^ε is bounded it suffices to prove estimate (49) for $g^\varepsilon \equiv 1$. Since assumptions (24)–(26) imply that the total number of sites $\int_\Omega S^\varepsilon(x)\,dx$ is uniformly bounded, for any positive δ

$$\sup_{x_0\in\Omega}\sup_{r\ge\delta}\frac{\int_{B_r(x_0)}|S^\varepsilon_{\rm tot}(x)|\,dx}{r^{d(1-1/p)}}$$

is uniformly bounded. By picking δ small enough so that the map

$$(y,\tau)\mapsto y+\tau\nu(y) \tag{50}$$

is one to one $\partial_i\Omega\times[0,\delta]$ and satisfies $|[y_1+\tau_1\nu(y_1)]-[y_2+\tau_2\nu(y_2)]|\ge c|y_1-y_2|$ there for some fixed positive c, we obtain that the intersection of the support of $S^\varepsilon_{\rm tot}$ with any ball $B_r(x_0)$ of radius at most δ is contained in a set of the form $\{y+\tau\nu(y)\mid y\in \partial_i\Omega\cap B_{kr}(y_0),\ \tau\in[0,\delta]\}$. Assumption (26) implies that the integral of $S^\varepsilon_{\rm tot}$ over such a set is bounded by a constant times the volume of a ball of radius kr in dimension $d-1$, which is a constant times r^{d-1}. Combining this with the bound for $r\ge\delta$ yields (49). ∎

Although integration against the Green's function for the Laplacian does not map L^1 into C^0 since functions in L^1 can tend weakly to a delta function, it does map $\mathcal{M}^p$ for p sufficiently large into the space $C^{0,\alpha}$ of Hölder-continuous functions for some appropriate positive α. We begin with a general result.

LEMMA 8. *Suppose that $f\in\mathcal{M}^p$ with $p>1$. Then for $\mu>\frac{1}{p}$, $Tf(x):=\int_\Omega\frac{f(y)}{|x-y|^{d(1-\mu)}}\,dy$ belongs to $C^{0,\alpha}$ for $\alpha<\min\{1,d(\mu-\frac{1}{p})\}$, where d is the spatial dimension. Furthermore, the $C^{0,\alpha}$ seminorm of Tf is bounded by a constant times the $\mathcal{M}^p$ norm of f.*

Proof. By interpolating between the elementary inequalities

$$\left|\frac{1}{|x_1-y|^\beta}-\frac{1}{|x_2-y|^\beta}\right|\le\left[\frac{1}{|x_1-y|^\beta}+\frac{1}{|x_2-y|^\beta}\right]$$

and

$$\left|\frac{1}{|x_1-y|^\beta}-\frac{1}{|x_2-y|^\beta}\right|\le c|x_1-x_2|\left[\frac{1}{|x_1-y|^{\beta+1}}+\frac{1}{|x_2-y|^{\beta+1}}\right],$$

we obtain that for any $\gamma\in[0,1]$,

$$\left|\frac{1}{|x_1-y|^\beta}-\frac{1}{|x_2-y|^\beta}\right|\le c(\gamma)|x_1-x_2|^\gamma\left[\frac{1}{|x_1-y|^{\beta+\gamma}}+\frac{1}{|x_2-y|^{\beta+\gamma}}\right]. \tag{51}$$

Pick $\alpha\in(0,1)$ such that $\alpha<d(\mu-\frac{1}{p})$. Applying (51) with $\gamma=\alpha$ and $\beta=d(1-\mu)$ yields

$$|[Tf](x_1)-[Tf](x_2)|\le c(\alpha)|x_1-x_2|^\alpha\sum_{j=1}^2\int_\Omega\frac{1}{|x_j-y|^{d(1-\lceil\mu-\frac{\alpha}{d}\rceil)}}f(y)\,dy. \tag{52}$$

Since $\mu - \frac{\alpha}{d} > \frac{1}{p}$ by construction, the integrals on the right side of (52) are bounded by [9, Lem. 7.18]. ∎

Using Lemma 8 we can show that N^ε and R^ε are uniformly bounded in some Hölder space.

LEMMA 9. *Under the conditions of Lemma* 5*, for bounded times the solutions* N^ε *and* R^ε *are uniformly bounded in* $C^{0,\alpha}$ *for* $\alpha < 1$.

Proof. Let G be the Neumann Green's function for the Laplacian in Ω, so that for any f having mean zero, the solutions to $\Delta u = f$ in Ω, $\frac{\partial u}{\partial \nu} = 0$ on $\partial\Omega$, are $u(x) = \int_\Omega G(x,y) f(y)\, dy + c$. The singularity of G when $x = y$ is of the same order as the Newtonian potential, i.e., $\frac{1}{|x-y|^{d-2}}$, or $\log(|x-y|)$ when $d = 2$. Since the smooth part of G makes a smooth contribution to u, it suffices to show that $\int_\Omega \frac{1}{|x-y|^{d(1-\mu)}} \Delta N^\varepsilon(y)\, dy$ belongs to $C^{0,\alpha}$ for $\alpha < 1$, where $\mu = \frac{2}{d}$ for $d > 2$ and is arbitrarily close to one for $d = 2$.

Now any bounded function belongs to $\mathcal{M}^\infty$ and hence also to $\mathcal{M}^p$ for any $p < \infty$. Hence (47) plus the bounds for N^ε, R^ε, and N_t^ε and Lemma 7 implies that ΔN^ε belongs to $\mathcal{M}^d$.

Lemma 8 therefore shows that $\int_\Omega \frac{1}{|x-y|^{d(1-\mu)}} \Delta N^\varepsilon(y)\, dy$ belongs to $C^{0,\alpha}$ for $\alpha < d(\frac{2}{d} - \frac{1}{d}) = 1$, and that its $C^{0,\alpha}$ seminorm is uniformly bounded. ∎

4.3. Taking the limit. By Ascoli's theorem, the uniform bounds obtained in the previous subsection imply the convergence along subsequences as $\varepsilon \to 0$.

COROLLARY 10. *Under the conditions of Lemma* 5*, for every sequence of values of* ε *there is a subsequence for which* N^ε *and* R^ε *converge uniformly in* Ω *for bounded times. The limits* N *and* R *satisfy the same bounds* (43) *as* N^ε *and* R^ε.

We first consider the limit in terms of the variables (N, R).

LEMMA 11. *The limits* (N, R) *satisfy*

$$N_t = D\Delta N \tag{53}$$

in Ω,

$$\nu \cdot \nabla N = \frac{k_b S_{tot}^0(y)(1 - NR)}{D} \tag{54}$$

on the inner boundary $\partial_i\Omega$ *bounding the tumor region,* (42) *on the outer boundary* $\partial_o\Omega$*, and*

$$R_t = k_d(1 - NR) \tag{55}$$

in $\overline{\Omega}$*, including in particular, on the inner boundary* $\partial_i\Omega$*. The initial values of* N *and* R *are the same as for the original system.*

Proof. Taking the weak limit of the PDE (40) yields (53) within the domain Ω, since N_t^ε converges weakly to N_t and the reaction term tends to zero in every compact subset of Ω. Since R^ε satisfies an ODE, the convergence of N^ε and R^ε implies that the limits satisfy (55). Since the convergence of N^ε and R^ε is uniform in time as well as space, their limits have the same initial values.

Finally, in order to obtain (54), let Ω_δ denote the subset of Ω whose distance to the inner boundary $\partial_i\Omega$ is less than δ. For sufficiently small δ, $\Omega_\delta = \{y + \tau\nu \mid y \in \partial_i\Omega,\ 0 < \tau < \delta\}$. The boundary of Ω_δ is then the disjoint union of $\partial_i\Omega$, and the set $\partial_\delta\Omega$ of points in Ω whose distance to $\partial_i\Omega$ is exactly δ. Since (53) implies that N is smooth in Ω, the derivative $\frac{\partial N}{\partial \nu}$ of N with respect to the outer normal on $\partial_\delta\Omega$ is well defined. As before, let $y(x)$ denote the mapping sending $x = y + \tau\nu$ to y.

By Green's formula, for any smooth function ψ

$$\int_{\partial_\delta\Omega}\psi(y(x))\frac{\partial N^\varepsilon}{\partial\nu}=\int_{\Omega_\delta}\psi\Delta N^\varepsilon-N^\varepsilon\Delta\psi+\int_{\partial_\delta\Omega}N^\varepsilon\frac{\partial\psi(y(x))}{\partial\nu} \tag{56}$$

since both $\frac{\partial N^\varepsilon}{\partial\nu}$ and $\frac{\partial\psi(y(x))}{\partial\nu}$ vanish on $\partial_i\Omega$.

Since $\frac{\partial\psi(y(x))}{\partial\nu}$ vanishes on $\partial_i\Omega$, it is $O(\delta)$ on $\partial_\delta\Omega$. Also, since N^ε is uniformly bounded and the volume of Ω_δ is $O(\delta)$, $\int_{\Omega_\delta}N^\varepsilon\Delta\psi=O(\delta)$. Similarly, upon substituting (47) into (56), the term involving N_t^ε contributes $O(\delta)$. Hence

$$\int_{\partial_\delta\Omega}\psi(y(x))\frac{\partial N^\varepsilon}{\partial\nu}=-\int_{\Omega_\delta}\psi(y(x))\frac{k_bS_{\rm tot}^\varepsilon(x)\,[1-N^\varepsilon R^\varepsilon]}{D}+O(\delta). \tag{57}$$

Now take the limit as first $\varepsilon\to0$ and then $\delta\to0$. The left side of (57) tends to $\int_{\partial_i\Omega}\psi(y)\frac{\partial N}{\partial\nu}$. Since the term $O(\delta)$ on the right side is uniform in ε, it contributes nothing to the combined limit. Hence

$$\begin{aligned}\int_{\partial_i\Omega}\psi(y)\frac{\partial N}{\partial\nu}\,d\sigma(y)&=-\lim_{\delta\to0}\lim_{\varepsilon\to0}\int_{\Omega_\delta}\psi(y(x))\frac{k_bS_{\rm tot}^\varepsilon(x)\,[1-N^\varepsilon R^\varepsilon]}{D}\,dx\\&=-\lim_{\delta\to0}\lim_{\varepsilon\to0}\int_{\Omega_\varepsilon}\psi(y(x))\frac{k_bS_{\rm tot}^\varepsilon(x)\,[1-N^\varepsilon R^\varepsilon]}{D}\,dx\\&=-\lim_{\delta\to0}\lim_{\varepsilon\to0}\int_{\partial_i\Omega}\psi(y)\int_0^\varepsilon\frac{k_bS_{\rm tot}^\varepsilon(y+\tau\nu)\,[1-N^\varepsilon(y+\tau\nu)R^\varepsilon(y+\tau\nu)]}{D}\,d\tau\,d\sigma(y),\end{aligned} \tag{58}$$

where we have used the fact that the difference between dx and $d\tau\,d\sigma(y)$ tends to zero with the distance from the boundary. Since the total integral of $S_{\rm tot}^\varepsilon$ is uniformly bounded, and $N^\varepsilon R^\varepsilon$ converge uniformly, we may replace that expression in (58) with its limit NR. In addition, $N(y+\tau\nu)R(y+\tau\nu)=N(y)R(y)+o(1)$, so in fact

$$\begin{aligned}&\int_{\partial_i\Omega}\psi(y)\frac{\partial N}{\partial\nu}\,d\sigma(y)\\&\quad=-\int_{\partial_i\Omega}\psi(y)\frac{k_b\,[1-N(y)R(y)]}{D}\left[\lim_{\varepsilon\to0}\int_0^\varepsilon S_{\rm tot}^\varepsilon(y+\tau\nu)\,d\tau\right]d\sigma(y)\\&\quad=-\int_{\partial_i\Omega}\psi(y)\frac{k_b\,[1-N(y)R(y)]}{D}S_{\rm tot}^{\varepsilon,0}(y)\,d\sigma(y)\\&\quad=-\int_{\partial_i\Omega}\psi(y)\frac{k_b\,[1-N(y)R(y)]}{D}S_{\rm tot}^0(y)\,d\sigma(y)\end{aligned} \tag{59}$$

by assumption (26). Since ψ is an arbitrary smooth function, (59) implies (54). ∎

In order to obtain convergence of the full sequence $(N^\varepsilon,R^\varepsilon)$ without restricting to some subsequence, it suffices to show that the limit obtained along different subsequences is unique.

LEMMA 12. *A bounded solution of* (53) *in* Ω, (54) *and* (55) *on the inner boundary* $\partial_i\Omega$, *and* (42) *on the outer boundary, with given initial data, is unique.*

Proof. Suppose that (N_j,R_j), $j=1,2$, are solutions having the same initial data. Define $N:=N1-N_2$ and $R:=R_1-R_2$. Multiplying the difference of (53) for N_1 and N_2 by N, integrating over Ω, and adding the integral over $\partial_i\Omega$ of R times the

difference of (55) for R_1 and R_2 yields

$$\begin{aligned}(60)\qquad &\frac{d}{dt}\frac{1}{2}\left[\int_\Omega N^2\,dx+\int_{\partial_i\Omega}R^2\,d\sigma(y)\right]\\ &\qquad=-D\int_\Omega\nabla N^2-\int_{\partial_i\Omega}\left[\frac{k_bS^0(y)}{D}N+k_dR\right]\,[R_1N+N_2R]\,d\sigma(y).\end{aligned}$$

Since both R_1 and N_2 are bounded and strictly positive, the elementary inequality $-c_1x^2+c_2xy-c_3y^2\le c_4y^2$ allows us to reduce (60) to

$$\begin{aligned}(61)\qquad &\frac{d}{dt}\frac{1}{2}\left[\int_\Omega N^2\,dx+\int_{\partial_i\Omega}R^2\,d\sigma(y)\right]\le c\int_{\partial_i\Omega}R^2\,d\sigma(y)\\ &\qquad\le c\cdot\frac{1}{2}\left[\int_\Omega N^2\,dx+\int_{\partial_i\Omega}R^2\,d\sigma(y)\right].\end{aligned}$$

Since $N\equiv0$ and $R\equiv0$ initially, (61) shows that they remain zero for all time. ∎

We are finally ready to prove the main result in terms of the original variables M^ε and S^ε.

Proof of Theorem 1. Lemma 12 implies the convergence of N^ε and R^ε to N and R holds without restricting to a subsequence. Upon transforming back to the variable M, (53) becomes (31). In view of the uniform convergence of R^ε, (26) implies the convergence of (34) to $S^0_{\text{tot}}R$, so (54)–(55) yield (32)–(33). The convergence of the initial data is obtained similarly. ∎

5. Possible extensions. In this study we have aimed to give both heuristic and rigorous justifications for using the laws of chemical kinetics to describe binding and dissociation reactions that take place on surfaces. Our results were obtained, however, for a specific model involving a single reaction taking place on the boundary, with purely diffusive dynamics away from the boundary. Furthermore, a more realistic model of the fluorophore-antibody–based imaging studied here should also include advection effects to account for the continuous drainage of the interstitial fluid. Other chemical and biological systems involve more complicated interactions, possibly including several reactions on the boundary. To what extent can our methods be applied in these more general situations?

Since both our heuristic and rigorous analyses implicitly or explicitly require uniform bounds on reaction concentrations or ratios, our methods seem to require the presence of an invariant region for the reaction dynamics. Models of a variety of chemical and biological systems for which the existence of such regions have been deduced or assumed have been studied [14, 16, 17].

Additional restrictions must be placed on reactants that concentrate at the boundary surface. First, the reaction terms must be at most linear in those reactants. In terms of our heuristic analysis, this condition arises because the volumetric concentrations of those reactants tend to Dirac delta functions, which makes superlinear functions of those concentrations diverge to infinity even when considered in the sense of distributions. In our rigorous analysis linearity is needed in order for the second change of variables in (37) to be helpful.

That reaction terms be at most linear in "surface" reactants seems to be a necessary condition for our results to hold, rather than a technical limitation. The presence of a superlinear growth term would make the reaction blow up as the reactant concentrates at the boundary, while superlinear decay terms would make reactions disappear in the limit.

Our methods can accommodate advection terms involving the "volumetric" reactants that do not concentrate at the boundary (e.g., advection of markers). Note, however, that our analysis does not apply to models that allow for advection or even diffusion of the "surface" reactants. Indeed, such advection or diffusion terms would change the model substantially, since they would cause the "surface" reactants to leave the region near the boundary.

REFERENCES

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 3rd ed., Garland Publishing, New York, 1994.

[2] D. A. Frank-Kamenetsky, *Diffusion and Heat Transfer in Chemical Kinetics*, Plenum, New York, 1969.

[3] G. Gannot, I. Gannot, A. Buchner, H. Vered, and Y. Keisari, *Increase in immune cell infiltration with progression of oral epithelium from hyperkeratosis to dysplasia and carcinoma*, British Journal of Cancer, 86 (2002), pp. 1444–1448.

[4] G. Gannot, A. Buchner, and Y. Keisari, *Interaction between the immune system and tongue squamous cell carcinoma induced by* 4*-nitroquinoline N-oxide in mice*, Oral Oncol., 40 (2004), pp. 287–297.

[5] I. Gannot, A. H. Gandjbakhche, G. Gannot, P. C. Fox, and R. F. Bonner, *Optical simulations experiments for development of a noninvasive technique for the diagnosis of diseased salivary glands in situ,* J. Med. Phy., 27 (1998), pp. 1139–1144.

[6] I. Gannot, G. Gannot, A. Garashi, A. Gandjbakhche, A. Buchner, and Y. Keisari, *Laser activated fluorescence measurements and morphological features—an in vivo study of clearance time of FITC tagged cell markers*, J. Biomed. Opt., 7 (2002), pp. 14–19.

[7] I. Gannot, A. Garashi, G. Gannot, V. Chernomordik, and A. Gandjbakhche, *Quantitative* 3*-D imaging of tumor labeled with exogenous specific fluorescence markers*, J. Appl. Opt., 42 (2003), pp. 3073–3080.

[8] I. Gannot, A. Garashi, V. Chernomordik, and A. Gandjbakhche, *Quantitative optical imaging of pharmacokinetics of specific fluorescent tumor markers through turbid media such as tissue*, Opt. Lett., 29 (2004), pp. 742–744.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[10] J. Hale and G. Raugel, *Reaction-diffusion equation on thin domains*, J. Math. Pures Appl., 71 (1992), pp. 33–95.

[11] A. Hammer, *Modeling, Analysis, and Optimization of Fluorescenated Antibody Based Imaging*, M.Sc. thesis, Tel Aviv University, Tel Aviv, 2003.

[12] D. Jones, H. V. Kojouharov, D. Le, and H. Smith, *Bacterial wall attachment in a flow reactor*, SIAM J. Appl. Math,, 62 (2002), pp. 1728–1771.

[13] L. S. Jung, J. S. Shumaker-Parry, C. T. Campbell, S. S. Yee, and M. H. Gelb, *Quantification of tight binding to surface-immobilized phospholipid vesicles using surface plasmon resonance: Binding constant of phospholipase* $A(2)$, J. Amer. Chem. Soc., 122 (2000), pp. 4177–4184.

[14] I. C. Kim, *Singular limits of chematoxis-growth model*, Nonlinear Anal. Ser. A: Theory Methods, 466 (2001), pp. 817–834.

[15] H. J. Kuiper, *Invariant sets for nonlinear elliptic and parabolic systems*, SIAM J. Math. Anal., 11 (1980), pp. 1075–1103.

[16] J. S. McGough and K. L. Riley, *A priori bounds for reaction-diffusion systems arising in chemical and biological dynamics*, Appl. Math. Comput., 163 (2005), pp. 1–16.

[17] M. Mincheva and D. Siegel, *Stability of mass action reaction-diffusion systems*, Nonlinear Anal., 56 (2004), pp. 1105–1131.

[18] P. F. Morrison, P. M. Bungay, J. K. Hsiao, B. A. Ball, I. N. Mefford, and R. L. Dedrick, *Quantitative microdialysis: Analysis of transients and application to pharmacokinetics in brain*, J. Neurochem., 57 (1991), pp. 103–119.

[19] H. Nygren, *Kinetics of antibody binding to surface-immobilized antigen. Analysis of data and an empiric model*, Biophys. Chem., 52 (1994), pp. 45–50.

[20] M. Praxmarer, C. Sung, P. M. Bungay, and W. W. van Osdol, *Computational models of antibody-based tumor imaging and treatment protocols*, Ann. Biomed. Eng., 29 (2001), pp. 340–358.

[21] R. Redlinger, *Invariant sets for strongly coupled reaction-diffusion systems under general boundary conditions*, Arch. Rational Mech. Anal., 108 (1989), pp. 281–291.

[22] A. R. Goldsby, T. J. Kindt, J. Kuby, and B. A. Osborne, *Immunology*, 4th ed., W. H. Freeman, San Francisco, 1997.
[23] J. Rauch and J. Smoller, *Qualitative theory of the FitzHugh-Nagumo equations*, Adv. in Math., 27 (1978), pp. 12–44.
[24] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Grundlehren Math. Wiss. 258, Springer, Berlin, 1983.
[25] W. Z. Wei, S. Ratner, A. M. Fulton, and G. H. Heppner, *Inflammatory infiltrates of experimental mammary cancers*, Biochim. Biophys. Acta, 865 (1986), pp. 13–26.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1389–1422

NONUNIFORM AVERAGE SAMPLING AND RECONSTRUCTION OF SIGNALS WITH FINITE RATE OF INNOVATION*

QIYU SUN†

Abstract. From an average (ideal) sampling/reconstruction process, the question arises whether the original signal can be recovered from its average (ideal) samples and, if so, how. We consider the above question under the assumption that the original signal comes from a prototypical space modeling signals with a finite rate of innovation, which includes finitely generated shift-invariant spaces, twisted shift-invariant spaces associated with Gabor frames and Wilson bases, and spaces of polynomial splines with nonuniform knots as its special cases. We show that the displayer associated with an average (ideal) sampling/reconstruction process, which has a well-localized average sampler, can be found to be well-localized. We prove that the reconstruction process associated with an average (ideal) sampling process is robust, locally behaved, and finitely implementable, and thus we conclude that the original signal can be approximately recovered from its incomplete average (ideal) samples with noise in real time. Most of our results in this paper are new even for the special case when the original signal comes from a finitely generated shift-invariant space.

Key words. average sampling, ideal sampling, signals with finite rate of innovation, shift-invariant spaces

AMS subject classifications. 42C15, 94A20, 46E30, 47B32, 41A15, 94A12

DOI. 10.1137/05063444X

1. Introduction.

Modern digital data processing of functions (or signals or images) uses a discretized version of the original function that is obtained by (average) sampling on a discrete set [2]. The classical model is the Shannon sampling and reconstruction on the band-limited space B_Ω, the space of all square-integrable functions on the real line with their Fourier transform supported in $[-\Omega, \Omega]$. From the Shannon sampling theorem, sampling a function f in B_π on the uniform grid $\mathbf{Z}$ yields an ℓ^2 sequence $(f(k))_{k\in\mathbf{Z}}$, and conversely the original function f can be recovered from its sampling data $\{f(k), k \in \mathbf{Z}\}$ by the following reconstruction formula:

$$f(x) = \sum_{k\in\mathbf{Z}} f(k)\mathrm{sinc}(x-k),\ x \in \mathbf{R}, \tag{1.1}$$

where the sinc function is defined by $\mathrm{sinc}(x) = \frac{\sin \pi x}{\pi x}$. The above sampling and reconstruction theorem gives a framework for converting analogue signals into sequences, which can be processed digitally and converted back into analogue signals via the reconstruction formula (1.1). For the ideal sampling and reconstruction on the band-limited space and the finitely generated shift-invariant spaces, there is an extensive literature (see, for example, the recent review papers [2, 57] and monographs [12, 14, 43]).

In most physical circumstances, due to the nonideal acquisition device at the sampling location, it is not realistic to measure the sample $f(\gamma)$ of the original signal f in a space V at the location γ exactly. So a better assumption is that the sampled data are of the form $\langle f, \psi_\gamma\rangle$,

$$A:\ V \ni f \longmapsto (\langle f, \psi_\gamma\rangle)_{\gamma\in\Gamma}, \tag{1.2}$$

*Received by the editors June 24, 2005; accepted for publication (in revised form) July 17, 2006; published electronically December 26, 2006.
http://www.siam.org/journals/sima/38-5/63444.html
†Department of Mathematics, University of Central Florida, Orlando, FL 32816 (qsun@mail.ucf.edu).

where ψ_γ, to be known as the *average sampling functional*, reflects the characteristic of the nonideal acquisition device at the sampling location γ. We call the above sampling process an *average sampling process*, and call the collection $\Psi := \{\psi_\gamma, \gamma \in \Gamma\}$ of average sampling functionals an *average sampler*. Clearly, the average sampling process becomes an ideal sampling process if the delta function δ_γ is used as the average sampling functional ψ_γ on every sampling location γ.

An easy model for the average sampling process is the discretization of the blurring process encountered in many practical situations, such as in the process of a remote camera imaging a scene and an observer viewing the sampled image [59]. For the average sampling and reconstruction on the band-limited space and on the finitely generated shift-invariant spaces, the reader may refer [1, 4, 5, 6, 22, 25, 28, 53, 54, 55, 61] and references cited therein.

The question arises from the average sampling process whether the original function can be recovered from its average samples and, if so, how. Specifically, the first part of the above question, which will be discussed in section 4, can be described as follows: Given a class of functions V on $\mathbf{R}^d$, find conditions on the average sampler $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ under which any function f in V can be reconstructed uniquely and stably from its average samples $\{\langle f, \psi_\gamma\rangle : \gamma \in \Gamma\}$.

The second part of the above question arising from the average sampling process is the reconstruction process from the average (ideal) samples:

$$D : (c_\gamma)_{\gamma\in\Gamma} \longmapsto \sum_{\gamma\in\Gamma} c_\gamma \tilde{\psi}_\gamma \in V \tag{1.3}$$

such that

$$DAf = f \quad \text{for all} \quad f \in V. \tag{1.4}$$

Here for each $\gamma \in \Gamma$, the function $\tilde{\psi}_\gamma$, to be known as the *display block* at the location γ, reflects the characteristic of the display device at the sampling location γ. We call the above reconstruction process an *average reconstruction process* and the collection $\Psi := \{\tilde{\psi}_\gamma,\ \gamma \in \Gamma\}$ of display blocks an *displayer*.

For the efficiency and stability of the reconstruction process (1.3), (1.4) to recover a function f in the space V from its averaging samples $\{\langle f, \psi_\gamma\rangle, \gamma \in \Gamma\}$ or from its ideal samples $\{f(\gamma), \gamma \in \Gamma\}$, we require the corresponding displayer $\tilde{\Psi} := \{\tilde{\psi}_\gamma, \gamma \in \Gamma\}$ to be *well-localized*, and the average sampling/reconstruction process (1.3), (1.4) to be *robust, local-behaved*, and *finitely implementable*. In this paper, we show that those natural requirements for the average (ideal) sampling/reconstruction process would be met when signals in the space V have a finite rate of innovation and the average sampler Ψ is well-localized; see section 2.3 for our reasons for considering sampling/reconstruction of signals with a finite rate of innovation. Here a signal is said to have a *finite rate of innovation* if it has a finite degree of freedom per unit of time; see [23, 34, 40, 42, 45, 46, 58].

The paper is organized as follows. We divide section 2 into five subsections. In the first three subsections, we make some basic assumptions on the sampling set Γ, the average sampler $\Psi = \{\psi_\gamma, \gamma \in \Gamma\}$, and the space V, which is where the original function f for the average sampling/reconstruction process comes from. Briefly, we assume that the sampling set Γ is a relatively separated subset of $\mathbf{R}^d$, the average sampler Ψ is well-localized in the sense that every average sampling functional ψ_γ in the average sampler Ψ is essentially located in a neighborhood of $\gamma \in \Gamma$, and the space V is the space $V_q(\Phi, \Lambda)$, that is, as originally introduced in [52] for modeling

signals with a finite rate of innovation. In the last two subsections, we recall some basic properties of the space $V_q(\Phi,\Lambda)$ from [52], and introduce a simplified model of our average (ideal) sampling/reconstruction process for the readers' convenience.

Since each display block $\tilde{\psi}_\gamma$ in the displayer $\tilde{\Psi} := \{\tilde{\psi}_\gamma, \gamma\in\Gamma\}$ reflects the characteristic of the display device at the sampling location $\gamma, \gamma\in\Gamma$, it is reasonable to *require* that for each $\gamma\in\Gamma$, the display block $\tilde{\psi}_\gamma$ be essentially supported in a neighborhood of the sampling location γ. In section 3, derived from a general theorem for localized frames (see [8, Theorem 1], [24, Theorem 3.6], and [32, Theorem 13]), it is shown that such a requirement would be met for average (ideal) sampling in the space $V_2(\Phi,\Lambda)$ if the average sampler Ψ and the generator Φ for the space $V_2(\Phi,\Lambda)$ are well-localized (Theorems 3.1 and 3.2); see Remark 3.1 for a more general formulation of the well-localization of a displayer. The well-localization of displayers will play a crucial role in our study of stable average sampling in $V_r(\Phi,\Lambda)$ with $r\neq 2$ (Corollary 3.4), the robustness and local convergence of the reconstruction process from average (ideal) samples (Theorems 5.1–5.3 and 6.1–6.3), and exponential convergence of an iterative algorithm for the reconstruction process from average (ideal) samples (Theorems 7.1 and 7.2).

In section 4, we find conditions on the average sampling sampler $\Psi = \{\psi_\gamma : \gamma\in\Gamma\}$ (respectively, on the ideal sampling set Γ) under which any function f in $V_2(\Phi,\Lambda)$ can be reconstructed uniquely and stably from its average samples $\{\langle f,\psi_\gamma\rangle : \gamma\in\Gamma\}$ (respectively, from its ideal samples $\{f(\gamma),\gamma\in\Gamma\}$); see Theorems 4.1 and 4.2.

In the average (ideal) sampling/reconstruction process, we should bring the following situations into our consideration: the average samples $\{\langle f,\psi_\gamma\rangle : \gamma\in\Gamma\}$ may involve some noises (caused by, for example, measurement, storage, or transmission), and the average sampler Ψ may not be exactly the same as the one we expect (because of the mathematical modeling or the measurement of the acquisition device). In section 5, we consider the numerical stability of the reconstruction process (1.3), (1.4). We show that if the average sampler Ψ and the generator Φ for the space $V_2(\Phi,\Lambda)$ are well-localized, then the reconstruction process (1.3), (1.4) for $f\in V_2(\Phi,\Lambda)$ is stable under the corruption of the average (ideal) sampling data, and under the perturbation of averaging samplers, ideal sampling sets and the displayers; see Theorems 5.1, 5.2, and 5.3 for details. Then we conclude that the reconstruction process (1.3), (1.4) for $f\in V_2(\Phi,\Lambda)$ is *robust*.

By the reconstruction process (1.3), (1.4), any function f in the space V can be recovered fully when its average (ideal) sampling data are received completely. In some situations (such as when data missing are in the transmission and in the real-time reconstruction process), we are required to recover the original function (signal) partially from incomplete (ideal) average samples. We observe from the well-localization of the average sampler Ψ and of the ideal sampling set Γ that the average sampling data $\langle f,\psi_\gamma\rangle$ and the ideal sampling data $f(\gamma)$ catch the information of the function f essentially in a neighborhood of the sampling location γ for every $\gamma\in\Gamma$, which implies that the average (ideal) sampling procedure is locally behaved. So a natural question is whether the reconstruction procedure is locally behaved, or particularly whether a function $f\in V_2(\Phi,\Lambda)$ on a certain region K can be recovered approximately (or exactly) from the average sampling data $\langle f,\psi_\gamma\rangle$ and the ideal sampling data $f(\gamma)$ for the sampling location γ in a neighborhood of that region. In section 6, it is proved that for any bounded region K, the original function in the space $V_2(\Phi,\Lambda)$ can be approximately recovered from its average (ideal) samples in an R-neighborhood $B(K,R) = \{y : \inf_{x\in K}|y-x|\le R\}$ of that region K via a finite algorithm (see

Theorems 6.1 and 6.2 for details), and moreover that the local convergence rate of the local reconstruction procedure is almost the same as the rate of polynomial (subexponential) decay of the generator Φ and of the average sampler Ψ. Therefore we conclude that the reconstruction process (1.3), (1.4) for $f \in V_2(\Phi, \Lambda)$ is *locally behaved* and *finitely implementable* (hence it could lead possibly to a real-time reconstruction algorithm) when the average sampler Ψ and the generator Φ are well-localized. As a by-product of the local reconstruction theorems, we obtain a necessary condition on the location Γ of average (ideal) sampling devices, which states that, for a stable average (ideal) sampling/reconstruction procedure on the space $V_2(\Phi, \Lambda)$, there exists a positive constant R_0 such that for any domain $K \subset \mathbf{R}^d$, the number of average (ideal) sampling devices located in R_0-neighborhood $B(K, R_0)$ of that domain K should exceed the degrees of freedom of the space $V_2(\Phi, \Lambda)$ in the domain K; see Theorem 6.3 for details. The above necessary condition, which is usually known as the density property, is established in [3] for the ideal sampling on the B-spline space (see [2, 3] and references cited therein for nonuniform sampling on the band-limited space, and see [8, 9] for nonuniform Gabor system).

In the average (ideal) sampling/reconstruction process, we need an efficient and fast numerical algorithm that recovers any function $f \in V_r(\Phi, \Lambda)$ from its average sampling values $\langle f, \psi_\gamma \rangle, \gamma \in \Gamma$, or from its ideal sampling values $f(\gamma), \gamma \in \Gamma$. In section 7, we modify the standard Richardson–Landweber iterative frame algorithm to implement the reconstruction process (1.3), (1.4) for signals $f \in V_r(\Phi, \Lambda)$ when average (ideal) samples are received completely, and show that the new iterative algorithm converges exponentially for any initial data in ℓ^r and that the limit agrees with the signal in the space $V_r(\Phi, \Lambda)$ whenever the initial data are obtained from average (ideal) sampling of that signal (see [2, 6, 22] for convergence results similar to the standard Richardson–Landweber iterative frame algorithm in the shift-invariant setting). The Richardson–Landweber iterative algorithm is easily implemented, but it provides slow convergence in general. Relaxation and acceleration techniques, such as the conjugate gradient acceleration, help to alleviate the convergence problem [29, 30], but their consideration is beyond the scope of this paper and will be discussed in a subsequent paper.

The proofs of all results are collected in section 8.

In this paper, an uppercase letter C, if unspecified, denotes an absolute constant which may be different at different occurrences.

2. Preliminaries.

2.1. The sampling set Γ. Every γ in the sampling set Γ is used as the location of a (non-)ideal sampling acquisition device, which has the average sampling characteristic ψ_γ. Then reasonable assumptions on the sampling set Γ are that only finitely many such sampling acquisition devices are located in any unit interval, and that the distribution of those devices is almost location invariant. So in this paper, we make the following basic assumption on the sampling set Γ:

(i) *The sampling set Γ is a relatively separated subset of $\mathbf{R}^d$.*

Here, given a subset $X = \{x_j\}$ of $\mathbf{R}^d$, we say that X is *relatively separated* if there exists a positive constant $D(X)$ such that

$$\sum_{x_j \in X} \chi_{x_j + [0,1]^d}(x) \le D(X) \quad \text{for all } x \in \mathbf{R}^d. \tag{2.1}$$

2.2. The average sampler Ψ. We say that a positive function u on $\mathbf{R}^d$ is a *weight* if it is continuous, symmetric, and satisfies $1 = u(0) \le u(x) < \infty$ for

all $x \in \mathbf{R}^d$, and the inequality $u(x+y) \le u(x)v(y), x, y \in \mathbf{R}^d$, holds for another continuous function v on $\mathbf{R}^d$. The model examples of weights convenient for our consideration of the sampling/reconstruction process are the *polynomial weights*

$$u_\alpha(x) = (1+|x|)^\alpha \tag{2.2}$$

with $\alpha \ge 0$, and the *subexponential weights*

$$e_{D,\delta}(x) = \exp(D|x|^\delta) \tag{2.3}$$

with $D > 0$ and $\delta \in (0,1)$.

Given $1 \le p, q \le \infty$, a weight u, a relatively separated subset Γ of $\mathbf{R}^d$, and a family $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ of functions on $\mathbf{R}^d$, we define $\|\Psi\|_{q,p,u}$ by

$$\begin{aligned}\|\Psi\|_{q,p,u} := &\sup_{\gamma\in\Gamma} \big\| \big(\|\psi_\gamma(\cdot)u(\cdot-\gamma)\|_{L^q(k+[0,1]^d)}\big)_{k\in\mathbf{Z}^d}\big\|_{\ell^p(\mathbf{Z}^d)} \\ &+ \sup_{k\in\mathbf{Z}^d} \big\| \big(\|\psi_\gamma(\cdot)u(\cdot-\gamma)\|_{L^q(k+[0,1]^d)}\big)_{\gamma\in\Gamma}\big\|_{\ell^p(\Gamma)},\end{aligned} \tag{2.4}$$

where, as usual, $\|\cdot\|_{L^q(K)}$ denotes the usual L^q norm on the space $L^q(K)$ of all q-integrable functions on a measurable set K, and $\|\cdot\|_{\ell^p(X)}$ (or $\|\cdot\|_{\ell^p}$ for short) is the usual $\ell^p(X)$ norm on the space of all q-summable sequences on the index set X. For $q = p = \infty$, it is obvious that $\|\Psi\|_{q,p,u} < \infty$ if and only if $|\psi_\gamma(x)| \le \|\Psi\|_{q,p,u}(u(x-\gamma))^{-1}$ for all $x \in \mathbf{R}^d$ and $\gamma \in \Gamma$. In general, for the family of functions $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ with $\|\Psi\|_{q,p,u} < \infty$, each function $\psi_\gamma, \gamma \in \Gamma$, is an L^q function "locally" and a weighted L^p function centered at γ "globally." Therefore for each $\gamma \in \Gamma$, the function ψ_γ in the collection $\Psi := \{\psi_\gamma, \gamma \in \Gamma\}$ with $\|\Psi\|_{q,p,u} < \infty$ can be thought of as essentially supported in a neighborhood of $\gamma \in \Gamma$.

For the average sampler $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$, each average sampling functional ψ_γ reflects the characteristic of the nonideal acquisition device at the location $\gamma \in \Gamma$, and hence it should be essentially supported in a neighborhood of the sampling location γ. So we make the following basic assumption on the average sampler Ψ:

(ii) *The average sampler* $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ *satisfies*

$$\|\Psi\|_{q,p,u} < \infty \tag{2.5}$$

for some $1 \le p, q \le \infty$, *and weight* u.

We interpret any average sampler that satisfies the basic assumption (ii) as having *polynomial* (*subexponential*) *decay*, due to the interpretation of the collection Ψ of average sampling functional $\psi_\gamma, \gamma \in \Gamma$, with $\|\Psi\|_{q,p,u} < \infty$ and the model assumption on the weight u that is convenient for our consideration of the sampling/reconstruction process in which u is a polynomial weight u_α or a subexponential weight $e_{D,\delta}$.

Remark 2.1. For adapting to different average (ideal) sampling situations, we add some flexibility to the basic assumption (ii) on the average sampler Ψ with variable exponents p and q and weights u. For instance, we may use $q = 1$ for approximating ideal sampling ($\psi_\gamma \approx \delta_\gamma$; see [6]), $q = 2$ for frame sampling (for instance, $\Psi = \{\phi(\cdot - k), k \in \mathbf{Z}^d\}$ for frame sampling in the shift-invariant space $V_2(\phi)$ generated by ϕ, [7, 13, 39, 52]), $q = \infty$ in local blurring or local averaging (for instance, $\psi_\gamma = h(\cdot-\gamma)$ for some compactly supported phase function h, [5, 25, 59]), and the subexponential weight $e_{D,\delta}$ for oversampling band-limited signals [35].

Remark 2.2. Given $p \in [1,\infty]$, a weight u, and two relatively separated subsets Γ, Γ' of $\mathbf{R}^d$, we define the *matrix algebra* $\mathcal{A}_{p,u}(\Gamma,\Gamma')$ *of Schur class* by

$$\mathcal{A}_{p,u}(\Gamma,\Gamma') := \big\{A := (A(\gamma,\gamma'))_{\gamma\in\Gamma,\gamma'\in\Gamma'} : \|A\|_{\mathcal{A}_{p,u}} < \infty\big\}, \tag{2.6}$$

where

$$\|A\|_{\mathcal{A}_{p,u}} := \sup_{\gamma\in\Gamma} \big\|\big(A(\gamma,\gamma')u(\gamma-\gamma')\big)_{\gamma'\in\Gamma'}\big\|_{\ell^p(\Gamma')} + \sup_{\gamma'\in\Gamma'} \big\|\big(A(\gamma,\gamma')u(\gamma-\gamma')\big)_{\gamma\in\Gamma}\big\|_{\ell^p(\Gamma)}$$

(see, e.g., [32, 37, 51]). Then the basic assumption $\|\Psi\|_{q,p,u} < \infty$ on the average sampler Ψ is characterized by

$$\|\Psi\|_{q,p,u} < \infty \text{ if and only if } (\|\psi_\gamma\|_{L^q(k+[0,1]^d)})_{\gamma\in\Gamma,k\in\mathbf{Z}^d} \in \mathcal{A}_{p,u}(\Gamma,\mathbf{Z}^d). \tag{2.7}$$

For the basic assumption $\|\Psi\|_{q,p,u} < \infty$ with different exponents p, q and weights u, we have the following results, which will be used frequently in the proofs:

$$\|\Psi\|_{q_1,p,u} \le C\|\Psi\|_{q_2,p,u} \tag{2.8}$$

if $q_1 \le q_2$, and

$$\|\Psi\|_{q,p_1,u} \le C\|\Psi\|_{q,p_2,v} \tag{2.9}$$

if $p_1 \le p_2$ and $\|uv^{-1}\|_{L^r} < \infty$, where $1/r = 1/p_1 - 1/p_2$; see [52] for details.

2.3. The space V in which functions are sampled and recovered. The band-limited space $B_\Omega, \Omega > 0$, is a prototypical space for sampling theory and for signal processing in the classical band-limited model [2]. By the Whittaker representation theorem, the band-limited space B_π is spanned by the shifted sinc function $\mathrm{sinc}(x-k) := \frac{\sin\pi(x-k)}{\pi(x-k)}, k \in \mathbf{Z}$, using ℓ^2 coefficients, i.e., $B_\Omega = \{\sum_{k\in\mathbf{Z}} c(k)\mathrm{sinc}(x-k) : (c(k)) \in \ell^2\}$.

Since the sinc function has infinite support and slow decay at infinity, the band-limited space is often unsuitable for numerical implementations (see, e.g., [2, 35]). Hence people consider other models that retain some of the *simplicity* and *structure* of the band-limited model, but are more *amenable* to numerical implementation and are more *flexible* for approximate real data (see [2, 11, 15, 36, 56] and references cited therein).

Other than the band-limited model, a widely used model is the finitely generated shift-invariant model; see, e.g., [1, 2, 4, 5, 6, 19, 28, 41, 53, 54, 55]. Here the finitely generated shift-invariant space $V_q(\phi_1,\dots,\phi_N)$, that has functions $\phi_1,\dots,\phi_N$ on $\mathbf{R}^d$ as its generators, is defined by

$$V_q(\phi_1,\dots,\phi_N) := \left\{ \sum_{n=1}^N \sum_{k\in\mathbf{Z}^d} c_n(k)\phi_n(\cdot-k) : (c_n(k))_{k\in\mathbf{Z}^d} \in \ell^q, 1 \le n \le N \right\}, \tag{2.10}$$

where $1 \le q \le \infty$ (see, e.g., [7, 13, 16, 21, 39] for the applications of finitely generated shift-invariant spaces in wavelet analysis and approximation theory). Clearly the finitely generated shift-invariant space $V_q(\phi_1,\dots,\phi_N)$ becomes the band-limited space B_π if we let $q = 2, N = 1$, and $\phi_1 = \mathrm{sinc}$.

The space $V_q(\Phi,\Lambda)$,

$$V_q(\Phi,\Lambda) := \left\{ \sum_{\lambda\in\Lambda} c(\lambda)\phi_\lambda : \ \big\|(c(\lambda))_{\lambda\in\Lambda}\big\|_{\ell^q(\Lambda)} < \infty \right\}, \tag{2.11}$$

that was recently introduced by the author in [52] is a new model (other than the above band-limited model and shift-invariant model), where $1 \le q \le \infty$, Λ is a relatively separated subset of $\mathbf{R}^d$, and $\Phi = \{\phi_\lambda, \lambda \in \Lambda\}$ satisfies $\|\Phi\|_{q,p,u} < \infty$ for

some $1 \le p, q \le \infty$ and some weight u. We call Φ the *generator* of the space $V_q(\Phi, \Lambda)$, and call Λ the *location of the generator*.

The prototypical space $V_q(\Phi, \Lambda)$ has shift-invariant spaces, twisted shift-invariant spaces generated by (non-)uniform Gabor frame system (or Wilson basis) in the time-frequency analysis (see, e.g., [8, 17, 26, 38, 47] and references cited therein), and spaces of polynomial splines (which are widely used as approximating spaces in data fitting problems and operator-equation problems [15, 36, 48]) as its special cases. Particularly the space $V_q(\Phi, \Lambda)$ and the shift-invariant space $V_q(\phi_1, \dots, \phi_N)$ are related as follows:

$$V_q(\Phi, \Lambda) = V_q(\phi_1, \dots, \phi_N) \tag{2.12}$$

and

$$\|\Phi\|_{q,p,u} < \infty \text{ if and only if } \phi_1, \dots, \phi_N \in W_q(L_{p,u}) \tag{2.13}$$

if we let $\Lambda := \{x_1, \dots, x_N\} + \mathbf{Z}^d$ and $\phi_\lambda := \phi_n(\cdot - k)$ if $\lambda = x_n + k$ for some $k \in \mathbf{Z}^d$, where $\{x_1, \dots, x_N\}$ is a discrete set in $\mathbf{R}^d/\mathbf{Z}^d$; see [52] for details. Here we recall that the *Wiener amalgam space* $W_q(L_{p,u})$, which consists of functions that are "locally" in L^q and "globally" in weighted L^p space with weight u [2], is defined by

$$W_q(L_{p,u}) := \Big\{ f : \ \|f\|_{W_q(L_{p,u})} := \big\| \big(\|fu\|_{L^q(k+[0,1]^d)} \big)_{k \in \mathbf{Z}^d} \big\|_{\ell^p(\mathbf{Z}^d)} < \infty \Big\}. \tag{2.14}$$

The prototypical space $V_q(\Phi, \Lambda)$ is suitable for modeling signals with a finite rate of innovations [23, 34, 40, 42, 44, 45, 46, 58] in, for instance, (i) stream of pulses $\sum_l a_l p(t - t_l)$ found in global positioning system (GPS) applications and cellular radio, where $p(t)$ is the antenna transmit pulse shape; (ii) stream of different pulses $\sum_l a_l p_l(t - t_l)$ found in modeling ultra wide-band, where different incoming paths are subjected to different frequency-selective attenuations; (iii) band-limited signals with additive shot noise $\sum_{k \in \mathbf{Z}} c(k)\mathrm{sinc}(t - k) + \sum_l d(l)\delta(t - t_l)$; (iv) sum of band-limited signals and nonuniform spline signals, convenient for modeling electrocardiogram signals.

The prototypical space $V_q(\Phi, \Lambda)$ retains some of the simplicity and structure of a finitely generated shift-invariant space of the form (2.10), is *amenable* to numerical implementation (see sections 5, 6, and 7), and is more *flexible* for approximating real data than the band-limited model and the shift-invariant model (see [52] for details).

So in this paper, we make the following basic assumption on the space V in which functions are sampled and recovered:

(iii) *The space V is of the form $V_q(\Phi, \Lambda)$, where $1 \le q \le \infty$, Λ is a relatively separated subset of $\mathbf{R}^d$, and $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$ is a family of functions on $\mathbf{R}^d$ satisfying $\|\Phi\|_{q,p,u} < \infty$ for some $1 \le p \le \infty$ and weight u.*

Remark 2.3. Signals in the space $V := V_q(\Phi, \Lambda)$, which satisfies the above basic assumption (iii), have a finite rate of innovation because a signal $f = \sum_{\lambda \in \Lambda} c(\lambda)\phi_\lambda \in V_q(\Phi, \Lambda)$ on a unit interval $t + [-1/2, 1/2)^d$ is essentially determined by the coefficients $c(\lambda)$ with $\lambda \in t + [-1/2, 1/2)^d$ because of the well-localization property of the generator Φ, and the total number of the locations $\lambda \in \Lambda$ on the unit interval $t + [-1/2, 1/2)^d$ is bounded by some constant C_0 independent of the center t of the unit interval due to the relative separatedness of the location Λ of the generator Φ.

Remark 2.4. We provide some flexibility on the assumption $\|\Phi\|_{q,p,u} < \infty$ on the generator Φ of the space $V_q(\Phi, \Lambda)$ for adapting to different modeling situations. For instance, we may use $q = 1$ and $p = \infty$ when modeling slow-varying signals with shot noises [58], $1 \le p, q \le \infty$ when modeling signals in a finitely generated shift-invariant space [2], and $q = \infty$ and $1 \le p \le \infty$ for decomposing a time signal via (non-)uniform Gabor frame system or Wilson basis [26].

2.4. The space $V_q(\Phi,\Lambda)$ for modeling signals with a finite rate of innovations. Let $\delta_{\lambda\lambda'}$ stand for the usual Kronecker symbol. For a Hilbert space H with $E=\{e_\lambda,\lambda\in\Lambda\}$ being its Riesz basis, we say that $E^d=\{e_\lambda^d:\ \lambda\in\Lambda\}\subset H$ is a *dual Riesz basis* of E if E^d is a Riesz basis of H and $\langle e_\lambda,e_{\lambda'}^d\rangle=\delta_{\lambda\lambda'}$ for all $\lambda,\lambda'\in\Lambda$, and we say that $E^o=\{e_\lambda^o:\ \lambda\in\Lambda\}$ is an *orthonormal basis* for H if E^o is a basis of H and $\langle e_\lambda^o,e_{\lambda'}^o\rangle=\delta_{\lambda\lambda'}$ for all $\lambda,\lambda'\in\Lambda$.

Let $1\le p\le\infty$. We say that a weight u is *p-admissible* if there exist a weight v and two positive constants $D:=D(u)\in(0,\infty)$ and $\theta:=\theta(u)\in(0,1)$ such that

$$u(x+y)\le D(u(x)v(y)+v(x)u(y))\quad\text{for all } x,y\in\mathbf{R}^d, \tag{2.15}$$

$$\|(vu^{-1})\|_{L^{p'}}\le D,\ \text{and} \tag{2.16}$$

$$\inf_{\tau>0}\|v\|_{L^1(B(\tau))}+t\|vu^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(\tau))}\le Dt^\theta\quad\text{for all } t\ge1, \tag{2.17}$$

where $p'=p/(p-1)$ and $B(\tau)=\{x\in\mathbf{R}^d:|x|\le\tau\}$. The p-admissibility of a weight u is a technical condition in [51] for establishing the Wiener lemma for matrix algebras of Schur class and of Sjöstrand class; see also Lemmas 8.1 and 8.2. It is verified in [51] that the polynomial weight u_α with $\alpha>d(1-1/p)$ and the subexponential weight $e_{D,\delta}$ with $D>0$ and $\delta\in(0,1)$ are p-admissible weights. The reader may refer to those two model examples for simplification; see also subsection 2.5.

Now we recall some properties of the space $V_q(\Phi,\Lambda)$ in [52]; see, e.g., [2, 39] for similar results for our familiar shift-invariant setting.

PROPOSITION 2.1 (see [52]). *Let $1\le q\le\infty$, $u_\alpha(x):=(1+|x|)^\alpha,\alpha\ge0$, be the polynomial weights, Λ be a relatively separated subset of $\mathbf{R}^d$, $\Phi=\{\phi_\lambda,\lambda\in\Lambda\}$ satisfy $\|\Phi\|_{q,1,u_0}<\infty$, and $V_q(\Phi,\Lambda)$ be defined as in* (2.11). *Then $V_q(\Phi,\Lambda)\subset L^q$. Moreover,*

$$\Big\|\sum_{\lambda\in\Lambda}c(\lambda)\phi_\lambda\Big\|_{L^r}\le C\big\|\big(c(\lambda)\big)_{\lambda\in\Lambda}\big\|_{\ell^r(\Lambda)}\|\Phi\|_{q,1,u_0} \tag{2.18}$$

for every sequence $(c(\lambda))_{\lambda\in\Lambda}\in\ell^r(\Lambda)$ with $1\le r\le q$, and

$$\big\|\big(\langle f,\phi_\lambda\rangle\big)_{\lambda\in\Lambda}\big\|_{\ell^r(\Lambda)}\le C\|f\|_{L^r}\|\Phi\|_{q,1,u_0} \tag{2.19}$$

for all $f\in L^r$ with $q/(q-1)\le r\le\infty$.

PROPOSITION 2.2 (see [52]). *Let $2\le q\le\infty$, $1\le p\le\infty$, u be a p-admissible weight, Λ be a relatively separated subset of $\mathbf{R}^d$, $\Phi=\{\phi_\lambda,\lambda\in\Lambda\}$ be a family of functions on $\mathbf{R}^d$, $V_2(\Phi,\Lambda)$ be as in* (2.11), *and the frame operator S on $V_2(\Phi,\Lambda)$ be defined by*

$$Sf=\sum_{\lambda\in\Lambda}\langle f,\phi_\lambda\rangle\phi_\lambda,\quad f\in V_2(\Phi,\Lambda). \tag{2.20}$$

Assume that Φ is a Riesz basis for $V_2(\Phi,\Lambda)$ and satisfies $\|\Phi\|_{q,p,u}<\infty$. Then $S^{-1}\Phi:=\{S^{-1}\phi_\lambda:\lambda\in\Lambda\}$ is a dual Riesz basis of Φ, $S^{-1/2}\Phi:=\{S^{-1/2}\phi_\lambda:\lambda\in\Lambda\}$ is an orthonormal basis of $V_2(\Phi,\Lambda)$, and $\|S^{-1}\Phi\|_{q,p,u}+\|S^{-1/2}\Phi\|_{q,p,u}<\infty$. If we further assume that Λ is a lattice, and that the generator Φ of the space $V_2(\Phi,\Lambda)$ is enveloped by a function in the Wiener amalgam space $W_\infty(L_{p,u})$ (hence $\|\Phi\|_{q,p,u}\le\|\Phi\|_{\infty,p,u}\le C\|h\|_{W_\infty(L_{p,u})}<\infty$), i.e., $|\phi_\lambda(x)|\le h(x-\lambda)$ for some $h\in W_\infty(L_{p,u})$, then so are dual Riesz basis $S^{-1}\Phi$ and orthonormal basis $S^{-1/2}\Phi$.

Remark 2.5. Let $1 \le p, r \le \infty$; we say that a weight u is (p,r)-*admissible* (or we say that $w(x,y) := u(x-y)$ is a (p,r)-admissible translation-invariant weight [51]) if there exist a weight v and two positive constants $D \in (0,\infty)$ and $\theta \in (0,1)$ such that (2.15), (2.16), and

$$\inf_{\tau>0} \|v\|_{L^{r'}(B(\tau))} + t\|vu^{-1}\|_{L^{p'}(\mathbf{R}^d \setminus B(\tau))} \le Dt^{\theta} \quad \text{for all } t \ge 1 \tag{2.21}$$

hold, where $p' = p/(p-1)$ and $r' = r/(r-1)$. Clearly the p-admissibility of a weight agrees with its (p,∞)-admissibility. Since for any weight v it holds that $\|v\|_{L^{r'}(B(\tau))} \le C\|v\|_{L^1(B(\tau))}$ for all $\tau \ge 1$, we then conclude that p-admissibility of a weight implies its (p,r)-admissibility for any $1 \le r \le \infty$.

2.5. Model. The reader may consider the following model for simplification:

(i) The generator $\Phi := \{\phi_\lambda, \lambda \in \Lambda\}$ of the space $V_r(\Phi,\Lambda)$ is enveloped by some function g in the Wiener amalgam space $W_q(L_{p,u})$ with $2 \le q \le \infty$, i.e.,

$$|\phi_\lambda(x)| \le g(x-\lambda) \quad \text{for all } x \in \mathbf{R}^d \text{ and } \lambda \in \Lambda. \tag{2.22}$$

(The above envelopment assumption (2.22) for the generator Φ implies that the basic assumption (iii) $\|\Phi\|_{q,p,u} < \infty$ for the generator Φ is satisfied. The converse is true in the shift-invariant setting, $V_r(\Phi,\Lambda) = V_r(\phi_1,\dots,\phi_N)$; particularly, the above envelopment property for the generator Φ of the space $V_r(\Phi,\Lambda)$ is equivalent to the basic assumption (iii) $\|\Phi\|_{q,p,u} < \infty$, and also equivalent to the property that the generators $\phi_1,\dots,\phi_N$ of the space $V_2(\phi_1,\dots,\phi_N)$ belong to the Wiener amalgam space $W_q(L_{p,u})$. The envelopment assumption (2.22) for the generator Φ is not satisfied when the space $V_q(\Phi,\Lambda)$ is used for modeling slow-varying signals with shot noises [42].)

(ii) The average sampler $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ is enveloped by some function h in the Wiener amalgam space $W_{q^*}(L_{p,u})$ with $q/(q-1) \le q^* \le \infty$, i.e.,

$$|\psi_\gamma(x)| \le h(x-\gamma) \quad \text{for all } x \in \mathbf{R}^d \text{ and } \gamma \in \Gamma. \tag{2.23}$$

(The above envelopment assumption (2.23) for the average sampler Ψ implies that the basic assumption (ii) $\|\Psi\|_{q^*,p,u} < \infty$ for the average sampler Ψ is satisfied. The above envelopment assumption (2.23) for the average sampler Ψ is not satisfied when it is a family of approximating delta functionals with variable width [28].)

(iii) The weight u in the above envelopment assumptions on the generator Φ and the average sampler Ψ is the polynomial weight $u_\alpha(x) := (1+|x|)^\alpha$ with $\alpha > d(1-1/p)$ or the exponential weight $e_{D,\delta}(x) := \exp(D|x|^\delta)$ with $D > 0$ and $\delta \in (0,1)$.

Remark 2.6. Given an exponent $p \in [1,\infty]$, a weight u, and two relatively separated subsets Γ, Γ' of $\mathbf{R}^d$, we define the *matrix algebra* $\mathcal{C}_{p,u}(\Gamma,\Gamma')$ *of Sjöstrand class* by

$$\mathcal{C}_{p,u}(\Gamma,\Gamma') := \big\{A := (A(\gamma,\gamma'))_{\gamma,\gamma'\in\Gamma},\ \|A\|_{\mathcal{C}_{p,u}} < \infty\big\}, \tag{2.24}$$

where

$$\|A\|_{\mathcal{C}_{p,u}} := \|(Au)_*(k))_{k\in\mathbf{Z}^d}\|_{\ell^p}$$

and

$$(Au)_*(k) = \sup_{\gamma\in m+[-1/2,1/2)^d, \gamma'\in m+k+[-1/2,1/2)^d, m\in\mathbf{Z}^d} |A(\gamma,\gamma')u(\gamma-\gamma')|,\ k \in \mathbf{Z}^d$$

(see, e.g., [8, 10, 27, 49, 51]). For the case when $q = \infty, 1 \le p \le \infty$, and Γ is a lattice, the envelopment property (2.23) for the average sampler $\Psi := \{\psi_\gamma,\ \gamma \in \Gamma\}$ is characterized by $(\|\psi_\gamma\|_{L^q(\mu+[0,1]^d)})_{\gamma,\mu\in\Gamma} \in \mathcal{C}_{p,u}(\Gamma,\Gamma)$. Here a subset X of $\mathbf{R}^d$ is said to be a *lattice* if $1 \le \sum_{x_j\in X} \chi_{x_j+[0,1]^d}(x) \le D(X)$ for all $x \in \mathbf{R}^d$ and some $D(X) < \infty$.

3. Well-localized displayer. Let $1 \le q \le \infty$ and V be a subspace of L^q. We say that Γ, a subset of $\mathbf{R}^d$, is a *stable ideal sampling set* for the space V if there exist two positive constants A, B such that

$$A\|f\|_{L^q} \le \|(f(\gamma))_{\gamma\in\Gamma}\|_{\ell^q(\Gamma)} \le B\|f\|_{L^q} \quad \text{for all } f \in V, \tag{3.1}$$

and that $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$, a family of average sampling functionals, is a *stable averaging sampler* for the space V if there exist two positive constants A', B' such that

$$A'\|f\|_{L^q} \le \|(\langle f, \psi_\gamma\rangle)_{\gamma\in\Gamma}\|_{\ell^q(\Gamma)} \le B'\|f\|_{L^q} \quad \text{for all } f \in V \tag{3.2}$$

(see [5]). From the above definitions of a stable ideal sampling set Γ and a stable average sampler Ψ, we have that any function $f \in V$ can be reconstructed uniquely and stably from its samples $\{f(\gamma) :\ \gamma \in \Gamma\}$ if Γ is a stable ideal sampling set for V, and similarly that any function $f \in V$ can be reconstructed uniquely and stably from its average samples $\{\langle f, \psi_\gamma\rangle :\ \gamma \in \Gamma\}$ if Ψ is a stable averaging sampler for V.

For average (ideal) sampling on the space $V_2(\Phi, \Lambda)$, derived from a general theorem for localized frames (see [8, Theorem 1], [24, Theorem 3.6], and [32, Theorem 13]), we have the following well-localization results for its displayer (Theorems 3.1 and 3.2); see Remark 3.1 for a more general formulation of the well-localization of a displayer. From those results, it concludes that the displayer $\tilde\Psi$ associated with a stable (ideal) averaging sampler Ψ has the same polynomial (subexponential) decay when both the average sampler Ψ and the generator Φ for the space $V_2(\Phi, \Lambda)$ have polynomial (subexponential) decay.

THEOREM 3.1. *Let $2 \le q \le \infty, q/(q-1) \le q^* \le \infty, 1 \le r \le q, 1 \le p \le \infty$, u be a p-admissible weight, the subsets Λ, Γ of $\mathbf{R}^d$ be relatively separated, the generator $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$ satisfy*

$$\|\Phi\|_{q,p,u} < \infty, \tag{3.3}$$

the average sampler $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ satisfy

$$\|\Psi\|_{q^*,p,u} < \infty, \tag{3.4}$$

and the space $V_r(\Phi, \Lambda)$ be as in (2.11). Assume that Φ is a Riesz basis of $V_2(\Phi, \Lambda)$ and that Ψ is a stable averaging sampler for $V_2(\Phi, \Lambda) \subset L^2$. Then there exists a displayer $\tilde\Psi = \{\tilde\psi_\gamma : \gamma \in \Gamma\} \subset V_1(\Phi, \Lambda)$ such that

$$\|\tilde\Psi\|_{q,p,u} < \infty, \tag{3.5}$$

and

$$f = \sum_{\gamma\in\Gamma} \langle f, \psi_\gamma\rangle \tilde\psi_\gamma \quad \text{for all } f \in V_r(\Phi, \Lambda). \tag{3.6}$$

If we further assume that Λ is a lattice and that the generator Φ and the average sampler Ψ are enveloped by some functions in the Wiener amalgam space $W_\infty(L_{p,u})$, then so is the displayer $\tilde\Psi$.

THEOREM 3.2. *Let $1 \le p, r \le \infty$, u be a p-admissible weight, the subsets Λ, Γ of $\mathbf{R}^d$ be relatively separated, $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$ be a family of continuous functions on $\mathbf{R}^d$ that satisfies*

$$\|\Phi\|_{\infty,p,u} < \infty, \tag{3.7}$$

and the space $V_r(\Phi, \Lambda)$ be as in (2.11). *Assume that Φ is a Riesz basis of $V_2(\Phi, \Lambda)$, and that Γ is a stable ideal sampling set for $V_2(\Phi, \Lambda) \subset L^2$. Then there exists a displayer $\tilde{\Psi} = \{\tilde{\psi}_\gamma : \gamma \in \Gamma\}$ such that $\tilde{\Psi} \subset V_1(\Phi, \Lambda)$, $\|\tilde{\Psi}\|_{\infty,p,u} < \infty$, and*

$$f = \sum_{\gamma\in\Gamma} f(\gamma)\tilde{\psi}_\gamma \quad \text{for all } f \in V_r(\Phi, \Lambda). \tag{3.8}$$

If we further assume that Γ is a lattice and that Φ is enveloped by a function in the Wiener amalgam space $W_\infty(L_{p,u})$, then so is the displayer $\tilde{\Psi}$.

Remark 3.1. For the well-localization of the displayer $\tilde{\Psi}$, the following general principle can be derived from theorems for localized frames (see [8, Theorem 1], [24, Theorem 3.6], and [32, Theorem 13]): Let Λ and Γ be two index sets, and let the families $\mathcal{A}(\Lambda) = \{(a(\lambda, \lambda'))_{\lambda,\lambda'\in\Lambda}\}$ and $\mathcal{A}(\Gamma, \Lambda) = \{(a(\gamma, \lambda))_{\gamma\in\Gamma,\lambda\in\Lambda}\}$ of infinite matrices have the following algebraic properties that (i) $\mathcal{A}(\Lambda)$ is an inverse-closed matrix algebra in $\mathcal{B}(\ell^2(\Lambda))$ (the space of all bounded operators on $\ell^2(\Lambda)$), (ii) $A^T A \in \mathcal{A}(\Lambda)$ for any $A \in \mathcal{A}(\Gamma, \Lambda)$, and (iii) $AB \in \mathcal{A}(\Gamma, \Lambda)$ for any $A \in \mathcal{A}(\Gamma, \Lambda)$ and $B \in \mathcal{A}(\Lambda)$. Then on the space $V_2(\Phi, \Lambda)$ whose generator $\Phi := (\phi_\lambda)_{\lambda\in\Lambda}$ is a frame, there exists an $\mathcal{A}(\Gamma, \Lambda)$-localized displayer $\tilde{\Psi} := (\tilde{\psi}_\gamma)_{\gamma\in\Gamma}$ associated with a stable average sampling processing

$$V_2(\Phi, \Lambda) \ni f \longmapsto \langle f, \Psi\rangle \in \ell^2(\Gamma) \tag{3.9}$$

whose average sampler $\Psi := (\psi_\gamma)_{\gamma\in\Gamma}$ is $\mathcal{A}(\Gamma, \Lambda)$-localized. Here we say that $\Psi := (\psi_\gamma)_{\gamma\in\Gamma}$ is $\mathcal{A}(\Gamma, \Lambda)$-localized on $V_2(\Phi, \Lambda)$ if $A_{\Psi,\Phi} := (\langle\psi_\gamma, \phi_\lambda\rangle)_{\gamma\in\Gamma,\lambda\in\Lambda} \in \mathcal{A}(\Gamma, \Lambda)$ and $A_{\Psi,\tilde{\Phi}} := (\langle\psi_\gamma, \tilde{\phi}_\lambda\rangle)_{\gamma\in\Gamma,\lambda\in\Lambda} \in \mathcal{A}(\Gamma, \Lambda)$, where $\tilde{\Phi} = (\tilde{\phi}_\lambda)_{\lambda\in\Lambda}$ is the dual frame generator $\tilde{\Phi}$ associated with the frame Φ. Thus the well-localization for the displayer in Theorems 3.1 and 3.2 becomes essentially a concrete example of the above general principle (particularly in Theorems 3.1 and 3.2, the Schur class $\mathcal{A}_{p,u}(\Gamma, \Lambda)$ and the Sjöstrand class $\mathcal{C}_{p,u}(\Gamma, \Lambda)$ are used as $\mathcal{A}(\Gamma, \Lambda)$ in the above principle, and $\mathcal{A}_{p,u}(\Lambda, \Lambda)$ and $\mathcal{C}_{p,u}(\Lambda, \Lambda)$ as $\mathcal{A}(\Lambda)$). The well-localization of the displayer in the above principle and the localization of the dual frame in the theory of frames are equivalent since for the case when $\Psi \in V_2(\Phi, \Lambda)$ (otherwise replacing Ψ with the projection $P\Psi$ of Ψ on $V_2(\Phi, \Lambda)$; see Remark 3.2 below) the stability of the average sampling process (3.9) is equivalent to the frame property for Ψ, and the displayer $A_{\Psi,\Phi}(A_{\Psi,\Phi}^T A_{\Psi,\Phi})^\dagger\Phi$ associated with the average sampling processing (3.9) is the canonical dual frame associated with the frame Ψ, where $(A_{\Psi,\Phi}^T A_{\Psi,\Phi})^\dagger$ is the pseudo-inverse of the matrix $A_{\Psi,\Phi}^T A_{\Psi,\Phi}$. Therefore the above principle for the well-localization of a displayer is established (i) in [8, Theorem 1] for the case when $\Psi = \Phi$, $\Gamma = \Lambda$, $\mathcal{A}(\Lambda, \Lambda)$, and $\mathcal{A}(\Lambda)$ are the Sjöstrand class $\mathcal{C}_{1,0}(\Lambda, \Lambda)$ defined by

$$\mathcal{C}_{1,0}(\Lambda, \Lambda) = \left\{(c(\lambda, \lambda')_{\lambda,\lambda'\in\Lambda} : \left(\sup_{a(\lambda)-a(\lambda')=g\in G} |c(\lambda, \lambda')|\right)_{g\in G} \in \ell^1(G)\right\},$$

where G is an additive discrete group of the form $B_0\mathbf{Z}^d \times (\mathbf{Z}^{d'}/B_1\mathbf{Z}^{d'})$ for some nonsingular diagonal matrices B_0 and B_1, and $a : \Lambda \to G$ is a map with $\sup_{g\in G} \#\{\lambda \in \Lambda : \ a(\lambda) = g\} < \infty$; (ii) in [24, Theorem 3.6] for the case when $\Psi = \Phi$, $\Gamma = \Lambda$, $\mathcal{A}(\Lambda, \Lambda)$ and $\mathcal{A}(\Lambda)$ are a solid, inverse-closed, involute Banach algebra; and (iii) in [32, Theorem 13] for the case that $\Lambda = \mathbf{Z}^d$, Γ is a relative separated subset of $\mathbf{R}^d$, $\mathcal{A}(\Lambda) = \mathcal{A}$, and $\mathcal{A}(\Gamma, \Lambda) = \{(a(\gamma, k))_{\gamma\in\Gamma,k\in\mathbf{Z}^d} : \ (\tilde{a}(m, n))_{m,n\in\mathbf{Z}^d} \in \mathcal{A}\}$, where $\mathcal{A}$ is a solid, inverse-closed, involute Banach algebra, and

$$\tilde{a}(m, n) = \begin{cases} \sup_{\gamma\in(m+[0,1)^d)\cap\Gamma} |a(\gamma, n)| & \text{if } (m + [0, 1)^d) \cap \Gamma \neq \emptyset, \\ 0 & \text{if } (m + [0, 1)^d) \cap \Gamma = \emptyset. \end{cases}$$

The above principle for the well-localization of the displayer can be derived from [32, Theorem 13] with identical proof.

Remark 3.2. Let $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$ and $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ be as in either Theorem 3.1 or Theorem 3.2, and let $S^{-1}\Phi = \{S^{-1}\phi_\lambda : \lambda \in \Lambda\}$ be the dual Riesz basis for the space $V_2(\Phi, \Lambda)$ in Proposition 2.2. By Propositions 2.1 and 2.2, the operator P defined by $Pf := \sum_{\lambda\in\Lambda}\langle f, S^{-1}\phi_\lambda\rangle\phi_\lambda$ is a projection operator from L^2 to $V_2(\Phi, \Lambda)$. We can extend the domain of the projection operator P so that $P\psi_\gamma$ is well defined for every sampling functional ψ_γ in the average sampling case and for the delta functional δ_γ in the ideal sampling case. Moreover, $\langle f, \psi_\gamma\rangle = \langle f, P\psi_\gamma\rangle$ for all $\gamma \in \Gamma$ and $f \in V_2(\Phi, \Lambda)$. We then have that if Ψ is a stable average sampler, then $P\Psi = \{P\psi_\gamma : \gamma \in \Gamma\}$ is a frame of $V_2(\Phi, \Lambda)$; that is, there exist two positive constants $A, B > 0$ such that

$$A\|f\|_2 \le \|(\langle f, P\psi_\gamma\rangle)_{\gamma\in\Gamma}\|_{\ell^2(\Gamma)} \le B\|f\|_2 \quad \text{for all } f \in V_2(\Phi, \Lambda). \tag{3.10}$$

So by the general frame theory, a displayer $\tilde{\Psi} \subset V_2(\Phi, \Lambda)$, which *may or may not* have polynomial (subexponential) decay, can be constructed, while we show in Theorems 3.1 and 3.2 that a displayer $\tilde{\Psi}$ can be constructed to has polynomial (subexponential) decay whenever the generator Φ and the average sampler Ψ have it. The reader may refer to [2] and references cited therein for the connection among average (ideal) sampling, reproducing kernel Hilbert space, and frame theory in the shift-invariant setting.

By Theorems 3.1 and 3.2, we have the following corollary for the uniform sampling in the familiar shift-invariant space.

COROLLARY 3.3. *Let $1 \le p \le \infty, 2 \le q \le \infty, 1 \le r \le q, q/(q-1) \le q^* \le \infty$, u be the polynomial weight u_α with $\alpha > d(1 - 1/p)$ or the subexponential weight $e_{D,\delta}$ with $D > 0$ and $\delta \in (0, 1)$, and $\phi_1, \dots, \phi_N \in W_q(L_{p,u})$, and $\Phi = \{\phi_n(\cdot - k) : \ 1 \le n \le N, k \in \mathbf{Z}^d\}$ be a Riesz basis of $V_2(\phi_1, \dots, \phi_N)$. Then we have that*

(i) *if $\psi_1, \dots, \psi_L \in W_{q^*}(L_{p,u})$, and $\Psi := \{\psi_l(\cdot-k) : 1 \le l \le L, k \in \mathbf{Z}^d\}$ is a stable average sampler for $V_2(\phi_1, \dots, \phi_N)$, then there exist functions $\tilde{\psi}_1, \dots, \tilde{\psi}_L \in W_q(L_{p,u})$ such that $f = \sum_{l=1}^L \sum_{k\in\mathbf{Z}^d}\langle f, \psi_l(\cdot - k)\rangle\tilde{\psi}_l(\cdot - k)$ holds for all $f \in V_r(\phi_1, \dots, \phi_N)$.*

(ii) *if $\phi_1, \dots, \phi_N$ are continuous functions in $W_\infty(L_{p,u})$, and $X_0 + \mathbf{Z}^d$ is a stable ideal sampling set for $V_2(\phi_1, \dots, \phi_N)$, where $X_0 = \{x_1, \dots, x_L\} \subset [0, 1)^d$, then there exist continuous functions $\tilde{\psi}_1, \dots, \tilde{\psi}_L \in W_\infty(L_{p,u})$ such that $f = \sum_{l=1}^L \sum_{k\in\mathbf{Z}^d} f(x_l + k)\tilde{\psi}_l(\cdot - k)$ holds for all $f \in V_r(\phi_1, \dots, \phi_N)$.*

Remark 3.3. For the shift-invariant setting $V_2(\Phi, \Lambda) = V_2(\phi_1, \dots, \phi_N)$ for some functions $\phi_1, \dots, \phi_N$, it is shown in [28] that if the generator $\phi_1, \dots, \phi_N$ satisfies $|\phi_n(x)| \le C_0(1 + |x|)^{-\alpha}$ for all $1 \le n \le N$ and $x \in \mathbf{R}^d$, (i.e., $\|\Phi\|_{\infty,\infty,u_\alpha} < \infty$, where $\Phi := \{\phi_n(\cdot - k), \ 1 \le n \le N, k \in \mathbf{Z}^d\}$), and if the average sampling functional

ψ_γ is supported in $\gamma + [-a, a]$ and $\|\psi_\gamma\|_{L^1} \le C_0$ for some positive constants C_0 and a (which implies that $\|\Psi\|_{1,\infty,u} < \infty$ for any weight u, where $\Psi := \{\psi_\gamma,\ \gamma \in \Gamma\}$), then the corresponding displayer $\tilde{\Psi} := \{\tilde{\psi}_\gamma : \gamma \in \Gamma\}$ can be chosen to satisfy $|\tilde{\psi}_\gamma(x)| \le C_1(1 + |x - \gamma|)^{-\alpha}$ for some positive constant C_1 (i.e., $\|\tilde{\Psi}\|_{\infty,\infty,u_\alpha} < \infty$). A similar result for the ideal sampling process is also established in [28]. The above results for the average (ideal) sampling process follow from Theorems 3.1 and 3.2 with $p = q = \infty, q^* = 1, u = u_\alpha$, and $V_2(\Phi, \Lambda)$ being a finitely generated shift-invariant space. Other than those exponents p, q, q^* and weight u mentioned above, the results in Theorems 3.1 and 3.2 are new even for our familiar shift-invariant setting.

By Theorems 3.1 and 3.2, we have the following result for the stability of the average sampler Ψ in the space $V_r(\Phi, \Lambda) \subset L^r$ with $r \ne 2$, which can also derived from [24, Theorem 2.7] and [28, Theorem 10].

COROLLARY 3.4. *Let $2 \le q \le \infty, 1 \le p \le \infty$, u be a p-admissible weight, and the subsets Λ, Γ of $\mathbf{R}^d$ be relatively separated. Then we have that*

(i) *if Φ is a Riesz basis of $V_2(\Phi, \Lambda)$ and satisfies* (3.3)*, and if Ψ is a stable averaging sampler for $V_2(\Phi, \Lambda) \subset L^2$ and satisfies* (3.4) *for some $q/(q-1) \le q^* \le \infty$, then the stable average sampler Ψ for $V_2(\Phi, \Lambda)$ is also a stable averaging sampler for $V_r(\Phi, \Lambda) \subset L^r$ for all $q^*/(q^*-1) \le r \le q$.*

(ii) *if the family $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$ of continuous functions on $\mathbf{R}^d$ is a Riesz basis of $V_2(\Phi, \Lambda)$ and satisfies* (3.7)*, and if Γ is a stable ideal sampling set for $V_2(\Phi, \Lambda)$, then the ideal sampling set Γ for $V_2(\Phi, \Lambda)$ is also a stable ideal sampling set for $V_r(\Phi, \Lambda) \subset L^r$ for all $1 \le r \le \infty$.*

Remark 3.4. By Corollary 3.4, the stability of the average (ideal) sampler for $V_2(\Phi, \Lambda) \subset L^2$ implies the stability of the average (ideal) sampler for $V_r(\Phi, \Lambda) \subset L^r$ with $r \ne 2$. Such an implication can also be derived from [24, Theorem 2.7] and [28, Theorem 10], where the formulation of localized frames are used. (The author thank the anonymous referee for pointing out that derivation.) The above implication is observed in [22, 28] for the shift-invariant setting under slightly stronger assumptions on the average sampler Ψ, the generator Φ, and the nonuniform sampling set Γ than the ones in Corollary 3.4. As for the case when the average sampler and the generator are identical and when the grid Λ and the sampling set Γ are $\mathbf{Z}^d$, the above implication has long been known (and even the converse is also true); see, for instance, [7, 39].

4. Stability of the average (ideal) sampling/reconstruction process. For a matrix $A := (A(\lambda, \lambda'))_{\lambda \in \Lambda, \lambda \in \Lambda'}$, we define its transpose A^* by $A^* := (\overline{A(\lambda, \lambda')})_{\lambda' \in \Lambda', \lambda \in \Lambda}$. For a space $V_2(\Phi, \Lambda)$ generated by $\Phi := \{\phi_\lambda, \lambda \in \Lambda\}$, an average sampling on the space $V_2(\Phi, \Lambda)$ with the average sampler $\Psi := \{\psi_\gamma, \gamma \in \Gamma\}$, and an ideal sampling on the space $V_2(\Phi, \Lambda)$ with the sampling set Γ, we define the subspace H of ℓ^2 by

$$H = \{(\langle f, \phi_\lambda \rangle)_{\lambda \in \Lambda} :\ f \in V_2(\Phi, \Lambda)\}, \tag{4.1}$$

and two Gram matrices $A_{\Psi,\Phi}$ and $A_{\delta_\Gamma,\Phi}$ by

$$A_{\Psi,\Phi} = \big(\langle \psi_\gamma, \phi_\lambda \rangle\big)_{\gamma \in \Gamma, \lambda \in \Lambda} \tag{4.2}$$

and

$$A_{\delta_\Gamma,\Phi} = \big(\phi_\lambda(\gamma)\big)_{\gamma \in \Gamma, \lambda \in \Lambda}. \tag{4.3}$$

THEOREM 4.1. *Let $2 \le q \le \infty, q/(q-1) \le q^* \le \infty; q^*/(q^*-1) \le r \le q; 1 \le p \le \infty$; u be a p-admissible weight; the subsets Λ, Γ of $\mathbf{R}^d$ be relatively separated; the*

generator $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$ *satisfy* $\|\Phi\|_{q,p,u} < \infty$*; the average sampler* $\Psi = \{\psi_\gamma : \gamma \in \Gamma\}$ *satisfy* $\|\Psi\|_{q^*,p,u} < \infty$*; and the space* $V_r(\Phi, \Lambda)$*, the subspace* H *of* ℓ^2*, and the matrix* $A_{\Psi,\Phi}$ *be as in* (2.11), (4.1)*, and* (4.2)*, respectively. Assume that* Φ *is a frame of* $V_2(\Phi, \Lambda)$*. Then* Ψ *is a stable averaging sampler for* $V_2(\Phi, \Lambda) \subset L^2$ *if and only if there exists a positive constant* C *such that*

$$C^{-1}\|c\|^2_{\ell^2(\Lambda)} \le \langle (A_{\Psi,\Phi})^* A_{\Psi,\Phi} c, c\rangle \le C\|c\|^2_{\ell^2(\Lambda)} \text{ for all } c \in H. \tag{4.4}$$

THEOREM 4.2. *Let* $1 \le p, r \le \infty$*;* u *be a* p*-admissible weight; the subsets* Λ, Γ *of* $\mathbf{R}^d$ *be relatively separated;* $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$ *be a family of continuous functions on* $\mathbf{R}^d$ *that satisfies* (3.7)*; and the space* $V_r(\Phi, \Lambda)$*, the space* H*, and the matrix* $A_{\delta_\Gamma,\Phi}$ *be as in* (2.11)*,* (4.1)*, and* (4.3)*, respectively. Assume that* Φ *is a frame of* $V_2(\Phi, \Lambda)$*. Then* Γ *is a stable ideal sampling set for* $V_2(\Phi, \Lambda)$ *if and only if there exists a positive constant* C *such that*

$$C^{-1}\|c\|^2_{\ell^2(\Lambda)} \le \langle (A_{\delta_\Gamma,\Phi})^* A_{\delta_\Gamma,\Phi} c, c\rangle \le C\|c\|^2_{\ell^2(\Lambda)} \text{ for all } c \in H. \tag{4.5}$$

Remark 4.1. For the uniform average sampling on finitely generated shift-invariant spaces, that is, $V_2(\Phi, \Lambda) = V_2(\phi_1, \dots, \phi_N)$ and $\Psi = \{\psi_l(\cdot - k), 1 \le l \le L, k \in \mathbf{Z}^d\}$ for some functions $\phi_n, 1 \le n \le N$, and $\psi_l, 1 \le l \le L$, the matrix $(A_{\Psi,\Phi})^* A_{\Psi,\Phi}$ in (4.4) can be written as

$$(A_{\Psi,\Phi})^* A_{\Psi,\Phi} = \left(\sum_{l=1}^{N} \sum_{j \in \mathbf{Z}^d} \langle \phi_n(\cdot - k), \psi_l(\cdot - j)\rangle \langle \psi_l(\cdot - j), \phi_{n'}(\cdot - k')\rangle \right)_{(n,k),(n',k') \in \{1,\dots,N\} \times \mathbf{Z}^d}.$$

Due to the shift-invariant structure of the matrix $(A_{\Psi,\Phi})^* A_{\Psi,\Phi}$, we may use the Fourier technique to interpret (4.4) in Theorem 4.1 as

$$C^{-1} G(\xi) \le A_{as}(\xi) \le C G(\xi) \text{ a.e. } \xi \in \mathbf{R}^d, \tag{4.6}$$

where the Fourier transform $\hat{f}$ of an integrable function f is defined by $\hat{f}(\xi) = \int_{\mathbf{R}^d} f(x) e^{-ix\xi} dx$, and the $N \times N$ matrices $G(\xi)$ and $A_{as}(\xi)$ are defined by

$$G(\xi) = \left(\sum_{k \in \mathbf{Z}^d} \widehat{\phi}_n(\xi + 2k\pi) \overline{\widehat{\phi}_{n'}(\xi + 2k\pi)} \right)_{1 \le n, n' \le N}$$

and

$$A_{as}(\xi) = \sum_{l=1}^{L} \sum_{k,k' \in \mathbf{Z}^d} \left(\widehat{\phi}_n(\xi + 2k\pi) \overline{\widehat{\psi}_l(\xi + 2k\pi)} \widehat{\psi}_l(\xi + 2k'\pi) \overline{\widehat{\phi}_{n'}(\xi + 2k'\pi)} \right)_{1 \le n, n' \le N}.$$

The above characterization (4.6) of the stable averaging sampler Ψ is given in [5] under weaker assumptions on the generator $\phi_1, \dots, \phi_N$ and the average sampler $\psi_1, \dots, \psi_L$ than the ones for the generator Φ and the average sampler Ψ in Theorem 4.1.

Remark 4.2. For the uniform ideal sampling on a single generated shift-invariant space, that is, $V_2(\Phi, \Lambda) = V_2(\phi)$ and $\Gamma = \mathbf{Z}^d$, the matrix $(A_{\delta_\Gamma,\Phi})^* A_{\delta_\Gamma,\Phi}$ in (4.5) can be

written as $(A_{\delta_\Gamma,\Phi})^* A_{\delta_\Gamma,\Phi} = (\sum_{j\in\mathbf{Z}^d} \phi(j-k)\phi(j-k'))_{k,k'\in\mathbf{Z}^d}$. Similar to the uniform average sampling case, we may use the Fourier technique to interpret (4.5) in Theorem 4.2 as follows: $C^{-1} \le |\sum_{k\in\mathbf{Z}^d} \widehat{\phi}(\xi+2k\pi)| \le C$ for almost all $\xi \in \mathbf{R}^d$, which was given in [60].

Remark 4.3. For the characterization of the stable average sampler and stable ideal sampling set for various spaces, there is an extensive literature (see, for instance, the recent review papers [2, 57] and monographs [12, 14, 43] for ideal sampling, and [1, 2, 4, 5, 6, 28, 53, 54, 55] for average sampling).

5. Robustness of the reconstruction process. For the numerical stability of the reconstruction formulas (3.6) and (3.8) when the average (ideal) sampling data and the displayer are corrupted, for instance, by the noise in the measurement, we have the following result.

THEOREM 5.1. *Let $p, q, q^*, r, u, \Lambda, \Gamma, \Phi, \Psi, \tilde{\Psi}$ be as either in Theorem* 3.1 *or in Theorem* 3.2. *Assume that the original function f belongs to $V_r(\Phi,\Lambda)$, and that $G = \{g_\gamma : \gamma\in\Gamma\}$ and $\tilde{\Psi}' = \{\tilde{\psi}'_\gamma : \gamma\in\Gamma\}$ are the corrupted average sampling data and displayer, respectively. Then there exists a positive constant C (independent of f, G, and $\tilde{\Psi}'$) such that*

$$\begin{aligned}\left\| f - \sum_{\gamma\in\Gamma} g_\gamma \tilde{\psi}'_\gamma \right\|_r &\le C\|(g_\gamma - \langle f, \psi_\gamma\rangle)_{\gamma\in\Gamma}\|_{\ell^r(\Gamma)} \|\tilde{\Psi}\|_{q,p,u} \\ &\quad + C\|\tilde{\Psi} - \tilde{\Psi}'\|_{q,p,u} \|(g_\gamma)_{\gamma\in\Gamma}\|_{\ell^r(\Gamma)}. \end{aligned} \tag{5.1}$$

For the numerical stability of the reconstruction processes (3.6) and (3.8) when there is certain perturbation for the average sampler and for the ideal sampling set, we have the following results.

THEOREM 5.2. *Let $p, q, q^*, u, \Lambda, \Gamma, \Phi, \Psi, \tilde{\Psi}, V_q(\Phi,\Lambda)$ be as in Theorem* 3.1. *Then there exist a sufficiently small positive number δ_0 and a positive constant C such that any average sampler $\Psi' = \{\psi'_\gamma : \gamma\in\Gamma\}$ with the property that*

$$\|\Psi' - \Psi\|_{q^*,p,u} \le \delta_0 \tag{5.2}$$

is a stable average sampler for the space $V_2(\Phi,\Lambda)$, and the corresponding displayer $\tilde{\Psi}'$ satisfies

$$\|\tilde{\Psi}' - \tilde{\Psi}\|_{q^*,p,u} \le C\|\Psi' - \Psi\|_{q^*,p,u}. \tag{5.3}$$

THEOREM 5.3. *Let $p, u, \Lambda, \Gamma, \Phi, \tilde{\Psi}, V_q(\Phi,\Lambda)$ be as in Theorem* 3.2. *Assume that $\Phi_\delta = \{\phi_{\gamma,\delta} : \gamma\in\Gamma\}$ satisfies*

$$\|\Phi_\delta\|_{\infty,p,u} \to 0 \text{ as } \delta\to 0, \tag{5.4}$$

where $\phi_{\gamma,\delta}(x) = \sup_{|t|\le\delta} |\phi_\gamma(x+t) - \phi_\gamma(x)|$. Then there exist a sufficiently small positive number δ_0 and a positive constant C such that any sampling set $\Gamma' := \{\gamma+\delta_\gamma : \gamma\in\Gamma\}$ with $\sup_{\gamma\in\Gamma}|\delta_\gamma| \le \delta_0$ is a stable ideal sampling set for the space $V_2(\Phi,\Lambda)$, and the corresponding displayer $\tilde{\Psi}'$ satisfies $\|\tilde{\Psi}' - \tilde{\Psi}\|_{\infty,p,u} \le C\|\Phi_{\delta_0}\|_{\infty,p,u}$.

By Theorems 5.2 and 5.3, we have the following results about perturbation for nonuniform average (ideal) sampling on a finitely generated shift-invariant space.

COROLLARY 5.4. *Let $1 \le p \le \infty, 2 \le q \le \infty, q/(q-1) \le q^* \le \infty$; u be the polynomial weight u_α with $\alpha > d(1-1/p)$ or the subexponential weight $e_{D,\delta}$ with*

$D > 0$ and $\delta \in (0,1)$, $\phi_1, \ldots, \phi_N \in W_q(L_{p,u})$; Γ be a relatively separated subset of $\mathbf{R}^d$; and $\Phi := \{\phi_n(\cdot - k),\ \ 1 \le n \le N, k \in \mathbf{Z}^d\}$ be a Riesz basis of the shift-invariant space $V_2(\phi_1, \ldots, \phi_N)$. Then

(i) *if $\psi_1, \ldots, \psi_L \in W_{q^*}(L_{p,u})$, and $\Psi := \{\psi_l(\cdot - \gamma),\ 1 \le l \le L, \gamma \in \Gamma\}$ is a stable average sampler for $V_2(\phi_1, \ldots, \phi_N)$, then there exists a positive constant δ_0 such that for any functions $\theta_1, \ldots, \theta_L$ with $\sum_{l=1}^{L} \|\psi_l - \theta_l\|_{W_{q^*}(L_{p,u})} \le \delta_0$, their generating average sampler $\Theta := \{\theta_l(\cdot - \gamma),\ 1 \le l \le L, \gamma \in \Gamma\}$ is a stable average sampler for $V_2(\phi_1, \ldots, \phi_N)$.*

(ii) *if $\phi_1, \ldots, \phi_N$ are continuous functions in $W_\infty(L_{p,u})$, and Γ is a stable ideal sampling set for $V_2(\phi_1, \ldots, \phi_N)$, then there exists a positive constant δ_0 such that any relatively separated set $\tilde{\Gamma} = \{\gamma + \delta_\gamma : \gamma \in \Gamma\}$ with $\sup_{\gamma \in \Gamma} |\delta_\gamma| \le \delta_0$ is a stable ideal sampling set for $V_2(\phi_1, \ldots, \phi_N)$.*

Remark 5.1. From Corollary 5.4, we see that for the shift-invariant setting, the stability of the nonuniform average (ideal) sampling is preserved under small perturbation. Such a phenomenon is observed in [19, 41, 57] for the ideal sampling process in the band-limited spaces and the finite-generated shift-invariant spaces, and in [4] for the average sampling process in the finitely generated shift-invariant spaces. We use a different approach than the ones in [4, 19, 41, 57] to consider the perturbation problem, and then the stability under small perturbation is shown to be preserved under weak assumptions on the generator Φ and the average sampler Ψ; see, for instance, [4] in which such a preservation is established only for the case when $q = \infty, q^* = 1, p = 1$, and $u = u_0$.

Remark 5.2. Unlike in the shift-invariant setting, the stability of the ideal sampling set is not preserved under small perturbation in our general setting, or in other words, assumption (5.4) in Theorem 5.3 cannot be eliminated if we expect that the stability of the ideal sampling set is preserved under small perturbation. For instance, let $\Lambda = \mathbf{Z}$ and $\Phi := \{\phi_k(x) := h(2(x-k))\cos^2(4k\pi(x-k)) : k \in \mathbf{Z}\}$, where $h(x) = \max(1 - |x|, 0)$ is the hat function. For the space $V_2(\Phi, \mathbf{Z})$ generated by that family of functions Φ, we see that $f(k) = c(k), k \in \mathbf{Z}$, for any $f = \sum_{k \in \mathbf{Z}} c(k)\phi_k \in V_2(\Phi, \mathbf{Z})$. Thus $\|f\|_2 \le \|(f(k))_{k \in \mathbf{Z}}\|_{\ell^2(\mathbf{Z})} \le 4\|f\|_2$ for all $f \in V_2(\Phi, \mathbf{Z})$, where we have also used the facts that $\phi_k, k \in \mathbf{Z}$, are supported in $k + [-1/2, 1/2]$ and satisfy $\frac{1}{4} \le \|\phi_k\|_2 \le 1$. This shows that $\mathbf{Z}$ is a stable sampling set for $V_2(\Phi, \mathbf{Z})$. Noting that $f(k + 1/(8k)) = 0, k \in \mathbf{Z}$, for any $f = \sum_{k \in \mathbf{Z}} c(k)\phi_k \in V_2(\Phi, \mathbf{Z})$, we then conclude that for *any* $0 < \delta < 1/2$, the small perturbation $\mathbf{Z}_\delta = \{k + (-1)^k \min(\delta, 1/(8|k|)),\ 0 \ne k \in \mathbf{Z}\}$ of the stable sampling set $\mathbf{Z}$ is *not* a stable sampling set for $V_2(\Phi, \Lambda)$. Moreover, assumption (5.4) does not hold for that family of functions Φ, since for any $\delta > 0$,

$$\|\Phi_\delta\|_{q,p,\alpha} \ge \|\phi_k - \phi_k\left(\cdot + \frac{1}{8k}\right)\|_{L^q(k+[0,1])} \ge \frac{1}{2}\int_0^{1/4} |\cos 8k\pi x| dx \ge \frac{1}{16},$$

where the integer k is chosen so that $8k\delta \ge 1$.

6. Locally finite reconstruction process. For a bounded set K and a positive number R, we let $B(K, R) := \{y \in \mathbf{R}^d : \inf_{x \in K} |y - x| \le R\}$ be the R-neighborhood of the set K. For an average sampling process on the space $V_r(\Phi, \Lambda)$ with the average sampler $\Psi := \{\psi_\gamma, \gamma \in \Gamma\}$, we define the *locally finite reconstruction approximation* of a function $f \in V_r(\Phi, \Lambda)$ on a bounded set K using average sampling data on the R-neighborhood $B(K, R)$ of the set K by

$$\tilde{f}^a_{K,R} = \sum_{\gamma \in \Gamma \cap B(K,R)} \langle f, \psi_\gamma \rangle \tilde{\psi}^a_{\gamma,K,R}, \tag{6.1}$$

where

$$\tilde{\psi}^a_{\gamma,K,R} = \sum_{\lambda_1,\lambda_2\in B(K,2R)} \langle \psi_\gamma, \phi_{\lambda_1}\rangle (\tilde{A}_{\Psi,\Phi,K,R})^{-1}(\lambda_1,\lambda_2)\phi_{\lambda_2} \tag{6.2}$$

and

$$\tilde{A}_{\Psi,\Phi,K,R} = \left(\sum_{\gamma\in\Gamma\cap B(K,4R)} \langle \phi_\lambda, \psi_\gamma\rangle\langle \psi_\gamma, \phi_{\lambda'}\rangle \right)_{\lambda,\lambda'\in B(K,3R)}. \tag{6.3}$$

Similarly for the ideal sampling process on the space $V_r(\Phi,\Lambda)$ with the sampling set Γ, we define the locally finite reconstruction approximation of a function $f\in V_r(\Phi,\Lambda)$ on a bounded set K using the ideal sampling data on the R-neighborhood $B(K,R)$ of the set K by

$$\tilde{f}^i_{K,R} = \sum_{\gamma\in\Gamma\cap B(K,R)} f(\gamma)\tilde{\psi}^i_{\gamma,K,R}, \tag{6.4}$$

where

$$\tilde{\psi}^i_{\gamma,K,R} = \sum_{\lambda_1,\lambda_2\in B(K,2R)} \phi_{\lambda_1}(\gamma)(\tilde{A}_{\delta_\Gamma,\Phi,K,R})^{-1}(\lambda_1,\lambda_2)\phi_{\lambda_2} \tag{6.5}$$

and

$$\tilde{A}_{\Psi,\Phi,K,R} = \left(\sum_{\gamma\in\Gamma\cap B(K,4R)} \phi_\lambda(\gamma)\phi_{\lambda'}(\gamma) \right)_{\lambda,\lambda'\in B(K,3R)}. \tag{6.6}$$

For any bounded set K, we observe that the locally finite reconstruction approximation $\tilde{f}^a_{K,R}$ for the average sampling/reconstruction process and $\tilde{f}^i_{K,R}$ for the ideal sampling/reconstruction process are obtained by using the samples in a finite neighborhood of that set K with finitely many steps. Then we conclude from Theorems 6.1 and 6.2 that the locally finite reconstruction approximation could be possibly used in the real-time reconstruction by selecting the parameter R properly.

Using an idea similar to the finite section method in frame theory (see, e.g., [18, 20, 35]), we have the following locally finite reconstruction approximation for the average (ideal) sampling/reconstruction process.

THEOREM 6.1. *Let $p,q,q^*,r,u,\Lambda,\Gamma,\Phi,\Psi,\tilde{\Psi}$ be as in Theorem* 3.1*, and let $\tilde{f}^a_{K,R}$ be defined as in* (6.1) *for any bounded set K, positive number $R\ge 1$, and function $f\in V_r(\Phi,\Lambda)$. Then there exists a positive constant C (independent of the bounded set K, the positive number $R\ge 1$, and the function $f\in V_r(\Phi,\Lambda)$) such that*

$$\begin{aligned}\|\tilde{f}^a_{K,R} - f\|_{L^r(K)} &\le C\|u^{-1}\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|(\langle f,\psi_\gamma\rangle)_{\gamma\in\Gamma}\|_{\ell^r(\Gamma)}\\ &\quad + C\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|(\langle f,\psi_\gamma\rangle)_{\gamma\in\Gamma\setminus B(K,R)}\|_{\ell^r(\Gamma\setminus B(K,R))}\end{aligned} \tag{6.7}$$

holds for any bounded set K, any positive number $R\ge 1$, and any $f\in V_r(\Phi,\Lambda)$.

THEOREM 6.2. *Let $p,r,u,\Lambda,\Gamma,\Phi,\tilde{\Psi}$ be as in Theorem* 3.2*, and let $\tilde{f}^i_{K,R}$ be defined as in* (6.4) *for any bounded set K, positive number $R\ge 1$, and function $f\in V_r(\Phi,\Lambda)$. Then there exists a positive constant C (independent of the bounded set K, the positive number $R\ge 1$, and the function $f\in V_r(\Phi,\Lambda)$) such that*

$$\begin{aligned}\|\tilde{f}^i_{K,R} - f\|_{L^r(K)} &\le C\|u^{-1}\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|(f(\gamma))_{\gamma\in\Gamma}\|_{\ell^r(\Gamma)}\\ &\quad + C\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|(f(\gamma))_{\gamma\in\Gamma\setminus B(K,R)}\|_{\ell^r(\Gamma\setminus B(K,R))}\end{aligned} \tag{6.8}$$

holds for any bounded set K, any positive number $R \geq 1$, and any $f \in V_r(\Phi, \Lambda)$.

As a by-product of Theorems 6.1 and 6.2, we have the following result about the location of (non-)ideal sampling devices.

THEOREM 6.3. *Let $p, q, q^*, r, u, \Lambda, \Gamma, \Phi, \Psi, \tilde{\Psi}$ be as in either Theorem* 3.1 *or Theorem* 3.2. *Then there exists a positive constant R_0 such that for any bounded set K, the number of (non-)ideal sampling devices located in $B(K, R_0)$ exceeds the degrees of freedom of the space $V_2(\Phi, \Lambda)$ in the domain K, that is,*

$$\#(\Gamma \cap B(K, R_0)) \geq \#(\Lambda \cap K), \tag{6.9}$$

where $\#(E)$ denotes the cardinality of a set E.

Remark 6.1. If $V_2(\Phi, \Lambda)$ is a shift-invariant space generated by a compactly supported continuous function ϕ, then one may verify that the matrix $\big(\sum_{\gamma\in\Gamma} \phi_\lambda(\gamma)\phi_{\lambda'}(\gamma)\big)$ associated with the ideal sampling on Γ is a band-limited matrix. Furthermore for the case when finite truncation of that matrix is invertible (which is true if ϕ is a B-spline and Γ is a sampling set with sampling density strictly less than the optimal density; see [3, 4]), Gröchenig and Schwab [33] proposed an efficient local reconstruction algorithm to recover the original function in a domain *exactly*, instead of *approximately* as in Theorems 6.1 and 6.2, from its samples in a neighborhood of that domain. Comparing with the locally perfect recovery in [33], we see that the locally finite approximation in Theorems 6.1 and 6.2 works for the well-localized average sampling process as well as the ideal sampling process, and for most signals with a finite rate of innovation instead of signals of B-spline type.

Remark 6.2. The density property (6.9) in Corollary 6.3 is established in [3] for the ideal sampling of signals in a B-spline space; that is, $V_2(\Phi, \Lambda)$ is the shift-invariant space generated by the integer shifts of a B-spline. The reader may refer to [2] for similar results for average (ideal) sampling in the band-limited space, and to [8, 9] for the nonuniform Gabor system.

7. The Richardson–Landweber iterative reconstruction process. Let Φ be the generator of the space $V_2(\Phi, \Lambda)$, Ψ be an average sampler, and Γ be an ideal sampling set. We define an iterative reconstruction algorithm from average sampling data $(a_\gamma)_{\gamma\in\Gamma} \in \ell^r$ by

$$\begin{cases} f_0 = A^{-2} \sum_{\gamma\in\Gamma,\lambda\in\Lambda} a_\gamma \langle \psi_\gamma, \phi_\lambda\rangle \phi_\lambda, \\ f_n = f_0 + f_{n-1} - A^{-2} T_{as} f_{n-1} \text{ if } n \geq 1, \end{cases} \tag{7.1}$$

where A is a positive parameter and the operator T_{as} is defined by

$$T_{as} f = \sum_{\gamma\in\Gamma,\lambda\in\Lambda} \langle f, \psi_\gamma\rangle \langle \psi_\gamma, \phi_\lambda\rangle \phi_\lambda. \tag{7.2}$$

Similarly we define an iterative reconstruction algorithm from ideal sampling data $(a_\gamma)_{\gamma\in\Gamma} \in \ell^r$ by

$$\begin{cases} f_0 = A^{-2} \sum_{\gamma\in\Gamma,\lambda\in\Lambda} a_\gamma \phi_\lambda(\gamma) \phi_\lambda, \\ f_n = f_0 + f_{n-1} - A^{-2} T_{is} f_{n-1}, \ n \geq 1, \end{cases} \tag{7.3}$$

where A is a positive parameter and the operator T_{is} is defined by

$$T_{is} f = \sum_{\gamma\in\Gamma,\lambda\in\Lambda} f(\gamma) \phi_\lambda(\gamma) \phi_\lambda. \tag{7.4}$$

Since the operators A_{as} in (7.2) and A_{is} in (7.4) can be written as $T_{as} = T_\Phi A^*_{\Psi,\Phi} A_{\Psi,\Phi}(T_\Phi)^{-1}$ and $T_{is} = T_\Phi A^*_{\delta_\Gamma,\Phi} A_{\delta_\Gamma,\Phi}(T_\Phi)^{-1}$, where $T_\Phi : (c(\lambda)) \longmapsto \sum_{\lambda\in\Lambda} c(\lambda)\phi_\lambda$, we then have that the iterative reconstruction algorithms from average sampling data and from ideal sampling data are equivalent to the familiar Richardson–Landweber iterative algorithms for the positive operators $A^*_{\Psi,\Phi}A_{\Psi,\Phi}$ and $A^*_{\delta_\Gamma,\Phi}A_{\delta_\Gamma,\Phi}$ on $\ell^2(\Lambda)$, respectively. For the iterative reconstruction algorithms from average sampling data and from ideal sampling data, we have the following exponential convergence for signals in $V_r(\Phi,\Lambda)$.

THEOREM 7.1. *Let $p, q, q^*, r, u, \Lambda, \Gamma, \Phi, \Psi, V_r(\Phi,\Lambda)$ be as in Theorem* 3.1. *We assume that the parameter A in* (7.1) *is a positive constant larger than the operator norm of the matrix $A_{\Psi,\Phi}$ from $\ell^2(\Lambda)$ to $\ell^2(\Gamma)$. Then the sequence $\{f_n, n \ge 1\}$ in* (7.1) *converges to a function f_∞ in $V_r(\Phi,\Lambda)$ in the L^r norm exponentially for any initial data $(a_\gamma)_{\gamma\in\Gamma} \in \ell^r$; that is, there exist two positive constants $C \in (0,\infty)$ and $s \in (0,1)$ such that*

$$\|f_n - f_\infty\|_r \le Cs^n \|(a_\gamma)_{\gamma\in\Gamma}\|_{\ell^r(\Gamma)} \quad \text{for all } n \ge 1. \tag{7.5}$$

Furthermore if the initial data $(a_\gamma)_{\gamma\in\Gamma}$ are obtained from average sampling a function f in $V_r(\Phi,\Lambda)$, that is, $a_\gamma = \langle f, \psi_\gamma\rangle$ for all $\gamma \in \Gamma$, then the limit function f_∞ of the sequence $\{f_n, n \ge 1\}$ agrees with the original function f.

THEOREM 7.2. *Let $p, r, u, \Lambda, \Gamma, \Phi, V_q(\Phi,\Lambda)$ be as in Theorem* 3.2. *Assume that the parameter A in* (7.3) *is a positive constant larger than the operator norm of the matrix $A_{\delta_\Gamma,\Phi} := (\phi_\lambda(\gamma))_{\gamma\in\Gamma,\lambda\in\Lambda}$ from $\ell^2(\Lambda)$ to $\ell^2(\Gamma)$. Then the sequence $\{f_n, n \ge 1\}$ in* (7.3) *converges to a function $f_\infty \in V_r(\Phi,\Lambda)$ in the L^r norm exponentially. Moreover if $a_\gamma = f(\gamma), \gamma \in \Gamma$, holds for some $f \in V_r(\Phi,\Lambda)$, then $f_\infty = f$.*

Remark 7.1. For $\Phi = \{\phi_\lambda : \lambda \in \Lambda\}$, we let P be the projection operator from L^2 to $V_2(\Phi,\Lambda)$; see Remark 3.2. Then for the average sampling/reconstruction process with Ψ as its average sampler (resp., the ideal sampling/reconstruction process with Γ as its sampling set), $P\Psi := \{P\psi_\gamma,\ \psi_\gamma \in \Psi\}$ (resp., $P\delta_\Gamma := \{P\delta_\gamma,\ \gamma \in \Gamma\}$) is a frame for $V_2(\Phi,\Lambda)$, and hence the corresponding frame algorithm is the familiar Richardson–Landweber iterative algorithm for the positive operator $(A_{\Phi,\Phi})^{-1/2}A^*_{\Psi,\Phi}A_{\Psi,\Phi}(A_{\Phi,\Phi})^{-1/2}$ (resp., $(A_{\Phi,\Phi})^{-1/2}A^*_{\delta_\Gamma,\Phi}A_{\delta_\Gamma,\Phi}(A_{\Phi,\Phi})^{-1/2}$) on the space $\ell^2(\Lambda)$; see [2]. Clearly the iterative frame algorithm becomes the iterative algorithm proposed in this article when Φ is an orthonormal basis of $V_2(\Phi,\Lambda)$. In general, we need more computation for each iterative step of the iterative frame algorithm than that of the iterative algorithm in the paper. The consideration of which iterative algorithm converges faster, and of other implementations of the reconstruction process associated with the average (ideal) sampling/reconstruction process, is beyond the scope of this paper and will be discussed in a subsequent paper. By the general frame theory, the iterative frame algorithm associated with the average (ideal) sampling/reconstruction process on the space $V_2(\Phi,\Lambda) \subset L^2$ converges exponentially. Similar to the proofs of Theorems 7.1 and 7.2, we have the exponential convergence of the iterative frame algorithm on the space $V_r(\Phi,\Lambda)$ with $r \ne 2$, under the assumption that $p, q, q^*, r, u, \Lambda, \Lambda, \Phi, \Psi$ are as in Theorems 3.1 or 3.2. The above exponential convergence theorem for the iterative frame algorithm is established in [2, 22] for the shift-invariant setting with some minor additional assumptions on the exponents p, q, q^* and weight u.

8. Proofs. In this section, we collect the proofs of all theorems and corollaries stated in Sections 3, 4, 5, 6, and 7.

8.1. Proof of Theorem 3.1. Unlike in the study of average (ideal) sampling signals in a shift-invariant space, the main obstacle when considering well-localization of the displayer comes from the nongroup structure on the generator Φ and the average sampler Ψ, which makes the standard approach from Fourier analysis inapplicable. In the proof of Theorem 3.1, we will use the procedure used in the study of localized frames (see [8, 24, 31, 32, 37, 50, 51, 52] and references cited therein) with some nonessential modification (see also Remark 3.1). For the completeness of this paper, we include a complete proof.

For a matrix $A = (a(\lambda, \lambda'))_{\lambda,\lambda' \in \Lambda}$, we denote by $\|A\|_{\mathcal{B}^2}$ its operator norm on $\ell^2(\Lambda)$. To prove Theorem 3.1, we recall some properties of the matrix algebras of the Schur class $\mathcal{A}_{p,u}(\Lambda, \Lambda')$ and of the Sjöstrand class $\mathcal{C}_{p,u}(\Lambda, \Lambda)$ in [51]. The third property in Lemmas 8.1 and 8.2 below is usually known as Wiener's lemma; see, for instance, [8, 10, 24, 27, 31, 32, 37, 49, 50, 51] and references cited therein for its recent development and various applications.

LEMMA 8.1 (see [51]). *Let $1 \le p \le \tilde{p} \le \infty$, u and $\tilde{u}$ be weights, and $\Lambda, \Lambda', \Lambda''$ be relatively separated subsets of $\mathbf{R}^d$. Then the following statements are true:*

(i) *If $\|u\tilde{u}^{-1}\|_{L^r} < \infty$, where $r = p\tilde{p}/(\tilde{p} - p)$, then*

$$\|A\|_{\mathcal{A}_{p,u}} \le C\|A\|_{\mathcal{A}_{\tilde{p},\tilde{u}}} \quad \text{for all } A \in \mathcal{A}_{\tilde{p},\tilde{u}}(\Lambda, \Lambda'). \tag{8.1}$$

(ii) *If there exists another weight v such that (2.15) and (2.16) hold, then*

$$\|AB\|_{\mathcal{A}_{p,u}} \le C\|A\|_{\mathcal{A}_{p,u}}\|B\|_{\mathcal{A}_{p,u}} \tag{8.2}$$

for all $A \in \mathcal{A}_{p,u}(\Lambda, \Lambda')$ and $B \in \mathcal{A}_{p,w}(\Lambda', \Lambda'')$, where

$$AB := \left(\sum_{\lambda' \in \Lambda'} A(\lambda, \lambda')B(\lambda', \lambda'') \right)_{\lambda \in \Lambda, \lambda'' \in \Lambda''}.$$

(iii) *If u is a p-admissible weight and A is a matrix in $\mathcal{A}_{p,u}(\Lambda, \Lambda)$ satisfying*

$$\|Ac\|_{\ell^2} \ge D_0\|c\|_{\ell^2} \text{ for all } c \in \ell^2 \tag{8.3}$$

for some positive constant D_0, then the inverse A^{-1} of the matrix A belongs to $\mathcal{A}_{p,u}(\Lambda, \Lambda)$.

(iv) *If u is a p-admissible weight, then there exist positive constants $C_1, C_2 \in (0, \infty)$ and $\theta \in (0, 1)$ such that the following estimate holds for all $A \in \mathcal{A}_{p,u}(\Lambda, \Lambda)$:*

$$\|A^n\|_{\mathcal{A}_{p,u}} \le \left(C_1 \frac{\|A\|_{\mathcal{A}_{p,u}}}{\|A\|_{\mathcal{B}^2}} \right)^{C_2 n^\theta} \|A\|_{\mathcal{B}^2}^n, \ n \ge 1. \tag{8.4}$$

LEMMA 8.2 (see [51]). *Let $1 \le p \le \tilde{p} \le \infty$, u and $\tilde{u}$ be weights on $\mathbf{R}^d$, and Λ be a lattice of $\mathbf{R}^d$. Then the following statements are true:*

(i) *If $\|u\tilde{u}^{-1}\|_{L^r} < \infty$, where $r = p\tilde{p}/(\tilde{p} - p)$, then*

$$\|A\|_{\mathcal{C}_{p,u}} \le C\|A\|_{\mathcal{C}_{\tilde{p},\tilde{u}}} \quad \text{for all } A \in \mathcal{C}_{\tilde{p},\tilde{u}}(\Lambda, \Lambda). \tag{8.5}$$

(ii) *If there exists another weight v such that (2.15) and (2.16) hold, then*

$$\|AB\|_{\mathcal{C}_{p,u}} \le C\|A\|_{\mathcal{C}_{p,u}}\|B\|_{\mathcal{C}_{p,u}} \quad \text{for all } A, B \in \mathcal{C}_{p,u}(\Lambda, \Lambda). \tag{8.6}$$

(iii) *If u is a p-admissible weight on $\mathbf{R}^d$ and if A is a matrix in $\mathcal{C}_{p,u}(\Lambda,\Lambda)$ that satisfies* (8.3)*, then the inverse A^{-1} of the matrix A belong to $\mathcal{C}_{p,u}(\Lambda,\Lambda)$.*

Now we are ready to start the proof of Theorem 3.1.

Proof of Theorem 3.1. By Proposition 2.2, without loss of generality, we may assume that Φ is an orthonormal basis of $V_2(\Phi,\Lambda)$ for otherwise replacing Φ with $S^{-1/2}\Phi$. Therefore

$$\big\|c\big\|_{\ell^2(\Lambda)} = \bigg\|\sum_{\lambda\in\Lambda} c(\lambda)\phi_\lambda\bigg\|_2 \quad \text{for all } c := \big(c(\lambda)\big)_{\lambda\in\Lambda} \in \ell^2(\Lambda) \tag{8.7}$$

and

$$f = \sum_{\lambda\in\Lambda} \langle f,\phi_\lambda\rangle\phi_\lambda \quad \text{for all } f \in V_2(\Phi,\Lambda). \tag{8.8}$$

From (2.8), (3.3), and (3.4), it follows that $\|\Psi\|_{q/(q-1),p,u} + \|\Phi\|_{q,p,u} < \infty$. This, together with (2.7), Lemma 8.1, and the trivial estimate

$$|A_{\Psi,\Phi}(\gamma,\lambda)| \le \sum_{k\in\mathbf{Z}^d} \|\psi_\gamma\|_{L^{q/(q-1)}(k+[0,1]^d)} \|\phi_\lambda\|_{L^q(k+[0,1)^d)},$$

proves that the matrix $A_{\Psi,\Phi} = (A_{\Psi,\Phi}(\gamma,\lambda))_{\gamma\in\Gamma,\lambda\in\Lambda}$ in (4.2) belongs to $\mathcal{A}_{p,u}(\Gamma,\Lambda)$:

$$A_{\Psi,\Phi} \in \mathcal{A}_{p,u}(\Gamma,\Lambda). \tag{8.9}$$

Furthermore, there exists a positive constant C such that

$$\|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}} \le C\|\Psi\|_{q/(q-1),p,u}\|\Phi\|_{q,p,u}. \tag{8.10}$$

Clearly the transpose A^* of a matrix $A \in \mathcal{A}_{p,u}$ has the same $\mathcal{A}_{p,u}$ norm as the one of the matrix A,

$$\|A^*\|_{\mathcal{A}_{p,u}} = \|A\|_{\mathcal{A}_{p,u}}. \tag{8.11}$$

Combining (8.2), (8.9), and (8.11) then yields

$$A^*_{\Psi,\Phi}A_{\Psi,\Phi} \in \mathcal{A}_{p,u}(\Lambda,\Lambda). \tag{8.12}$$

For the matrix $A_{\Psi,\Phi}$, we obtain from (8.7) and the stable assumption on the averaging sampler Ψ that

$$C^{-1}\|c\|_2 \le \bigg\|\bigg(\sum_{\lambda\in\Lambda} A_{\Psi,\Phi}(\gamma,\lambda)c(\lambda)\bigg)_{\gamma\in\Gamma}\bigg\|_{\ell^2(\Gamma)} \le C\|c\|_2 \tag{8.13}$$

for any $\ell^2(\Lambda)$-sequence $c = (c(\lambda))_{\lambda\in\Lambda}$.

Combining (8.2), (8.12), and (8.13), and applying Lemma 8.1 to the matrix $A^*_{\Psi,\Phi}A_{\Psi,\Phi}$, we conclude that

$$(A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1} \in \mathcal{A}_{p,u}(\Lambda,\Lambda). \tag{8.14}$$

Let

$$R := A_{\Psi,\Phi}(A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1}, \tag{8.15}$$

write $R = (R(\gamma,\lambda))_{\gamma\in\Gamma,\lambda\in\Lambda}$, and define $\tilde{\Psi} := \{\tilde{\psi}_\gamma : \gamma \in \Gamma\}$ by

$$\tilde{\psi}_\gamma = \sum_{\lambda\in\Lambda} R(\gamma,\lambda)\phi_\lambda, \quad \gamma \in \Gamma. \tag{8.16}$$

Now we prove that $\tilde{\Psi}$ satisfies all requirements of the displayer associated with the average sampler Ψ. By (8.2), (8.9), and (8.14), we have

$$R \in \mathcal{A}_{p,u}(\Gamma,\Lambda). \tag{8.17}$$

This implies that the sequence $(R(\gamma,\lambda))_{\lambda\in\Lambda} \in \ell^1(\Lambda)$ for any $\gamma \in \Gamma$, and hence $\tilde{\Psi} \subset V_1(\Phi,\Lambda)$.

From (2.7), (8.2), (8.17), and the trivial estimate for $\tilde{\Psi}$: $\|\tilde{\psi}_\gamma\|_{L^q(k+[0,1]^d)} \le \sum_{\lambda\in\Lambda}|R(\gamma,\lambda)|\|\psi_\lambda\|_{L^q(k+[0,1]^d)}$, we have

$$\|\tilde{\Psi}\|_{q,p,u} \le C\|R\|_{\mathcal{A}_{p,u}}\|\Phi\|_{q,p,u} < \infty. \tag{8.18}$$

For any $f \in V_2(\Phi,\lambda)$, it follows from (8.8), (8.15), and (8.17) that

$$\begin{aligned}
&\left(\left\langle \sum_{\gamma\in\Gamma}\langle f,\psi_\gamma\rangle\tilde{\psi}_\gamma, \phi_\lambda\right\rangle\right)_{\lambda\in\Lambda}\\
&\quad= \left(\sum_{\lambda_1\in\Lambda}\langle f,\phi_{\lambda_1}\rangle \times \sum_{\lambda_2\in\Lambda}\langle\phi_{\lambda_2},\phi_\lambda\rangle\left(\sum_{\gamma\in\Gamma}A^*_{\Psi,\Phi}(\lambda_1,\gamma)R(\gamma,\lambda_2)\right)\right)_{\lambda\in\Lambda}\\
&\quad= \left(\sum_{\lambda_1\in\Lambda}\langle f,\phi_{\lambda_1}\rangle\langle\phi_{\lambda_1},\phi_\lambda\rangle\right)_{\lambda\in\Lambda} = \big(\langle f,\phi_\lambda\rangle\big)_{\lambda\in\Lambda}.
\end{aligned} \tag{8.19}$$

This proves the reconstruction formula (3.6) for $r = 2$,

$$f = \sum_{\gamma\in\Gamma}\langle f,\psi_\gamma\rangle\tilde{\psi}_\gamma \quad \text{for any } f \in V_2(\Phi,\Lambda). \tag{8.20}$$

For $1 \le r < \infty$, we obtain from (2.8), (2.9), (8.2), (8.9), (8.17), and Proposition 2.1 that

$$\begin{aligned}
&\left\|\sum_{\gamma\in\Gamma}\sum_{\lambda,\lambda'\in\Lambda}|c(\lambda)| \times |\langle\phi_\lambda,\psi_\gamma\rangle| \times |R(\gamma,\lambda')| \times |\phi_{\lambda'}|\right\|_{L^r}\\
&\quad\le C\|c\|_{\ell^r(\Lambda)}\|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}}\|R\|_{\mathcal{A}_{p,u}}\|\Phi\|_{r,p,u}\\
&\quad\le C\|c\|_{\ell^r(\Lambda)}\|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}}\|R\|_{\mathcal{A}_{p,u}}\|\Phi\|_{q,p,u} < \infty
\end{aligned} \tag{8.21}$$

for any sequence $c := (c(\lambda))_{\lambda\in\Lambda} \in \ell^r(\Lambda)$. Then for $1 \le r < \infty$, the reconstruction formula (3.6) follows easily from (8.20), (8.21), and the density of $\ell^2 \cap \ell^r$ in ℓ^r.

For $r=\infty$, we have that $p=\infty$. Take $c=(c(\lambda))_{\lambda\in\Lambda}\in\ell^\infty(\Lambda)$. We let $f=\sum_{\lambda\in\Lambda}c(\lambda)\phi_\lambda$ and $f_{k,\tau}=\sum_{|\lambda-k|\le\tau}c(\lambda)\phi_\lambda$ for $\tau\ge 1$. Then there exists a positive constant C (independent of $k\in\mathbf{Z}^d$ and $\tau\ge 1$) such that

$$\begin{aligned}\sup_{x\in k+[0,1]^d}|f(x)-f_{k,\tau}(x)|&\le\sum_{|\gamma-k|>\tau}|c(\lambda)|\sup_{x\in k+[0,1]^d}|\phi_\lambda(x)|\\&\le C\|c\|_{\ell^\infty(\Lambda)}\|\Phi\|_{\infty,p,u}\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(\tau))}\end{aligned}$$

and

$$\begin{aligned}&\sup_{x\in k+[0,1]^d}\Big|\sum_{\gamma\in\Gamma}\langle f-f_{k,\tau},\psi_\gamma\rangle\tilde{\psi}_\gamma(x)\Big|\\&\le\|c\|_{\ell^\infty}\Bigg(\sum_{|\gamma-k|\ge\tau/2,\gamma\in\Gamma}\sum_{\lambda\in\Lambda}+\sum_{\gamma\in\Gamma}\sum_{|\lambda-\gamma|\ge\tau/2,\lambda\in\Lambda}\Bigg)|\langle\phi_\lambda,\psi_\gamma\rangle|\sup_{x\in k+[0,1]^d}|\tilde{\psi}_\gamma(x)|\\&\le C\|c\|_{\ell^\infty(\Lambda)}\|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}}\|\tilde{\Psi}\|_{\infty,p,u}\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(\tau/2))}.\end{aligned}$$

The above two estimates, together with (8.20), lead to

$$\begin{aligned}&\sup_{x\in k+[0,1]^d}\Big|f(x)-\sum_{\gamma\in\Gamma}\langle f,\psi_\gamma\rangle\tilde{\psi}_\gamma(x)\Big|\\&=\sup_{x\in k+[0,1]^d}\Big|(f-f_{k,\tau})(x)-\sum_{\gamma\in\Gamma}\langle f-f_{k,\tau},\psi_\gamma\rangle\tilde{\psi}_\gamma(x)\Big|\\&\le C\|c\|_{\ell^\infty(\Lambda)}\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(\tau/2))}\to 0\text{ as }\tau\to\infty\end{aligned}$$

for all $k\in\mathbf{Z}^d$, where we have used assumption (2.17) to obtain the last limit. This proves the reconstruction formula (3.6) for $r=\infty$, and hence completes the verification that the collection $\tilde{\Psi}$ in (8.16) is the desired displayer associated with the average sampler Ψ.

Now we prove that the displayer $\tilde{\Psi}$ in (8.16) is enveloped by a function in $W_\infty(L_{p,u})$ when Λ is a lattice, and the average sampler Ψ and the generator Φ for the space $V_2(\Phi,\Lambda)$ are enveloped by some functions $g,h\in W_\infty(L_{p,u})$, respectively. Let $A_{\Psi,\Phi}=(A_{\Psi,\Phi}(\gamma,\lambda))_{\gamma\in\Gamma,\lambda\in\Lambda}$ be as in (4.2). Then for any $\lambda\in m+[-1/2,1/2)^d$ and $\lambda'\in m'+[-1/2,1/2)^d$ with $m,m'\in\mathbf{Z}^d$,

$$\begin{aligned}|(A^*_{\Psi,\Phi}A_{\Psi,\Phi})(\lambda,\lambda')|&\le\sum_{k,l\in\mathbf{Z}^d}\sum_{\mu\in\Gamma}\|h(\cdot+\lambda-k)\|_{L^\infty([0,1]^d)}\|g(\cdot+\mu-k)\|_{L^\infty([0,1]^d)}\\&\qquad\times\|g(\cdot+\mu-l)\|_{L^\infty([0,1]^d)}\|h(\cdot+\lambda'-l)\|_{L^\infty([0,1]^d)}\\&\le C\sum_{k,l,n\in\mathbf{Z}^d}\|h(\cdot+m-k)\|_{L^\infty([-1,2]^d)}\|g(\cdot+n-k)\|_{L^\infty([-1,2]^d)}\\&\qquad\times\|g(\cdot+n-l)\|_{L^\infty([-1,2]^d)}\|h(\cdot+m'-l)\|_{L^\infty([-1,2]^d)}\\&\le d(m-m')\end{aligned}$$

for some sequence $(d(m))_{m\in\mathbf{Z}^d}$ with $(d(m)u(m))_{m\in\mathbf{Z}^d}\in\ell^p$. This implies that

$$A^*_{\Psi,\Phi}A_{\Psi,\Phi}\in\mathcal{C}_{p,u}(\Lambda,\Lambda).\tag{8.22}$$

Therefore using (8.13) and (8.22), and applying Lemma 8.2 to the matrix $A^*_{\Psi,\Phi}A_{\Psi,\Phi}$, we conclude that

$$(A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1} \in \mathcal{C}_{p,u}(\Lambda,\Lambda). \tag{8.23}$$

Let R be the matrix in (8.15) and $\tilde{\Psi} := \{\tilde{\psi}_\gamma, \gamma \in \Gamma\}$ be the displayer defined in (8.16). Then for $\gamma \in k + [0,1)^d$ and $x \in l + [0,1)^d$,

$$\begin{aligned}
|\tilde{\psi}_\gamma(x)| &\le \sum_{\lambda,\lambda'\in\Lambda}\sum_{m\in\mathbf{Z}^d} \|g(\cdot+\gamma-m)\|_{L^\infty([0,1]^d)}\|h(\cdot+\lambda'-m)\|_{L^\infty([0,1]^d)} \\
&\quad \times |(A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1}(\lambda',\lambda)|h(x-\lambda) \\
&\le C \sum_{n,n',m\in\mathbf{Z}^d} \|g\|_{L^\infty(m-k+[-1,2]^d)}\|h\|_{L^\infty(m-n'+[-1,2]^d)} \\
&\quad \times b(n-n') \sup_{y\in l-n+[-1,2]^d} |h(y)| \le c(k-l),
\end{aligned}$$

where $(b(k))_{k\in\mathbf{Z}^d}$ and $(c(k))_{k\in\mathbf{Z}^d}$ are sequences with $(b(k)u(k))_{k\in\mathbf{Z}^d}$ and $(c(k)u(k))_{k\in\mathbf{Z}^d}$ belonging to ℓ^p. This proves that $\tilde{\Psi} := \{\tilde{\psi}_\gamma, \gamma \in \Gamma\}$ is enveloped by some function in $W_\infty(L_{p,u})$, and hence completes the proof. □

8.2. Proof of Theorem 3.2. By (3.7) and the relative separatedness of the sets Γ and Λ, the matrix $A_{\delta_\Gamma,\Phi} = (\phi_\lambda(\gamma))_{\gamma\in\Gamma,\lambda\in\Lambda}$ in (4.3) belongs to $\mathcal{A}_{p,u}(\Gamma,\Lambda)$. Then we may reach the conclusion of Theorem 3.2 using the same argument as the one in the proof of Theorem 3.1, except we replace the average sampler Ψ with the ideal sampler δ_Γ. We omit the details of the proof here. □

8.3. Proof of Theorem 4.1. Let S be the frame operator (2.20) on the space $V_2(\Phi,\Lambda)$. By the frame assumption on the generator Φ, we have

$$C^{-1}\|f\|_2 \le \|Sf\|_2 \le C\|f\|_2 \quad \text{for all } f \in V_2(\Phi,\Lambda) \tag{8.24}$$

and

$$C^{-1}\|f\|_2 \le \big\|\big(\langle f,\phi_\lambda\rangle\big)_{\lambda\in\Lambda}\big\|_{\ell^2(\Lambda)} \le C\|f\|_2 \quad \text{for all } f \in V_2(\Phi,\Lambda). \tag{8.25}$$

Then the conclusion follows from (8.24), (8.25), and

$$\langle A^*_{\Psi,\Phi}A_{\Psi,\Phi}c, c\rangle = \big\|\big(\langle Sf,\psi_\gamma\rangle\big)_{\gamma\in\Gamma}\big\|^2_{\ell^2(\Lambda)}$$

for any $c = (\langle f,\phi_\lambda\rangle)_{\lambda\in\Lambda}$, where $f \in V_2(\Phi,\Lambda)$. □

8.4. Proof of Theorem 4.2. Note that $\langle A^*_{\delta_\Gamma,\Phi}A_{\delta_\Gamma,\Phi}c, c\rangle = \|(Sf(\gamma))_{\gamma\in\Gamma}\|^2_{\ell^2(\Lambda)}$ for $c = (\langle f,\phi_\lambda\rangle)_{\lambda\in\Lambda}$, where $f \in V_2(\Phi,\Lambda)$. This, together with (8.24) and (8.25), proves the conclusion. □

8.5. Proof of Theorem 5.1. The estimate (5.1) follows from (3.6), (3.8), Proposition 2.1, and the following inequality:

$$\left|f - \sum_{\gamma\in\Gamma} g_\gamma\tilde{\psi}'_\gamma\right| \le \sum_{\gamma\in\Gamma}|\langle f,\psi_\gamma\rangle - g_\gamma||\tilde{\psi}_\gamma| + \sum_{\gamma\in\Gamma}|g_\gamma||\tilde{\psi}_\gamma - \tilde{\psi}'_\gamma|. \qquad \square$$

8.6. Proof of Theorem 5.2. Let $A_{\Psi,\Phi}$ and $A_{\Psi',\Phi}$ be the Gram matrices as in (4.2). Then there exists a positive constant C independent of the average samplers Ψ' and Ψ by (2.8) and (8.10) such that

$$\begin{aligned}\|A_{\Psi,\Phi}-A_{\Psi',\Phi}\|_{\mathcal{A}_{p,u}} &= \|A_{\Psi-\Psi',\Phi}\|_{\mathcal{A}_{p,u}}\\ &\le C\|\Psi-\Psi'\|_{q/(q-1),p,\alpha}\|\Phi\|_{q,p,u} \le C\|\Psi-\Psi'\|_{q^*,p,u}\|\Psi\|_{q,p,u}.\end{aligned} \tag{8.26}$$

This, together with (8.2), yields the following estimate:

$$\|(A_{\Psi,\Phi})^*A_{\Psi,\Phi}-(A_{\Psi',\Phi})^*A_{\Psi',\Phi}\|_{\mathcal{A}_{p,u}} \le C\|\Psi-\Psi'\|_{q^*,p,u}, \tag{8.27}$$

which then implies that

$$\begin{aligned}&\|((A_{\Psi',\Phi})^*A_{\Psi',\Phi})^{-1}-((A_{\Psi,\Phi})^*A_{\Psi,\Phi})^{-1}\|_{\mathcal{A}_{p,u}}\\ &\quad\le C\|((A_{\Psi,\Phi})^*A_{\Psi,\Phi})^{-1}\|_{\mathcal{A}_{p,u}}\\ &\qquad\times\sum_{k=1}^{\infty}\Big\|\Big(\big((A_{\Psi,\Phi})^*A_{\Psi,\Phi}\big)^{-1}\big((A_{\Psi,\Phi})^*A_{\Psi,\Phi}-(A_{\Psi',\Phi})^*A_{\Psi',\Phi}\big)\Big)^k\Big\|_{\mathcal{A}_{p,u}}\\ &\quad\le C\|((A_{\Psi,\Phi})^*A_{\Psi,\Phi})^{-1}\|_{\mathcal{A}_{p,u}}\\ &\qquad\times\sum_{k=1}^{\infty}\big(C\|((A_{\Psi,\Phi})^*A_{\Psi,\Phi})^{-1}\|_{\mathcal{A}_{p,u}}\|\big((A_{\Psi,\Phi})^*A_{\Psi,\Phi}-(A_{\Psi',\Phi})^*A_{\Psi',\Phi}\big)\|_{\mathcal{A}_{p,u}}\big)^k\\ &\quad\le C\sum_{k=1}^{\infty}\big(C\|\Psi-\Psi'\|_{q^*,p,u}\big)^k \le \frac{C^2\|\Psi-\Psi'\|_{q^*,p,u}}{1-C\|\Psi-\Psi'\|_{q^*,p,u}} \le \frac{C^2\delta_0}{1-C\delta_0}<\infty\end{aligned}$$

when δ_0 in (5.2) is sufficiently small, where we have used the estimates in (5.2), (8.2), (8.14), and (8.27). This proves that $((A_{\Psi',\Phi})^*A_{\Psi',\Phi})^{-1}\in\mathcal{A}_{p,u}(\Lambda,\Lambda)$ as δ_0 in (5.2) is sufficiently small. Moreover using the argument in the proof of Theorem 3.1, we conclude that Ψ' is a stable average sampler for the space $V_2(\Phi,\Lambda)$, and

$$\|R'-R\|_{p,\alpha} \le C\|\Psi-\Psi'\|_{q^*,p,\alpha} \tag{8.28}$$

for some positive constant C, where $R'=A_{\Psi',\Phi}((A_{\Psi',\Phi})^*A_{\Psi',\Phi})^{-1}=(R'(\gamma,\lambda))_{\gamma\in\Gamma,\lambda\in\Lambda}$ and $R=A_{\Psi,\Phi}((A_{\Psi,\Phi})^*A_{\Psi,\Phi})^{-1}=(R(\gamma,\lambda))_{\gamma\in\Gamma,\lambda\in\Lambda}$. Therefore the displayers $\tilde{\Psi}'=\{\tilde{\psi}'_\gamma:\gamma\in\Gamma\}$ and $\tilde{\Psi}=\{\tilde{\psi}_\gamma:\gamma\in\Gamma\}$ associated with the stable average samplers Ψ' and Ψ, respectively, which are defined by $\psi'_\gamma=\sum_{\lambda\in\Lambda}R'(\gamma,\lambda)\phi_\lambda$ and $\psi_\gamma=\sum_{\lambda\in\Lambda}R(\gamma,\lambda)\phi_\lambda$, $\gamma\in\Gamma$, satisfy

$$\|\tilde{\Psi}'-\tilde{\Psi}\|_{q^*,p,u} \le C\|R-R'\|_{\mathcal{A}_{p,u}}\|\Phi\|_{q^*,p,u} \le C\|\Psi-\Psi'\|_{q^*,p,u}. \tag{8.29}$$

Hence (5.3) follows. □

8.7. Proof of Theorem 5.3. We can use the same technique as the one in the proof of Theorem 5.2, except the matrices $A_{\Psi,\Phi}$ and $A_{\Psi',\Phi}$ and the estimate (8.26) are replaced with the matrices $A_{\delta_\Gamma,\Phi}=(\phi_\lambda(\gamma))_{\gamma\in\Gamma,\lambda\in\Lambda}$ and $A_{\delta_{\Gamma'},\Phi}=(\phi_\lambda(\gamma+\delta_\gamma))_{\gamma\in\Gamma,\lambda\in\Lambda}$, and the estimate $\|A_{\delta_\Gamma,\Phi}-A_{\delta_{\Gamma'},\Phi}\|_{\mathcal{A}_{p,u}}\le\|\Phi_{\delta_0}\|_{\infty,p,u}$ for sequences $\{\delta_\gamma\}$ with $\sup_{\gamma\in\Gamma}|\delta_\gamma|\le\delta_0$, respectively. We omit the details of the proof here. □

8.8. Proof of Corollary 5.4. The first conclusion follows from Theorem 5.2 and the equivalence between $\sum_{l=1}^{L}\|\theta_n\|_{W_{q^*}(L_{p,u})}$ and $\|\Theta\|_{q*,p,\alpha}$, where $\Theta=\{\theta_l(\cdot-\gamma):1\le l\le L,\gamma\in\Gamma\}$.

For any continuous function ϕ in $W_\infty(L_{p,u})$, there exist continuous functions $\phi_n, n \ge 1$, with compact support such that $\lim_{n\to\infty} \|\phi_n - \phi\|_{W_\infty(L_{p,u})} = 0$. Therefore the modulus of continuity $\omega(\phi,\delta)(x) := \sup_{|t|\le\delta} |\phi(x+t) - \phi(x)|$ of the continuous function ϕ in $W_\infty(L_{p,u})$ has the property that $\|\omega(\phi,\delta)\|_{W_\infty(L_{p,u})} \to 0$ as $\delta \to 0$ [2]. This together with (2.13) and Theorem 5.3 proves the second conclusion. □

8.9. Proof of Theorem 6.1. For average sampling on the space $V_r(\Phi,\Lambda)$ with the average sampler $\Psi := \{\psi_\gamma, \gamma \in \Gamma\}$, we introduce two local reconstruction approximations of a function $f \in V_r(\Phi,\Lambda)$ on a bounded set K using average sampling data on the R-neighborhood $B(K,R)$ of the set K by

$$f_{K,R} = \sum_{\gamma\in\Gamma\cap B(K,R)} \langle f, \psi_\gamma\rangle \tilde{\psi}_\gamma \tag{8.30}$$

and

$$f^1_{K,R} = \sum_{\gamma\in\Gamma\cap B(K,R)} \langle f, \psi_\gamma\rangle \tilde{\psi}^1_{\gamma,K,R}, \tag{8.31}$$

where

$$\tilde{\psi}_\gamma = \sum_{\lambda_1,\lambda_2\in\Lambda} \langle \psi_\gamma, \phi_{\lambda_1}\rangle (A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1}(\lambda_1,\lambda_2)\phi_{\lambda_2}$$

and

$$\tilde{\psi}^1_{\gamma,K,R} = \sum_{\lambda_1,\lambda_2\in B(K,2R)\cap\Lambda} \langle \psi_\gamma, \phi_{\lambda_1}\rangle (A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1}(\lambda_1,\lambda_2)\phi_{\lambda_2}.$$

For a bounded domain K and a positive number $R > 0$, define the projection matrices $P_{K,R}$ and $Q_{K,R}$ by

$$(P_{K,R}c)(\lambda) = \begin{cases} c(\lambda) & \text{if } \lambda \in \Lambda\cap B(K,R), \\ 0 & \text{if } \lambda \notin \Lambda\cap B(K,R) \end{cases}$$

for any $c := (c(\lambda))_{\lambda\in\Lambda}$, and

$$(Q_{K,R}d)(\gamma) = \begin{cases} d(\gamma) & \text{if } \gamma \in \Gamma\cap B(K,R), \\ 0 & \text{if } \gamma \notin \Gamma\cap B(K,R) \end{cases}$$

for any $d := (d(\gamma))_{\gamma\in\Gamma}$.

To prove Theorem 6.1, we need the following estimates for $f - f_{K,R}$ and $f_{K,R} - f^1_{K,R}$.

LEMMA 8.3. *Let $p, q, q^*, r, u, \Lambda, \Gamma, \Phi, \Psi, \tilde{\Psi}$ be as in Theorem 3.1, and set $p' = p/(p-1)$. Then there exists a positive constant C (independent of the bounded set K, the positive number $R \ge 1$, and the function $f \in V_r(\Phi,\Lambda)$) such that*

$$\|f_{K,R} - f\|_{L^r(K)} \le C\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(R))} \|(\langle f, \psi_\gamma\rangle)_{\gamma\in\Gamma\setminus B(K,R)}\|_{\ell^r(\Gamma\setminus B(K,R))} \tag{8.32}$$

and

$$\|f_{K,R} - f^1_{K,R}\|_{L^r(K)} \le C\|u^{-1}\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))} \|(\langle f, \psi_\gamma\rangle)_{\gamma\in B(K,R)}\|_{\ell^r(B(K,R))} \tag{8.33}$$

for any compact set K, any positive number $R \ge 1$, and any $f \in V_r(\Phi,\Lambda)$.

Proof. Let $\tilde{K} \subset \mathbf{Z}^d$ be the minimal subset of $\mathbf{Z}^d$ such that $K \subset \tilde{K} + [0,1)^d$. For $1 \le r \le q$ and $r < \infty$,

$$\begin{aligned}
\|f - f_{K,R}\|_{L^r(K)}^r &= \left\| \sum_{\gamma \notin \Gamma_{K,R}} \langle f, \psi_\gamma \rangle \tilde{\psi}_\gamma \right\|_{L^r(K)}^r \\
&\le \int_{\tilde{K}+[0,1]^d} \left(\sum_{\gamma \notin \Gamma_{K,R}} |\langle f, \psi_\gamma \rangle|^r |\tilde{\psi}_\gamma(x)| \right) \times \left(\sum_{\gamma \notin \Gamma_{K,R}} |\tilde{\psi}_\gamma(x)| \right)^{r-1} dx \\
&\le \sum_{\gamma \notin \Gamma_{K,R}} |\langle f, \psi_\gamma \rangle|^r \sum_{k \in \tilde{K}} \|\tilde{\psi}_\gamma\|_{L^r(k+[0,1]^d)} \times \left(\sum_{\gamma \notin \Gamma_{K,R}} \|\tilde{\psi}_\gamma\|_{L^r(k+[0,1]^d)} \right)^{r-1} \\
&\le \|(\langle f, \psi_\gamma \rangle)_{\gamma \in \Gamma \setminus \Gamma_{K,R}}\|_{\ell^r(\Gamma \setminus \Gamma_{K,R})}^r \times \left(\sup_{\gamma \notin \Gamma_{K,R}} \sum_{k \in \tilde{K}} \|\tilde{\phi}_\gamma\|_{L^r(k+[0,1]^d)} \right) \\
&\quad \times \left(\sup_{k \in \tilde{K}} \sum_{\gamma \notin \Gamma_{K,R}} \|\tilde{\psi}_\gamma\|_{L^r(k+[0,1]^d)} \right)^{r-1} \qquad (8.34)
\end{aligned}$$

where we set $\Gamma_{K,R} = \Gamma \cap B(K,R)$. Then for all $r \in [1,\infty)$ with $1 \le r \le q$, we have

$$\begin{aligned}
\|f - f_{K,R}\|_{L^r(K)} &\le C \|(\langle f, \psi_\gamma \rangle)_{\gamma \in \Gamma \setminus \Gamma_{K,R}}\|_{\ell^r(\Gamma \setminus \Gamma_{K,R})} \\
&\quad \times \|u\|_{L^{p'}(\mathbf{R}^d \setminus B(R))} \|\tilde{\Psi}\|_{r,p,u}. \qquad (8.35)
\end{aligned}$$

For $r = \infty$, it follows from $r \le q$ that $q = \infty$. Then using standard modification to the estimate (8.35), we obtain

$$\begin{aligned}
\|f - f_{K,R}\|_{L^\infty(K)} &\le C \|(\langle f, \psi_\gamma \rangle)_{\gamma \in \Gamma \setminus \Gamma_{K,R}}\|_{\ell^\infty(\Gamma \setminus \Gamma_{K,R})} \\
&\quad \times \|u\|_{L^{p'}(\mathbf{R}^d \setminus B(R))} \|\tilde{\Psi}\|_{\infty,p,u}. \qquad (8.36)
\end{aligned}$$

Then the local estimate (8.32) follows from (8.18), (8.35), and (8.36).

Set $R_{\Phi,r} = (\|\phi_\lambda\|_{L^r(k+[0,1]^d)})_{\lambda \in \Lambda, k \in \mathbf{Z}^d}$ and $|A| = (|a_{\lambda,\lambda'}|)$ for a matrix $A = (a_{\lambda,\lambda'})$. Then it follows from (2.7), (2.9), and Lemma 8.1 that

$$\begin{aligned}
&\sup_{k \in \tilde{K}} \sum_{\gamma \in B(K,R) \cap \Gamma} \|\tilde{\psi}_\gamma - \tilde{\psi}^1_{\gamma,K,R}\|_{L^r(k+[0,1]^d)} \\
&\qquad + \sup_{\gamma \in B(K,R) \cap \Gamma} \sum_{k \in \tilde{K}} \|\tilde{\psi}_\gamma - \tilde{\psi}^1_{\gamma,K,R}\|_{L^r(k+[0,1]^d)} \\
&\le C \left(\sup_{k \in \tilde{K}} \sum_{\gamma \in B(K,R) \cap \Gamma} + \sup_{\gamma \in B(K,R) \cap \Gamma} \sum_{k \in \tilde{K}} \right) \\
&\qquad \left(\sum_{\lambda \in B(K,2R) \cap \Lambda} \sum_{\lambda' \in \Lambda \setminus B(K,2R)} + \sum_{\lambda \in \Lambda \setminus B(K,2R)} \sum_{\lambda' \in \Lambda} \right) \\
&\qquad |\langle \psi_\gamma, \phi_\lambda \rangle| \, |(A^*_{\Psi,\Phi} A_{\Psi,\Phi})^{-1}(\lambda,\lambda')| \, \|\phi_{\lambda'}\|_{L^r(k+[0,1]^d)} \\
&\le C \|Q_{K,R} \, |A_{\Psi,\Phi}| \, P_{K,2R} \, |(A^*_{\Psi,\Phi} A_{\Psi,\Phi})^{-1}| \, (I - P_{K,2R}) R_{\Phi,r} P_{K,1}\|_{\mathcal{A}_{1,u_0}} \\
&\quad + C \|Q_{K,R} \, |A_{\Psi,\Phi}| \, (I - P_{K,2R}) \, |(A^*_{\Psi,\Phi} A_{\Psi,\Phi})^{-1}| \, R_{\Phi,r} P_{K,1}\|_{\mathcal{A}_{1,u_0}} \\
&\le C \|u\|^2_{L^{p'}(\mathbf{R}^d \setminus B(R))} \| \, |A_{\Psi,\Phi}| \, P_{K,2R} \, |(A^*_{\Psi,\Phi} A_{\Psi,\Phi})^{-1}| \, \|_{\mathcal{A}_{p,u}} \|R_{\Phi,r}\|_{\mathcal{A}_{p,u}} \\
&\quad + C \|u\|^2_{L^{p'}(\mathbf{R}^d \setminus B(R))} \|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}} \| \, |(A^*_{\Psi,\Phi} A_{\Psi,\Phi})^{-1}| \, R_{\Phi,r}\|_{\mathcal{A}_{p,u}} \\
&\le C \|u\|^2_{L^{p'}(\mathbf{R}^d \setminus B(R))} \|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}} \|(A^*_{\Psi,\Phi} A_{\Psi,\Phi})^{-1}\|_{\mathcal{A}_{p,u}} \|\Phi\|_{r,p,u}. \qquad (8.37)
\end{aligned}$$

This together with (8.10) and (8.14) proves (8.33). □

To prove Theorem 6.1, we need another lemma.

LEMMA 8.4. *Let $p, q, q^*, r, u, \Lambda, \Gamma, \Phi, \Psi, \tilde{\Psi}$ be as in Theorem 3.1, and $A_{\Psi,\Phi}$ be as in (4.2). Then there exist positive constants C and δ_0 (independent of the bounded set K and the positive number $R \geq 1$) such that*

$$\begin{aligned} &\|P_{K,2R}(A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1}P_{K,2R} - P_{K,2R}R_{\Psi,\Phi,K,3R}P_{K,2R}\|_{\mathcal{A}_{1,u_0}} \\ &\quad \leq C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))} \end{aligned} \tag{8.38}$$

holds for all bounded sets K and all positive numbers R with $\|u\|_{L^{p'}(\mathbf{R}^d\setminus B(R))} \leq \delta_0$, where $R_{\Psi,\Phi,K,3R}$ is the generalized inverse of the matrix $P_{K,3R}\tilde{A}_{\Psi,\Phi,K,R}P_{K,3R}$, that is, $R_{\Psi,\Phi,K,3R}P_{K,3R}\tilde{A}_{\Psi,\Phi,K,R}P_{K,3R} = P_{K,3R}$ and $P_{K,3R}R_{\Psi,\Phi,K,3R}P_{K,3R} = R_{\Psi,\Phi,K,3R}$.

Proof. By (8.13), for any $A \geq (\|A^*_{\Psi,\Phi}A_{\Psi,\Phi}\|_{\mathcal{B}^2})^{1/2}$ there exists a matrix $B \in \mathcal{A}_{p,u}(\Lambda,\Lambda)$ such that

$$A^*_{\Psi,\Phi}A_{\Psi,\Phi} = A^2(I - B) \tag{8.39}$$

and

$$\|B\|_{\mathcal{B}^2} < 1, \tag{8.40}$$

where I is the usual unit matrix. By (8.4), (8.40), and the estimates $\|P_{K,3R}BP_{K,3R}\|_{\mathcal{B}^2} \leq \|B\|_{\mathcal{B}^2}$ and $\|P_{K,3R}BP_{K,3R}\|_{\mathcal{A}_{p,u}} \leq \|B\|_{\mathcal{A}_{p,u}}$, we have

$$\|B^k\|_{\mathcal{A}_{p,u}} + \|(P_{K,3R}BP_{K,3R})^k\|_{\mathcal{A}_{p,u}} \leq C\bigg(\frac{\|B\|_{\mathcal{B}^2}+1}{2}\bigg)^k,\ k \geq 1. \tag{8.41}$$

Therefore for $k \geq 2$,

$$\begin{aligned} &\bigg\|P_{K,2R}\sum_{k=1}^{\infty}B^kP_{K,2R} - P_{K,2R}\sum_{k=1}^{\infty}(P_{K,3R}BP_{K,3R})^kP_{K,2R}\bigg\|_{\mathcal{A}_{1,u_0}} \\ &\leq \sum_{k=2}^{\infty}\sum_{l=0}^{k-2}\|P_{K,2R}(P_{K,3R}BP_{K,3R})^lB(I-P_{K,3R})B^{k-1-l}P_{K,2R}\|_{\mathcal{A}_{1,u_0}} \\ &\leq C\sum_{k=2}^{\infty}\sum_{l=0}^{k-2}\|P_{K,2R}(P_{K,3R}BP_{K,3R})^lB(I-P_{K,3R})\|_{\mathcal{A}_{1,u_0}} \\ &\qquad\qquad \times\|(I-P_{K,3R})B^{k-1-l}P_{K,2R}\|_{\mathcal{A}_{1,u_0}} \\ &\leq C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\sum_{k=2}^{\infty}\sum_{l=0}^{k-2}\|(P_{K,3R}BP_{K,3R})^lB\|_{\mathcal{A}_{p,u}}\|B^{k-1-l}\|_{\mathcal{A}_{p,u}} \\ &\leq C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|B\|_{\mathcal{A}_{p,u}}\sum_{k=2}^{\infty}(k-1)\bigg(\frac{\|B\|_{\mathcal{B}^2}+1}{2}\bigg)^{k-1} \\ &\leq C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}, \end{aligned} \tag{8.42}$$

where we have used (8.40) and (8.41) to obtain the last inequality.

Write

$$P_{K,3R}\tilde{A}_{\Psi,\Phi,K,R}P_{K,3R} = P_{K,3R}A^*_{\Psi,\Phi}Q_{K,4R}A_{\Psi,\Phi}P_{K,3R} = A^2(P_{K,3R} - B'), \tag{8.43}$$

where A is the positive constant in (8.39). Since

$$\begin{aligned}&\|P_{K,3R}A^*_{\Psi,\Phi}Q_{K,4R}A_{\Psi,\Phi}P_{K,3R}-P_{K,3R}A^*_{\Psi,\Phi}A_{\Psi,\Phi}P_{K,3R}\|_{\mathcal{B}^2}\\&\quad\le\|P_{K,3R}A^*_{\Psi,\Phi}Q_{K,4R}A_{\Psi,\Phi}P_{K,3R}-P_{K,3R}A^*_{\Psi,\Phi}A_{\Psi,\Phi}P_{K,3R}\|_{\mathcal{A}_{1,u_0}}\\&\quad\le C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|A_{\Psi,\Phi}\|^2_{\mathcal{A}_{p,u}},\end{aligned}\tag{8.44}$$

we have that

$$\|P_{K,3R}BP_{K,3R}-B'\|_{\mathcal{B}^2}\le D_0\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|A_{\Psi,\Phi}\|^2_{\mathcal{A}_{p,u}}\le\frac{1-\|B\|_{\mathcal{B}^2}}{3},\tag{8.45}$$

when $\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\le\delta_0$ for some sufficiently small positive number δ_0, where D_0 is a positive constant. Therefore

$$\|(B')^k\|_{\mathcal{A}_{p,u}}\le C\bigg(\frac{\|B\|_{\mathcal{B}^2}+1}{2}\bigg)^k,\quad k\ge1,\tag{8.46}$$

by (8.4), (8.40), and (8.45). Similar to the argument in the proof of the estimate (8.42), we have

$$\begin{aligned}&\bigg\|P_{K,2R}\sum_{k=1}^{\infty}(B')^kP_{K,2R}-P_{K,2R}\sum_{k=1}^{\infty}(P_{K,3R}BP_{K,3R})^kP_{K,2R}\bigg\|_{\mathcal{A}_{1,u_0}}\\&\quad\le\sum_{k=1}^{\infty}\sum_{l=0}^{k-1}\|(B')^l(B'-P_{K,3R}BP_{K,3R})(P_{K,3R}BP_{K,3R})^{k-1-l}\|_{\mathcal{A}_{1,u_0}}\\&\quad\le C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R)}\|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}}\sum_{k=1}^{\infty}k\bigg(\frac{\|B\|_{\mathcal{B}^2}+1}{2}\bigg)^k\le C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))},\end{aligned}\tag{8.47}$$

where the second inequality follows from (8.41), (8.44), (8.45), and (8.46). Combining (8.39), (8.42), (8.43), and (8.47) proves the desired estimate (8.38). □

Now we start to prove Theorem 6.1.

Proof of Theorem 6.1. By Lemma 8.4,

$$\sup_{\lambda\in B(K,2R)}\sum_{\lambda'\in B(K,2R)}|(A^*_{\Psi,\Phi}A_{\Psi,\Phi})^{-1}(\lambda,\lambda')-(\tilde A_{\Psi,\Phi,K,R})^{-1}(\lambda,\lambda')|\le C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))},$$

which, together with (8.9) and Lemma 8.1, implies that

$$\begin{aligned}&\bigg(\sup_{k\in\tilde K}\sum_{\gamma\in B(K,R)\cap\Gamma}+\sup_{\gamma\in B(K,R)\cap\Gamma}\sum_{k\in\tilde K}\bigg)\|\tilde\psi^a_{\gamma,K,R}-\tilde\psi^1_{\gamma,K,R}\|_{L^r(k+[0,1]^d)}\\&\quad\le C\|u\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}.\end{aligned}\tag{8.48}$$

Therefore estimate (6.7) follows from (8.48), Proposition 2.1, and Lemma 8.3. □

8.10. Proof of Theorem 6.2. Theorem 6.2 can be proved using an argument similar to the one in the proof of Theorem 6.1, except the average sampler Ψ is replaced with the ideal sampler δ_Γ. We omit the details of the proof here. □

8.11. Proof of Theorem 6.3. For any bounded set K, we let

$$V_2(\Phi, \Lambda\cap K)=\bigg\{\sum_{\lambda\in\Lambda\cap K} c(\lambda)\phi_\lambda,\ \sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2<\infty\bigg\}.$$

By the Riesz assumption on Φ,

$$C^{-1}\|c\|_{\ell^2(\Lambda\cap K)}\le\bigg\|\sum_{\lambda\in\Lambda\cap K} c(\lambda)\phi_\lambda\bigg\|_{L^2(\mathbf{R}^d)}\le C\|c\|_{\ell^2(\Lambda\cap K)} \tag{8.49}$$

for any sequence $c:=(c(\lambda))_{\lambda\in\Lambda\cap K}$. For any $R\ge 1$, it follows by the localization assumption on the generator Φ that

$$\begin{aligned}&\bigg\|\sum_{\lambda\in\Lambda\cap K} c(\lambda)\phi_\lambda\bigg\|^2_{L^2(\mathbf{R}^d\setminus B(K,R))}\\ &\quad\le\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2\sum_{k\in\mathbf{Z}^d\setminus B(K,R-1)}\|\phi_\lambda\|_{L^2(k+[0,1]^d)}\bigg(\sum_{\lambda'\in\Lambda\cap K}\|\phi_{\lambda'}\|_{L^2(k+[0,1]^d)}\bigg)\\ &\quad\le C\|\Phi\|^2_{2,p,u}\|u^{-1}\|^2_{L^{p'}(\mathbf{R}^d\setminus B(R))}\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2.\end{aligned} \tag{8.50}$$

We recall that

$$\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(R))}\to 0\text{ as }R\to\infty \tag{8.51}$$

by (2.17). Therefore by (8.49), (8.50), and (8.51), there exist positive constants C and R_1 such that

$$C^{-1}\|c\|_{\ell^2(\Lambda\cap K)}\le\bigg\|\sum_{\lambda\in\Lambda\cap K} c(\lambda)\phi_\lambda\bigg\|_{L^2(B(K,R_1))}\le C\|c\|_{\ell^2(\Lambda\cap K)} \tag{8.52}$$

for all sequences $c:=(c(\lambda))_{\lambda\in\Lambda\cap K}$.

Set $K_1=B(K,R_1)$. For the average sampling/reconstruction process, there exists a positive constant C by Theorem 6.1 and Lemma 8.1 such that for any $R\ge 1$ and $f=\sum_{\lambda\in\Lambda\cap K}c(\lambda)\phi_\lambda$,

$$\begin{aligned}\|\tilde f^a_{K_1,R}-f\|_{L^2(K_1)}&\le C\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|(\langle f,\psi_\gamma\rangle)_{\gamma\in\Gamma}\|_{\ell^2(\Gamma)}\\ &\le C\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}}\bigg(\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2\bigg)^{1/2}.\end{aligned} \tag{8.53}$$

Similarly for the ideal sampling/reconstruction procedure, there exists a positive constant C such that for any $R\ge 1$ and $f=\sum_{\lambda\in\Lambda\cap K}c(\lambda)\phi_\lambda$,

$$\|\tilde f^i_{K_1,R}-f\|_{L^2(K_1)}\le C\|u^{-1}\|_{L^{p'}(\mathbf{R}^d\setminus B(R))}\|A_{\delta_\Gamma,\Phi}\|_{\mathcal{A}_{p,u}}\bigg(\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2\bigg)^{1/2}. \tag{8.54}$$

Therefore there exists a positive constant R_2 independent of K by (8.9) and (8.51)–(8.54) so that

$$C^{-1}\bigg(\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2\bigg)^{1/2}\le\|f^a_{K_1,R_2}\|_{L^2(K_1)}\le C\bigg(\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2\bigg)^{1/2} \tag{8.55}$$

for the average sampling/reconstruction process, and

$$C^{-1}\left(\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2\right)^{1/2}\le\|f^i_{K_1,R_2}\|_{L^2(K_1)}\le C\left(\sum_{\lambda\in\Lambda\cap K}|c(\lambda)|^2\right)^{1/2} \tag{8.56}$$

for the ideal sampling/reconstruction process. Therefore conclusion (6.9) follows by letting $R_0=R_1+R_2$. □

8.12. Proof of Theorem 7.1. Let $A_{\Psi,\Phi}$ be as in (4.2), and the matrix $B\in\mathcal{A}_{p,u}(\Lambda,\Lambda)$ be as in (8.39) and (8.40). Write the sequence $(a_\gamma)_{\gamma\in\Gamma}$ as a vector, to be denoted by a, and the family of functions Φ as a vector, which is still denoted by Φ. We claim that

$$f_n=\sum_{k=0}^n a^TA_{\Psi,\Phi}B^k\Phi,\quad n\ge0, \tag{8.57}$$

where a^T denotes the transpose of the vector a. The above claim is obviously true for $n=0$. Inductively we assume that the claim is true for n. By (7.1), (8.39), and the inductive hypothesis, we have

$$\begin{aligned}f_{n+1}&=a^TA_{\Psi,\Phi}\Phi+\sum_{k=0}^n a^TA_{\Psi,\Phi}B^k\Phi-A^{-2}\sum_{k=0}^n a^TA_{\Psi,\Phi}B^kA^*_{\Psi,\Phi}A_{\Psi,\Phi}\Phi\\&=a^TA_{\Psi,\Phi}\Phi+\sum_{k=0}^n a^TA_{\Psi,\Phi}B^k\Phi-\sum_{k=0}^n a^TA_{\Psi,\Phi}B^k(I-B)\Phi\\&=\sum_{k=0}^{n+1}a^TA_{\Psi,\Phi}B^k\Phi.\end{aligned}$$

This proves claim (8.57) by induction.

By (8.9), (8.41), (8.57), and Proposition 2.1, we have

$$\begin{aligned}\|f_{n+1}-f_n\|_r&=\|a^TA_{\Psi,\Phi}B^{n+1}\Phi\|_r\\&\le C\|a\|_{\ell^r(\Gamma)}\|A_{\Psi,\Phi}\|_{\mathcal{A}_{p,u}}\|B^{n+1}\|_{\mathcal{A}_{p,u}}\|\Phi\|_{q,p,u}\\&\le C\left(\frac{\|B\|_{\mathcal{B}^2}+1}{2}\right)^n\|a\|_{\ell^r(\Gamma)}\quad\text{for all } n\ge0,\end{aligned} \tag{8.58}$$

where $a:=(a_\gamma)_{\gamma\in\Gamma}$. The first conclusion then follows from (8.58).

Now we assume that the initial data $a:=(a_\gamma)_{\gamma\in\Gamma}$ are obtained from average sampling a function $f\in V_r(\Phi,\Lambda)$. Taking limits at both sides of $f_n=f_0+f_{n-1}-A^{-2}T_{as}f_{n-1}$ and using the Riesz property of Φ, we obtain

$$\sum_{\gamma\in\Gamma}\langle f-f_\infty,\psi_\gamma\rangle\langle\psi_\gamma,\phi_\lambda\rangle=0\ \text{ for all } \lambda\in\Lambda. \tag{8.59}$$

Write

$$f-f_\infty=\sum_{\lambda\in\Lambda}d_\lambda\phi_\lambda \tag{8.60}$$

for some ℓ^r sequence $d=(d_\lambda)_{\lambda\in\Lambda}$. We then may write (8.59) as

$$d^TA^*_{\Psi,\Phi}A_{\Psi,\Phi}=0. \tag{8.61}$$

Combining (8.13), (8.60), and (8.61) leads to the second conclusion of the theorem that the limit function f_∞ agrees with the original function f. □

8.13. Proof of Theorem 7.2. We may use the same argument as in the proof of Theorem 7.1 with standard modification, for instance, the matrix $A_{\Psi,\Phi}$ in the proof of Theorem 7.1 by the matrix $A_{\delta_\Gamma,\Phi} := (\phi_\lambda(\gamma))_{\gamma\in\Gamma,\lambda\in\Lambda}$. We omit the details of the proof here. □

Acknowledgments. The author would like to thank Professors A. Aldroubi, K. Gröchenig, and D. Han for their help in preparing this paper. The author also thanks the anonymous referees for their valuable comments and suggestions which led to an improvement of the results and the presentation in the paper. In particular, the author thanks the anonymous referee for the suggestion that leads to the general principle in Remark 3.1 for the well-localization of the displayer, and for pointing out that the results in Theorems 3.1 and 3.2 can be derived from theorems for localized frames in [8, 24, 32].

REFERENCES

[1] A. Aldroubi, *Non-uniform weighted average sampling and reconstruction in shift-invariant and wavelet spaces*, Appl. Comput. Harmon. Anal., 13 (2002), pp. 151–161.
[2] A. Aldroubi and K. Gröchenig, *Nonuniform sampling and reconstruction in shift-invariant spaces*, SIAM Rev., 43 (2001), pp. 585–620.
[3] A. Aldroubi and K. Gröchenig, *Beurling-Landau-type theorems for non-uniform sampling in shift invariant spaces*, J. Fourier Anal. Appl., 6 (2000), pp. 93–103.
[4] A. Aldroubi and I. Krishtal, *Robustness of sampling and reconstruction and Beurling-Landau-type theorems for shift invariant spaces*, Appl. Comput. Harmon. Anal., 20 (2006), pp. 250–260.
[5] A. Aldroubi, Q. Sun, and W.-S. Tang, *Convolution, average sampling and a Calderon resolution of the identity for shift-invariant spaces*, J. Fourier Anal. Appl., 11 (2005), pp. 215–244.
[6] A. Aldroubi, Q. Sun, and W.-S. Tang, *Non-uniform average sampling and reconstruction in multiply generated shift-invariant spaces*, Constr. Approx., 20 (2004), pp. 173–189.
[7] A. Aldroubi, Q. Sun, and W.-S. Tang, *p-frames and shift invariant subspaces of L^p*, J. Fourier Anal. Appl., 7 (2001), pp. 1–21.
[8] R. Balan, P. G. Casazza, C. Heil, and Z. Landau, *Density, overcompleteness and localization of frames* I: *theory*, J. Fourier Anal. Appl., 12 (2006), pp. 105–143.
[9] R. Balan, P. G. Casazza, C. Heil, and Z. Landau, *Density, overcompeleteness and localization of frames* II: *Gabor System*, J. Fourier Anal. Appl., 12 (2006), pp. 309–344.
[10] A. G. Baskakov, *Wiener's theorem and asymptotic estimates for elements in inverse matrices*, Funktsional Anal. i. Prilozhen., 24 (1990), pp. 64–65.
[11] E. Beller and G. de Haan, *New algorithms for motion estimation on interlaced video*, in Visual Communication and Image Processing, 3309, 1998, pp. 111–121.
[12] J. J. Benedetto and P. J. S. G. Ferreira, eds., *Modern Sampling Theory: Mathematics and Applications*, Birkhäuser Boston, Boston, MA, 2001.
[13] J. J. Benedetto and S. Li, *The theory of multiresolution analysis frames and applications to filter banks*, Appl. Comput. Harmon. Anal., 5 (1998), pp. 389–427.
[14] J. J. Benedetto and A. I. Zayed, eds., *Sampling, Wavelets, and Tomography*, Birkhäuser Boston, Boston, MA, 2003.
[15] T. Blu, P. Thévenaz, and M. Unser, *Linear interpolation revitalized*, IEEE Trans. Image Process., 13 (2004), pp. 710–719.
[16] M. Bownik, *The structure of shift-invariant subspaces of $L^2(R^n)$*, J. Funct. Anal., 177 (2000), pp. 282–309.
[17] P. G. Casazza, *The art of frame theory*, Taiwanese J. Math., 4 (2000), pp. 129–201.
[18] P. G. Casazza and O. Christensen, *Approximation of the inverse frame operator and applications to Gabor frames*, J. Approx. Theory, 103 (2000), pp. 338–356.
[19] W. Chen, S. Itoh, and J. Shiki, *Irregular sampling theorems for wavelet subspaces*, IEEE Trans. Inform. Theory, 44 (1998), pp. 1131–1142.
[20] O. Christensen and T. Strohmer, *The finite section method and problems in frame theory*, J. Approx. Theory, 133 (2005), pp. 221–237.
[21] C. K. Chui and Q. Sun, *Affine frame decompositions and shift-invariant spaces*, Appl. Comput. Harmon. Anal., 20 (2006), pp. 74–107.

[22] E. Cordero and K. Gröchenig, *Localization of frames* II, Appl. Comput. Harmonic Anal., 17 (2004), pp. 29–47.
[23] R. J.-M. Cramer, R. A. Scholtz, and M. Z. Win, *Evaluation of an ultra wide-band propagation channel*, IEEE Trans. Antennas and Propagation, 50 (2002), pp. 561–569.
[24] M. Fornasier and K. Gröchenig, *Intrinsic localization of frames*, Constr. Approx., 22 (2005), pp. 395–415.
[25] H. G. Feichtinger and K. Gröchenig, *Irregular sampling theorems and series expansions of band-limited functions*, J. Math. Anal. Appl., 167 (1992), pp. 530–556.
[26] K. Gröchenig, *Foundation of Time-Frequency Analysis*, Birkhäuser Boston, Boston, MA, 2001.
[27] K. Gröchenig, *Time-frequency analysis of Sjöstrand class*, Rev. Mat. Iberoamericana, 22 (2006), pp. 703–724.
[28] K. Gröchenig, *Localization of frames, Banach frames, and the invertibility of the frame operator*, J. Fourier Anal. Appl., 10 (2004), pp. 105–132.
[29] K. Gröchenig, *Acceleration of the frame algorithm*, IEEE Trans. Signal Process., 41 (1993), pp. 3331–3340.
[30] K. Gröchenig, *Reconstruction algorithms in irregular sampling*, Math. Comp., 59 (1992), pp. 181–194.
[31] K. Gröchenig and M. Leinert, *Wiener's lemma for twisted convolution and Gabor frames*, J. Amer. Math. Soc., 17 (2003), pp. 1–18.
[32] K. Gröchenig and M. Leinert, *Symmetry of matrix algebras and symbolic calculus for infinite matrices*, Trans. Amer. Math. Soc., 358 (2006), pp. 2695–2711.
[33] K. Gröchenig and H. Schwab, *Fast local reconstruction methods for nonuniform sampling in shift-invariant spaces*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 899–913.
[34] Y. Hao, P. Marziliano, M. Vetterli, and T. Blu, *Sampling and reconstruction of ECG as a signal with a finite rate of innovation*, IEEE Trans. Biomedical Engineering, submitted.
[35] J. A. Hogan, *Frame-based non-uniform sampling in Paley-Wiener spaces*, J. Fourier Anal. Appl., to appear.
[36] H. S. Hou and H. C. Andrews, *Cubic splines for image interpolation and digital filtering*, IEEE Trans. Acoust. Speech Signal Process., 26 (1978), pp. 508–517.
[37] S. Jaffard, *Properiétés des matrices bien localisées prés de leur diagonale et quelques applications*, Ann. Inst. H. Poincaré, 7 (1990), pp. 461–476.
[38] A. J. E. M. Janssen, *Duality and biorthogonality for Weyl-Heisenberg frames*, J. Fourier Anal. Appl., 1 (1995), pp. 403–436.
[39] R.-Q. Jia and C. A. Micchelli, *Using the refinement equations for the construction of prewavelets*. II. *Powers of two*, in Curves and Surfaces (Chamonix-Mont-Blanc, 1990), Academic Press, Boston, MA, 1991, pp. 209–246.
[40] J. Kusuma, I. Maravic, and M. Vetterli, *Sampling with finite rate of innovation: Channel and timing estimatation for UWB and GPS*, in Proceedings of the IEEE Conference on Communication, Anchorage, AK, 2003, pp. 3540–3544.
[41] Y. Liu, *Irregular sampling for spline wavelet subspaces*, IEEE Trans. Inform. Theory, 42 (1996), pp. 623–627.
[42] I. Maravic and M. Vetterli, *Sampling and reconstruction of signals with finite rate of innovation in the presence of noise*, IEEE Trans. Signal Process., 53 (2005), pp. 2788–2805.
[43] F. A. Marvasti, ed., *Nonuniform Sampling: Theory and Practice, Inf. Technol. Transm., Process. Storage*, Kluwer Academic/Plenum Publishers, New York, 2001.
[44] P. Marziliano and M. Vetterli, *Reconstruction of irregular sampled discrete-time bandlimited signals with unknown sampling locations*, IEEE Trans. Signal Process., 48 (2000), pp. 3462–3471.
[45] P. Marziliano, M. Vetterli, and T. Blu, *Sampling and exact reconstruction of bandlimited signals with shot noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 2230–2233.
[46] P. Marziliano, *Sampling Innovations*, Ph.D. thesis, Swiss Federal Institute of Technology, Zurich, Switzerland, 2001.
[47] A. Ron and Z. Shen, *Weyl-Heisenberg frames and Riesz bases in* $L_2(R^d)$, Duke Math. J., 89 (1997), pp. 237–282.
[48] L. L. Schumaker, *Spline Functions: Basic Theory*, John Wiley & Sons, New York, 1981.
[49] J. Sjöstrand, *An algebra of pseudodifferential operators*, Math. Res. Lett., 1 (1994), pp. 185–192.
[50] Q. Sun, *Wiener's lemma for infinite matrices with polynomial off-diagonal decay*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 567–570.
[51] Q. Sun, *Wiener's lemma for infinite matrices with polynomial and subexponential off-diagonal decay*, Trans. Amer. Math. Soc., to appear.

[52] Q. Sun, *Frame in spaces with finite rate of innovation*, Adv. Comput. Math., http://www.springerlink.com/content/rxg9553u61641005 (2006).
[53] W. Sun and X. Zhou, *Average sampling in shift invariant subspaces with symmetric averaging functions*, J. Math. Anal. Appl., 287 (2003), pp. 279–295.
[54] W. Sun and X. Zhou, *Reconstruction of band-limited functions from local averages*, Constr. Approx., 18 (2002), pp. 205–222.
[55] W. Sun and X. Zhou, *Average sampling theorems for shift-invariant subspaces*, Sci. China Ser. E 43 (2000), pp. 524–530.
[56] G. Thomas, *A comparison of motion-compensated interlace-to-progressive conversion method*, Image Communication, 12 (1998), pp. 209–229.
[57] M. Unser, *Sampling—50 years after Shannon*, Proc. IEEE, 88 (2000), pp. 569–587.
[58] M. Vetterli, P. Marziliano, and T. Blu, *Sampling signals with finite rate of innovation*, IEEE Trans. Signal Process., 50 (2002), pp. 1417–1428.
[59] R. H. Vollmerhausen and R. G. Driggers, *Analysis of Sampled Imaging Systems*, SPIE Press, Bellingham, WA, 2000.
[60] G. G. Walter, *A sampling theorem for wavelet subspaces*, IEEE Trans. Inform. Theory, 38 (1992), pp. 881–884.
[61] R. G. Wiley, *Recovery of bandlimited signals from unequally spaced samples*, IEEE Trans. Comm., 26 (1978), pp. 135–137.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1423–1449

ON THE PARTIAL DIFFERENTIAL EQUATIONS OF ELECTROSTATIC MEMS DEVICES: STATIONARY CASE*

NASSIF GHOUSSOUB† AND YUJIN GUO†

Abstract. We analyze the nonlinear elliptic problem $\Delta u = \frac{\lambda f(x)}{(1+u)^2}$ on a bounded domain Ω of R^N with Dirichlet boundary conditions. This equation models a simple electrostatic micro-electromechanical system (MEMS) device consisting of a thin dielectric elastic membrane with boundary supported at 0 above a rigid ground plate located at -1. When a voltage—represented here by λ—is applied, the membrane deflects towards the ground plate, and a snap-through may occur when it exceeds a certain critical value λ^* (pull-in voltage). This creates a so-called pull-in instability, which greatly affects the design of many devices. The mathematical model leads to a nonlinear parabolic problem for the dynamic deflection of the elastic membrane, which will be considered in a forthcoming paper. Here, we focus on the stationary equation and on estimates for λ^* in terms of material properties of the membrane, which can be fabricated with a spatially varying dielectric permittivity profile f. Applying analytical and numerical techniques, we establish upper and lower bounds for λ^* in terms of the volume and shape of the domain, the dimension of the ambient space, as well as the permittivity profile. We analyze the first branch of stable steady states when $\lambda < \lambda^*$ and prove that a semistable (extremal) solution exists at $\lambda = \lambda^*$ in dimension $1 \le N \le 7$, and that classical extremal solutions may not exist for dimension $N \ge 8$. More refined properties of stable steady states—such as regularity, stability, uniqueness, multiplicity, energy estimates, and comparison results—are also established. The analysis of branches of unstable solutions is more elaborate and is tackled in the companion paper [P. Esposito, N. Ghoussoub, and Y. Guo, *Comm. Pure Appl. Math.*, (2006), to appear].

Key words. MEMS, pull-in voltage, power law permittivity profile, minimal solutions

AMS subject classifications. 35J60, 35B40, 35J20

DOI. 10.1137/050647803

1. Introduction.

Micro-electromechanical systems (MEMS) are often used to combine electronics with microsize mechanical devices in the design of various types of microscopic machinery. MEMS devices have therefore become key components of many commercial systems, including accelerometers for airbag deployment in automobiles, ink jet printer heads, optical switches, chemical sensors, and so on (see, for example, [20]). The key component of many modern MEMS is a simple idealized electrostatic device consisting of a thin and deformable elastic membrane that is held fixed along its boundary and which lies above a rigid grounded plate. This elastic membrane is modeled as a dielectric with a small but finite thickness. The upper surface of the membrane is coated with a negligibly thin metallic conducting film. When a voltage V is applied to the conducting film, the thin dielectric membrane deflects towards the bottom plate, and when V is increased beyond a certain critical value V^*—known as the pull-in voltage—the steady state of the elastic membrane is lost and proceeds to touchdown or snap through at a finite time, creating the so-called pull-in instability.

A mathematical model of the physical phenomena, leading to a partial differential

*Received by the editor December 16, 2005; accepted for publication (in revised form) June 12, 2006; published electronically January 8, 2007. The first author was partially supported by the Natural Science and Engineering Research Council of Canada. The second author was partially supported by the Natural Science Foundation of P. R. China (10171036) and by a U.B.C. Graduate Fellowship.

http://www.siam.org/journals/sima/38-5/64780.html

†Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2 (nassif@math.ubc.ca, yjguo@math.ubc.ca).

equation for the dimensionless dynamic deflection of the membrane, was derived and analyzed in [10] and [12]. In the damping-dominated limit, and using a narrow-gap asymptotic analysis, the dimensionless dynamic deflection $u = u(x,t)$ of the membrane on a bounded domain Ω in R^2 is found to satisfy the following parabolic problem:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Delta u - \frac{\lambda f(x)}{(1+u)^2} \quad \text{for} \quad x \in \Omega, \\ u(x,0) &= 0 \qquad\qquad\qquad\quad \text{for} \quad x \in \Omega, \\ u(x,t) &= 0 \qquad\qquad\qquad\quad \text{for} \quad x \in \partial\Omega. \end{aligned} \tag{1.1}$$

The initial condition in (1.1) assumes that the membrane is initially undeflected and that the voltage is suddenly applied to the upper surface of the membrane at time $t = 0$. The parameter $\lambda > 0$ in (1.1) characterizes the relative strength of the electrostatic and mechanical forces in the system, and is given in terms of the applied voltage V by $\lambda = \frac{\varepsilon_0 V^2 L^2}{2T_e d^3}$, where d is the undeflected gap size, L is the length scale of the membrane, T_e is the tension of the membrane, and ε_0 is the permittivity of free space in the gap between the membrane and the bottom plate. From now on we shall use the parameter λ and λ^* to represent the applied voltage V and pull-in voltage V^*, respectively. Referred to as the *permittivity profile*, $f(x)$ in (1.1) is defined by the ratio $f(x) = \frac{\varepsilon_0}{\varepsilon_2(x)}$, where $\varepsilon_2(x)$ is the dielectric permittivity of the thin membrane.

There are several issues that must be considered in the actual design of MEMS devices. Typically one of the primary goals is to achieve the maximum possible stable deflection before touchdown occurs, which is referred to as the *pull-in distance* (cf. [12] and [19]). Another consideration is to increase the stable operating range of the device by improving the pull-in voltage λ^* subject to the constraint that the range of the applied voltage is limited by the available power supply. Such improvements in the stable operating range are important for the design of certain MEMS devices such as microresonators. One way of achieving larger values of λ^*, while simultaneously increasing the pull-in distance, was studied in [19] and [12] and consists of introducing a spatially varying dielectric permittivity $\varepsilon_2(x)$ of the membrane. The idea is to locate the region where the membrane deflection would normally be largest under a spatially uniform permittivity, and then make sure that a new dielectric permittivity $\varepsilon_2(x)$ is largest—and consequently the profile $f(x)$ smallest—in that region.

Pelesko studied in [19] the steady states of (1.1), when $f(x)$ is assumed to be bounded away from zero, i.e., $0 < C \le f(x) \le 1$ for all $x \in \Omega$. He established in this case an upper bound for λ^*, and derived numerical results for the power-law permittivity profile, from which the larger pull-in voltage and thereby the larger pull-in distance, existence, and multiplicity of the steady states were observed. From the strictly mathematical point of view, it turned out that—at least for $f \equiv 1$—there already exist in the literature many interesting results concerning the properties of the branch of semistable solutions for Dirichlet boundary value problems of the form $-\Delta u = \lambda h(u)$, where h is a regular nonlinearity (e.g., $h(u) = e^u$ or $(1+u)^p$ for $p > 1$). See, for example, the seminal papers [8, 14, 15] and also [6] for a survey on the subject and an exhaustive list of related references. After a first version of this paper was circulated, X. Cabre informed us that even the case of singular nonlinearities involved in MEMS devices had already been considered in [4] and in a more general context in [17].

Recently, Guo, Pan, and Ward studied in [12] the dynamic behavior of (1.1), which is also of great practical interest. They considered a more general class of profiles $f(x)$, where the membrane is allowed to be perfectly conducting, i.e., $0 \le$

$f(x) \le 1$ for all $x \in \Omega$, with $f(x) > 0$ on a subset of positive measure of Ω. By using both analytical and numerical techniques, they obtained larger pull-in voltage λ^* and larger pull-in distance for different classes of varying permittivity profiles. Besides the above practical considerations, the model turned out to be a very rich source of interesting mathematical phenomena. Numerics give lots of information and point to many conjectures, but the available arsenal of nonlinear analysis and PDE techniques can tackle only a precious few, even in the case of power-law permittivity profiles $f(x) = |x|^\alpha$.

This paper is a first in a series where we try to provide a rigorous mathematical analysis for various phenomena related to this model, which were observed numerically. Estimates on the pull-in voltage λ^* depend on the size and geometry of the domain, but also on the dimension of the ambient space and the permittivity profile f. A similar dependence occurs for the refined properties of steady states—such as regularity, stability, uniqueness, multiplicity, energy estimates, and comparison results. The same complexity carries to the dynamic case, where issues related to the "touchdown profile"—in finite or infinite time—or to global convergence towards a stable steady state present many interesting mathematical challenges.

In this first paper, we focus on the stable and semistable stationary deflections of the membrane, while the unstable case is considered in [9], and the dynamic case in [11]. For convenience, we shall set $v = -u$ in such a way that our discussion will center on the following elliptic problem:

$$(S)_\lambda \qquad -\Delta v = \frac{\lambda f(x)}{(1-v)^2} \quad \text{and} \quad 0 < v < 1 \ \text{ for } x \in \Omega\,, \ \text{ while } v = 0 \ \text{ on } \ \partial\Omega.$$

Throughout the paper and unless mentioned otherwise, solutions for $(S)_\lambda$ will be taken in the classical sense. The permittivity profile f will be allowed to vanish somewhere, and will be assumed to satisfy

$$\begin{aligned} &f \in C^\alpha(\bar\Omega) \quad \text{for some } \alpha \in (0,1],\ 0 \le f \le 1, \text{ and} \\ &f > 0 \quad \text{on a subset of } \Omega \text{ of positive measure.} \end{aligned} \tag{1.2}$$

This paper is organized as follows. In section 2 we mainly show the existence of a specific pull-in voltage and study its dependence on the size and shape of the domain as well as on the permittivity profile. These monotonicity properties will help us establish in section 3 new lower and upper bound estimates on the pull-in voltage. We shall write $|\Omega|$ for the volume of a domain Ω in R^N and $P(\Omega) := \int_{\partial\Omega} ds$ for its "perimeter," with ω_N referring to the volume of the unit ball $B_1(0)$ in R^N. We denote by μ_Ω the first eigenvalue of $-\Delta$ on $H_0^1(\Omega)$, and by ϕ_Ω the corresponding positive eigenfunction normalized with $\int_\Omega \phi_\Omega dx = 1$.

THEOREM 1.1. *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω; then there exists a finite pull-in voltage $\lambda^* := \lambda^*(\Omega, f) > 0$ such that we have the following:*

1. *If $0 \le \lambda < \lambda^*$, there exists at least one solution for $(S)_\lambda$.*
2. *If $\lambda > \lambda^*$, there is no solution for $(S)_\lambda$.*
3. *The following bounds on λ^* hold for any bounded domain Ω:*

$$\max\left\{\frac{8N}{27},\ \frac{6N-8}{9}\right\} \frac{1}{\sup_\Omega f} \left(\frac{\omega_N}{|\Omega|}\right)^{\frac{2}{N}} \le \lambda^*(\Omega), \tag{1.3}$$

$$\min\left\{\bar\lambda_1 := \frac{4\mu_\Omega}{27 \inf\limits_{x\in\Omega} f(x)}, \bar\lambda_2 := \frac{\mu_\Omega}{3\int_\Omega f\phi_\Omega\, dx}\right\} \ge \lambda^*(\Omega). \tag{1.4}$$

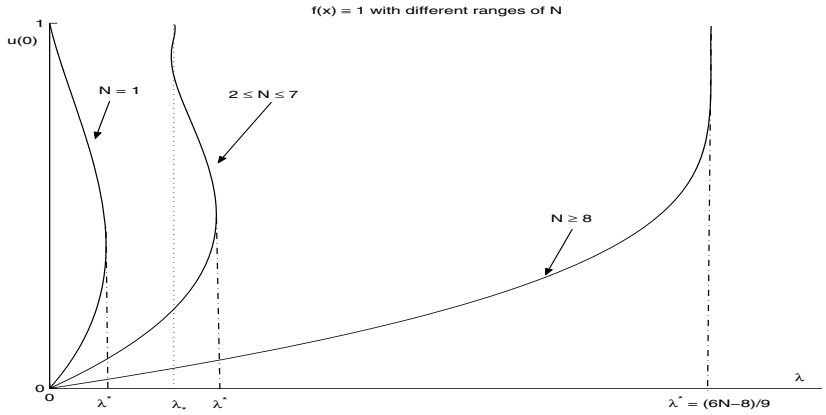


FIG. 1.1. *Plots of $u(0)$ versus λ for the constant permittivity profile $f(x) \equiv 1$ defined in the unit ball $B_1(0) \subset R^N$ with different ranges of N. In the case of $N \geq 8$, we have $\lambda^* = (6N-8)/9$.*

4. *If Ω is a strictly star-shaped domain, that is if $x \cdot \nu(x) \geq a > 0$ for all $x \in \partial\Omega$, where $\nu(x)$ is the unit outer normal at $x \in \partial\Omega$, and if $f \equiv 1$, then*

$$\lambda^*(\Omega) \leq \bar{\lambda}_3 = \frac{(N+2)^2 P(\Omega)}{8aN|\Omega|}.$$

In particular, if $\Omega = B_1(0) \subset R^N$, then we have the bound

$$\lambda^*(B_1(0)) \leq \frac{(N+2)^2}{8}.$$

5. *If $f(x) \equiv |x|^\alpha$ with $\alpha \geq 0$ and Ω is a ball of radius R, then we have*

$$\lambda^*(B_R, |x|^\alpha) \geq \max\left\{\frac{4(2+\alpha)(N+\alpha)}{27R^{(2+\alpha)}}, \frac{(2+\alpha)(3N+\alpha-4)}{9R^{(2+\alpha)}}\right\}. \tag{1.5}$$

Moreover, if $N \geq 8$ and $0 \leq \alpha \leq \alpha^{**}(N) := \frac{4-6N+3\sqrt{6}(N-2)}{4}$, we have

$$\lambda^*(B_1, |x|^\alpha) = \frac{(2+\alpha)(3N+\alpha-4)}{9}. \tag{1.6}$$

In section 3.3 we give some numerical estimates on λ^* to compare them with the analytic bounds given in Theorem 1.1 above. Note that the upper bound $\bar{\lambda}_1$ is relevant only when f is bounded away from 0, while the upper bound $\bar{\lambda}_2$ is valid for all permittivity profiles. However, the order between these two upper bounds can vary in general. For example, in the case of exponential permittivity profiles of the form $f(x) = e^{\alpha(|x|^2-1)}$ on the unit disc, one can see that $\bar{\lambda}_1$ is a better upper bound than $\bar{\lambda}_2$ for small α, while the reverse holds true for larger values of α. The lower bounds in (1.3) and (1.5) can be improved in small dimensions, but they are optimal—at least for the ball—in dimension larger than 8.

We also consider issues of uniqueness and multiplicity of solutions for $(S)_\lambda$ with $0 < \lambda \leq \lambda^*$. The bifurcation diagrams in Figure 1.1 show the complexity of the situation, even in the radially symmetric case. One can see that the number of branches—and of solutions—is closely connected to the space dimension, a fact that

we establish analytically in section section 4, by focussing on the very first branch of solutions considered to be "minimal" in the following way.

DEFINITION 1.2. *A solution $u_\lambda(x)$ of $(S)_\lambda$ is said to be minimal if for any other solution u of $(S)_\lambda$ we have $u_\lambda(x) \le u(x)$ for all $x \in \Omega$.*

One can also consider for any solution u of $(S)_\lambda$ the linearized operator at u defined by $L_{u,\lambda} = -\Delta - \frac{2\lambda f(x)}{(1-u)^3}$ and its eigenvalues $\{\mu_{k,\lambda}(u); k = 1, 2, \ldots\}$. The first eigenvalue is then simple and is given by

$$\mu_{1,\lambda}(u) = \inf \left\{ \langle L_{u,\lambda}\phi, \phi \rangle_{H_0^1(\Omega)} \, ; \, \phi \in C_0^\infty(\Omega), \int_\Omega |\phi(x)|^2 dx = 1 \right\}.$$

Stable solutions (*resp.*, *semi-stable solutions*) of $(S)_\lambda$ are those solutions u such that $\mu_{1,\lambda}(u) > 0$ (*resp.*, $\mu_{1,\lambda}(u) \ge 0$). Our main results in this direction can be stated as follows.

THEOREM 1.3. *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω, and consider $\lambda^* := \lambda^*(\Omega, f)$ as defined in Theorem* 1.1. *Then the following hold:*

1. *For any $0 \le \lambda < \lambda^*$ there exists a unique minimal solution u_λ of $(S)_\lambda$ such that $\mu_{1,\lambda}(u_\lambda) > 0$. Moreover, for each $x \in \Omega$ the function $\lambda \to u_\lambda(x)$ is strictly increasing and differentiable on $(0, \lambda^*)$.*
2. *If $1 \le N \le 7$, then one has $\sup_{\lambda \in (0,\lambda^*)} \| u_\lambda \|_\infty < 1$, and consequently $u^* = \lim_{\lambda \uparrow \lambda^*} u_\lambda$ exists in $C^{2,\alpha}(\bar{\Omega})$ and is a solution for $(S)_{\lambda^*}$ such that $\mu_{1,\lambda^*}(u^*) = 0$. In particular, u^*—often referred to as the extremal solution of problem $(S)_\lambda$—is unique.*
3. *On the other hand, if $N \ge 8$, $f(x) = |x|^\alpha$ with $0 \le \alpha \le \alpha^{**}(N) := \frac{4-6N+3\sqrt{6}(N-2)}{4}$, and Ω is the unit ball, then the extremal solution is necessarily $u^*(x) = 1 - |x|^{\frac{2+\alpha}{3}}$ and is therefore singular.*

We note that, in general, the function u^* exists in any dimension, does solve $(S)_{\lambda^*}$ in a suitable weak sense, and is the unique solution in an appropriate class. The above theorem says that it is, however, a classical solution in dimensions $1 \le N \le 7$, and this will allow us to start another branch of nonminimal (unstable) solutions. Indeed, we show in section 5—following ideas of Crandall and Rabinowitz [8]—that, for $1 \le N \le 7$ and for λ close enough to λ^*, there exists a unique second branch U_λ of solutions for $(S)_\lambda$, bifurcating from u^*, with

$$\mu_{1,\lambda}(U_\lambda) < 0 \quad \text{while } \mu_{2,\lambda}(U_\lambda) > 0. \tag{1.7}$$

In the companion paper [9], we shall provide a variational (mountain pass) characterization of these unstable solutions, and more importantly we establish—under the same dimension restriction as above—a compactness result along the second branch of unstable solutions leading to a—nonzero—second bifurcation point.

Issues of uniqueness, multiplicity, and other qualitative properties of the solutions for $(S)_\lambda$ are still far from being well understood, even in the radially symmetric case which we consider in section 6. Some of the classical work of Joseph and Lundgren [14] and many that followed can be adapted to this situation when the permittivity profile is constant. However, the case of a power-law permittivity profile $f(x) = |x|^\alpha$ defined in a unit ball already presents a much richer situation. In section 6 we present some numerical evidence for various conjectures relating to this case, some of which have been tackled in [9]. A detailed and involved analysis of compactness along the unstable branches will be discussed there, as well as some information about the second bifurcation point.

2. The pull-in voltage. In this section, we establish the existence and some monotonicity properties for the pull-in voltage λ^*, which is defined as

$$\lambda^*(\Omega, f) = \sup\{\lambda > 0 \mid (S)_\lambda \text{ possesses at least one solution}\}. \tag{2.1}$$

2.1. Existence of the pull-in voltage. For any bounded domain Γ in R^N, we denote by μ_Γ the first eigenvalue of $-\Delta$ on $H_0^1(\Gamma)$ and by ψ_Γ the corresponding positive eigenfunction normalized with $\sup_{x\in\Gamma}\psi_\Gamma = 1$. We also associate with any domain Ω in R^N the following parameter:

$$\nu_\Omega = \sup\Big\{\mu_\Gamma H(\inf_\Omega \psi_\Gamma);\ \Gamma \text{ domain of } R^N,\ \Gamma \supset \bar{\Omega}\Big\}, \tag{2.2}$$

where H is the function $H(t) = \frac{t(t+1+2\sqrt{t})}{(t+1+\sqrt{t})^3}$.

THEOREM 2.1. *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω in R^N; then there exists a finite pull-in voltage $\lambda^* := \lambda^*(\Omega, f) > 0$ such that*

1. *if $\lambda < \lambda^*$, there exists at least one solution for $(S)_\lambda$;*
2. *if $\lambda > \lambda^*$, there is no solution for $(S)_\lambda$.*

Moreover, we have the lower bound

$$\lambda^*(\Omega, f) \geq \frac{\nu_\Omega}{\sup_{x\in\Omega} f(x)}. \tag{2.3}$$

Proof. We need to show that $(S)_\lambda$ has at least one solution when $\lambda < \frac{\nu_\Omega}{\sup_\Omega f(x)}$. It is clear that $u \equiv 0$ is a subsolution of $(S)_\lambda$ for all $\lambda > 0$. To construct a super-solution, consider a bounded domain $\Gamma \supset \bar{\Omega}$ with smooth boundary, and let $(\mu_\Gamma, \psi_\Gamma)$ be its first eigenpair normalized in such a way that

$$\sup_{x\in\Gamma}\psi_\Gamma(x) = 1 \quad \text{and} \quad \inf_{x\in\Omega}\psi_\Gamma(x) := s_1 > 0.$$

We construct a supersolution in the form $\psi = A\psi_\Gamma$, where A is a scalar to be chosen later. First, we must have $A\psi_\Gamma \geq 0$ on $\partial\Omega$ and $0 < 1 - A\psi_\Gamma < 1$ in Ω, which requires that $0 < A < 1$. We also require

$$-\Delta\psi - \frac{\lambda f(x)}{(1-A\psi)^2} \geq 0 \quad \text{in } \Omega, \tag{2.4}$$

which can be satisfied as long as $\mu_\Gamma A\ \psi_\Gamma \geq \frac{\lambda \sup_\Omega f(x)}{(1-A\ \psi_\Gamma)^2}$ in Ω or

$$\lambda \sup_\Omega f(x) < \beta(A,\Gamma) := \mu_\Gamma \inf\{g(sA);\ s \in [s_1(\Gamma), 1]\}, \tag{2.5}$$

where $g(s) = s(1-s)^2$. In other words, $\lambda^* \sup_\Omega f(x) \geq \sup\{\beta(A,\Gamma);\ 0 < A < 1,\ \Gamma \supset \bar{\Omega}\}$, and therefore it remains to show that

$$\nu_\Omega = \sup\{\beta(A,\Gamma);\ 0 < A < 1, \Gamma \supset \bar{\Omega}\}. \tag{2.6}$$

For that, we note first that $\inf_{s\in[s_1,1]} g(As) = \min\big\{g(As_1), g(A)\big\}$. We also have that $g(As_1) \leq g(A)$ if and only if $A^2(s_1^3 - 1) - 2A(s_1^2 - 1) + (s_1 - 1) \leq 0$, which happens if and only if $A^2(s_1^2 + s_1 + 1) - 2A(s_1 + 1) + 1 \geq 0$ or if and only if either $A \leq A_-$ or $A \geq A_+$, where

$$A_+ = \frac{s_1 + 1 + \sqrt{s_1}}{s_1^2 + 1 + s_1} = \frac{1}{s_1 + 1 - \sqrt{s_1}}, \qquad A_- = \frac{s_1 + 1 - \sqrt{s_1}}{s_1^2 + 1 + s_1} = \frac{1}{s_1 + 1 + \sqrt{s_1}}.$$

Since $A_- < 1 < A_+$, we get that

$$G(A) = \inf_{s\in[s_1,1]} g(As) = \begin{cases} g(As_1) & \text{if } 0 \le A \le A_-, \\ g(A) & \text{if } A_- \le A \le 1. \end{cases} \tag{2.7}$$

We now have that $\frac{dG}{dA} = g'(As_1)s_1 \ge 0$ for all $0 \le A \le A_-$. And since $A_- \ge \frac{1}{3}$, we have $\frac{dG}{dA} = g'(A) \le 0$ for all $A_- \le A \le 1$. It follows that

$$\begin{aligned} \sup_{0<A<1} \inf_{s\in[s_1,1]} g(As) &= \sup_{0<A<1} G(A) = G(A_-) = g(A_-) \\ &= \frac{1}{s_1+1+\sqrt{s_1}}\left(1 - \frac{1}{s_1+1+\sqrt{s_1}}\right)^2 \\ &= \frac{s_1(s_1+1+2\sqrt{s_1})}{(s_1+1+\sqrt{s_1})^3} = H\left(\inf_\Omega \psi_\Gamma\right), \end{aligned}$$

which proves our lower estimate.

Now we know that $\lambda^* > 0$, so we can pick $\lambda \in (0, \lambda^*)$ and use the definition of λ^* to find a $\bar\lambda \in (\lambda, \lambda^*)$ such that $(S)_{\bar\lambda}$ has a solution $u_{\bar\lambda}$; i.e.,

$$-\Delta u_{\bar\lambda} = \tfrac{\bar\lambda f(x)}{(1-u_{\bar\lambda})^2},\ 0 \le u_{\bar\lambda} < 1, \text{ on } \Omega \quad \text{and} \quad u_{\bar\lambda} = 0 \text{ on } \partial\Omega,$$

and in particular $-\Delta u_{\bar\lambda} \ge \frac{\lambda f(x)}{(1-u_{\bar\lambda})^2}$ on Ω, which then implies that $u_{\bar\lambda}$ is a supersolution of $(S)_\lambda$. Since $u \equiv 0$ is a subsolution of $(S)_\lambda$, we can again conclude that there is a solution u_λ of $(S)_\lambda$.

It is also easy to show that λ^* is finite, since if $(S)_\lambda$ has at least one solution $0 < u < 1$, then, by integrating against the first (positive) eigenfunction ϕ_Ω, we get

$$\begin{aligned} +\infty > \mu_\Omega \ge \mu_\Omega \int_\Omega u\phi_\Omega &= -\int_\Omega u\Delta\phi_\Omega = -\int_\Omega \phi_\Omega \Delta u \\ &= \lambda \int_\Omega \frac{\phi_\Omega f}{(1-u)^2} dx \ge \lambda \int_\Omega \phi_\Omega f\, dx \end{aligned} \tag{2.8}$$

and therefore $\lambda^* < +\infty$. The definition of λ^* implies that there is no solution of $(S)_\lambda$ for any $\lambda > \lambda^*$. □

2.2. Monotonicity results for the pull-in voltage. In this subsection, we give a more precise characterization of λ^*, namely as the endpoint for the branch of minimal solutions. This will allow us to establish various monotonicity properties for λ^* that will help in the estimates given in the next subsections. First we give a recursive scheme for the construction of minimal solutions.

THEOREM 2.2. *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω in R^N; then for any $0 < \lambda < \lambda^*(\Omega, f)$ there exists a unique minimal positive solution u_λ for $(S)_\lambda$. It is obtained as the limit of the sequence $\{u_n(\lambda; x)\}$ constructed recursively as follows: $u_0 \equiv 0$ in Ω and, for each $n \ge 1$,*

$$\begin{aligned} -\Delta u_n &= \frac{\lambda f(x)}{(1-u_{n-1})^2}, \quad x \in \Omega; \\ 0 \le u_n < 1, \quad x \in \Omega; &\qquad u_n = 0, \qquad x \in \partial\Omega. \end{aligned} \tag{2.9}$$

Proof. Let u be any positive solution for $(S)_\lambda$, and consider the sequence $\{u_n(\lambda;x)\}$ defined in (2.9). Clearly $u(x) > u_0 \equiv 0$ in Ω, and whenever $u(x) \geq u_{n-1}$ in Ω, then

$$-\Delta(u-u_n) = \lambda f(x)\left[\frac{1}{(1-u)^2} - \frac{1}{(1-u_{n-1})^2}\right] \geq 0, \quad x \in \Omega\,; \quad u-u_n = 0, \quad x \in \partial\Omega.$$

The maximum principle and an immediate induction yield that $1 > u(x) \geq u_n$ in Ω for all $n \geq 0$. In a similar way, the maximum principle implies that the sequence $\{u_n(\lambda;x)\}$ is monotone increasing. Therefore, $\{u_n(\lambda;x)\}$ converges uniformly to a positive solution $u_\lambda(x)$, satisfying $u(x) \geq u_\lambda(x)$ in Ω, which is a minimal positive solution of $(S)_\lambda$. It is also clear that u_λ is unique in this class of solutions. □

Remark 2.1. Let $g(x,\xi,\Omega)$ be Green's function of the Laplace operator, with $g(x,\xi,\Omega) = 0$ on $\partial\Omega$. Then the iteration in (2.9) can be replaced by $u_0 \equiv 0$ in Ω, and for each $n \geq 1$,

$$u_n(\lambda;x) = \lambda \int_\Omega \frac{f(\xi)g(x,\xi,\Omega)}{(1-u_{n-1}(\lambda;\xi))^2} d\xi, \quad x \in \Omega\,; \qquad u_n(\lambda;x) = 0, \quad x \in \partial\Omega. \tag{2.10}$$

The same reasoning as above yields that $\lim_{n\to\infty} u_n(\lambda;x) = u_\lambda(x)$ for all $x \in \Omega$.

The above construction of solutions yields the following monotonicity result for the pull-in voltage.

PROPOSITION 2.3. *If $\Omega_1 \subset \Omega_2$ and if f is a function satisfying* (1.2) *on Ω_2, then $\lambda^*(\Omega_1) \geq \lambda^*(\Omega_2)$ and the corresponding minimal solutions satisfy $u_{\Omega_1}(\lambda,x) \leq u_{\Omega_2}(\lambda,x)$ on Ω_1 for every $0 < \lambda < \lambda^*(\Omega_2)$.*

Proof. Again the method of sub/supersolutions immediately yields that $\lambda^*(\Omega_1) \geq \lambda^*(\Omega_2)$. Now consider, for $i = 1, 2$, the sequences $\{u_n(\lambda,x,\Omega_i)\}$ on Ω_i defined by (2.10), where $g(x,\xi,\Omega_i)$ are the corresponding Green's functions on Ω_i. Since $\Omega_1 \subset \Omega_2$, we have that $g(x,\xi,\Omega_1) \leq g(x,\xi,\Omega_2)$ on Ω_1. Hence, it follows that

$$u_1(\lambda,x,\Omega_2) = \lambda \int_{\Omega_2} f(\xi)g(x,\xi,\Omega_2)d\xi \geq \lambda \int_{\Omega_1} f(\xi)g(x,\xi,\Omega_1)d\xi = u_1(\lambda,x,\Omega_1)$$

on Ω_1. By induction we conclude that $u_n(\lambda,x,\Omega_2) \geq u_n(\lambda,x,\Omega_1)$ on Ω_1 for all n. On the other hand, since $u_n(\lambda,x,\Omega_2) \leq u_{n+1}(\lambda,x,\Omega_2)$ on Ω_2 for n, we get that $u_n(\lambda,x,\Omega_1) \leq u_{\Omega_2}(\lambda,x)$ on Ω_1, and we are done. □

We also note the following easy comparison results, and we omit the details.

COROLLARY 2.4. *Suppose that $f_1, f_2 : \Omega \to R$ are two functions satisfying* (1.2) *and such that $f_1(x) \leq f_2(x)$ on Ω; then $\lambda^*(\Omega,f_1) \geq \lambda^*(\Omega,f_2)$, and for $0 < \lambda < \lambda^*(\Omega,f_2)$ we have $u_1(\lambda,x) \leq u_2(\lambda,x)$ on Ω, where $u_1(\lambda,x)$ (resp., $u_2(\lambda,x)$) are the unique minimal positive solution of*

$$-\Delta u = \frac{\lambda f_1(x)}{(1-u)^2} \ \left(\textit{resp.}, -\Delta u = \frac{\lambda f_2(x)}{(1-u)^2}\right) \textit{ on } \Omega \quad \textit{and} \quad u = 0 \textit{ on } \partial\Omega.$$

Moreover, if $f_2(x) > f_1(x)$ on a subset of positive measure, then $u_1(\lambda,x) < u_2(\lambda,x)$ for all $x \in \Omega$.

The following result is similar to Theorem 4.10 of [2], which deals only with nonsingular nonlinearities. It can, however, be extended to our setting, but the adaptation of the proof is left to the interested reader.

PROPOSITION 2.5. *For any bounded domain Γ in R^N and any function f satisfying* (1.2) *on Γ, we have $\lambda^*(\Gamma,f) \geq \lambda^*(B_R,f^*)$, where $B_R = B_R(0)$ is the Euclidean ball in R^N with radius $R > 0$ and with volume $|B_R| = |\Gamma|$, and where f^* is the Schwarz symmetrization of f.*

3. Estimates for the pull-in voltage. While the lower bound in (2.3) is useful to prove existence, it is not easy to compute. The following subsection gives more computationally accessible lower estimates for λ^*.

3.1. Lower bounds for λ^*.

PROPOSITION 3.1. *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω in R^N; then we have the following lower bound:*

$$\lambda^*(\Omega, f) \geq \max\left\{\frac{8N}{27}, \frac{6N-8}{9}\right\}\frac{1}{\sup_\Omega f}\left(\frac{\omega_N}{|\Omega|}\right)^{\frac{2}{N}}. \tag{3.1}$$

Moreover, if $f(x) \equiv |x|^\alpha$ with $\alpha \geq 0$ and Ω is a ball of radius R, then we have

$$\lambda^*(B_R, |x|^\alpha) \geq \max\left\{\frac{4(2+\alpha)(N+\alpha)}{27}, \frac{(2+\alpha)(3N+\alpha-4)}{9}\right\}R^{-(2+\alpha)}. \tag{3.2}$$

Finally, if $N \geq 8$ and $0 \leq \alpha \leq \alpha^{**}(N) := \frac{4-6N+3\sqrt{6}(N-2)}{4}$, we have

$$\lambda^*(B_1, |x|^\alpha) = \frac{(2+\alpha)(3N+\alpha-4)}{9}. \tag{3.3}$$

Proof. Setting $R = \left(\frac{|\Omega|}{\omega_N}\right)^{\frac{1}{N}}$, it suffices—in view of Proposition 2.5 and since $\sup_{B_R} f^* = \sup_\Omega f$—to show that

$$\lambda^*(B_R, f^*) \geq \max\left\{\frac{8N}{27R^2\sup_\Omega f^*}, \frac{6N-8}{9R^2\sup_\Omega f^*}\right\} \tag{3.4}$$

for the case where $\Omega = B_R$. In fact, the function $w(x) = \frac{1}{3}(1 - \frac{|x|^2}{R^2})$ satisfies on B_R

$$-\Delta w = \frac{2N}{3R^2} = \frac{2N(1-\frac{1}{3})^2}{3R^2}\frac{1}{(1-\frac{1}{3})^2} \geq \frac{8N}{27R^2\sup_\Omega f}\frac{f(x)}{[1-\frac{1}{3}(1-\frac{|x|^2}{R^2})]^2}$$
$$= \frac{8N}{27R^2\sup_\Omega f}\frac{f(x)}{(1-w)^2}.$$

So for $\lambda \leq \frac{8N}{27R^2\sup_\Omega f}$, w is a supersolution of $(S)_\lambda$ in B_R. Since, on the other hand, $w_0 \equiv 0$ is a subsolution of $(S)_\lambda$ and $w_0 \leq w$ in B_R, then there exists a solution of $(S)_\lambda$ in B_R, which proves a part of (3.4).

A similar computation applied to the function $v(x) = 1 - (\frac{|x|}{R})^{\frac{2}{3}}$ shows that v is also a supersolution as long as $\lambda \leq \frac{6N-8}{9R^2\sup_\Omega f}$.

In order to prove (3.2), it suffices to note that $w(x) = \frac{1}{3}\left(1 - \frac{|x|^{2+\alpha}}{R^{2+\alpha}}\right)$ is a supersolution for $(S)_\lambda$ on B_R, provided $\lambda \leq \frac{4(2+\alpha)(N+\alpha)}{27R^{2+\alpha}}$, and that $v(x) = 1 - (\frac{|x|}{R})^{\frac{2+\alpha}{3}}$ is a supersolution for $(S)_\lambda$ on B_R, provided $\lambda \leq \frac{(2+\alpha)(3N+\alpha-4)}{9R^{2+\alpha}}$.

In order to complete the proof of Proposition 3.1, we need to establish that the function $u^*(x) = 1 - |x|^{\frac{2+\alpha}{3}}$ is the extremal function as long as $N \geq 8$ and $0 \leq \alpha \leq \alpha^{**}(N) = \frac{4-6N+3\sqrt{6}(N-2)}{4}$. This will then yield that for such dimensions and these values of α, the voltage $\lambda = \frac{(2+\alpha)(3N+\alpha-4)}{9}$ is exactly the pull-in voltage λ^*.

First, it is easy to check that u^* is a $H_0^1(\Omega)$-weak solution of $(S)_{\lambda^*}$. Since $\|u^*\|_\infty = 1$, and by the characterization of Theorem 5.1 below, we need to prove only that

$$\int_\Omega |\nabla\phi|^2 \geq \int_\Omega \frac{2\lambda|x|^\alpha}{(1-u^*)^3}\phi^2 \quad \text{for all } \phi \in H_0^1(\Omega). \tag{3.5}$$

However, Hardy's inequality gives for $N \geq 2$ that $\int_{B_1} |\nabla\phi|^2 \geq \frac{(N-2)^2}{4}\int_{B_1} \frac{\phi^2}{|x|^2}$ for any $\phi \in H_0^1(B_1)$, which means that (3.5) holds whenever $2\lambda^* \leq \frac{(N-2)^2}{4}$ or, equivalently, if $N \geq 8$ and $0 \leq \alpha \leq \alpha^{**} = \frac{4-6N+3\sqrt{6}(N-2)}{4}$. □

Remark 3.1. The above lower bounds can be improved at least in low dimensions. First note that if $N > \frac{12+\alpha}{5}$, then $\lambda_2 = \frac{(2+\alpha)(3N+\alpha-4)}{9}$ is the better lower bound and is actually sharp on the ball as soon as $N \geq 8$ and $\alpha \leq \alpha^{**}$. For lower dimensions, the best lower bounds are more complicated even when one considers supersolutions of the form $v(x) = a(1-(\frac{|x|}{R})^k)$ and optimizes $\lambda(a,k,R)$ over a and k. For example, in the case where $\alpha = 0$, $N = 2$, and $R = 1$, one can see that a better lower bound can be obtained via the supersolution $v(x) = \frac{1}{2.4}(1-|x|^{1.6})$.

3.2. Upper bounds for λ^*. We note that (2.8) already yields a finite upper bound for λ^*. However, Pohozaev-type arguments can be used to establish better and more computable upper bounds. In this subsection, we establish the upper estimates claimed in Theorem 1.1. For a general domain Ω, the following upper bounds on $\lambda^*(\Omega)$ were established in Theorem 3.1 of [19] and in Theorem 2.2 of [12], respectively.

PROPOSITION 3.2. (1) *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω in R^N such that* $\inf_\Omega f > 0$*; then*

$$\lambda^*(\Omega, f) \leq \bar{\lambda}_1 \equiv \frac{4\mu_\Omega}{27}\left(\inf_\Omega f\right)^{-1}. \tag{3.6}$$

(2) *If we only suppose that $f > 0$ on a set of positive measure, then*

$$\lambda^*(\Omega, f) \leq \bar{\lambda}_2 \equiv \frac{\mu_\Omega}{3}\left(\int_\Omega f\phi_\Omega\, dx\right)^{-1}. \tag{3.7}$$

If, in addition, $\Omega \subset R^N$ is a strictly star-shaped domain around 0, then we can prove the following estimate.

PROPOSITION 3.3. *Suppose $f \equiv 1$ and that $\Omega \subset R^N$ is a strictly star-shaped smooth domain such that $x \cdot \nu \geq a > 0$ for $x \in \partial\Omega$, where ν is the unit outer normal to $\partial\Omega$, then*

$$\lambda^*(\Omega) \leq \bar{\lambda}_3 = \frac{(N+2)^2 P(\Omega)}{8aN|\Omega|}. \tag{3.8}$$

In particular, if Ω is the Euclidean unit ball in R^N, then we have the bound

$$\lambda^*(B_1(0)) \leq \frac{(N+2)^2}{8}.$$

Proof. Recall Pohozaev's identity: If u is a solution of

$$\Delta u + \lambda g(u) = 0 \quad \text{for } x \in \Omega\,; \qquad u = 0 \quad \text{for } x \in \partial\Omega,$$

then

$$N\lambda\int_\Omega G(u)dx-\frac{N-2}{2}\lambda\int_\Omega ug(u)dx=\frac{1}{2}\int_{\partial\Omega}(x\cdot\nu)\left(\frac{\partial u}{\partial\nu}\right)^2 ds, \tag{3.9}$$

where $G(u)=\int_0^u g(s)ds$. Applying this with $g(u)=\frac{1}{(1-u)^2}$ and $G(u)=\frac{u}{1-u}$ yields

$$\begin{aligned}\frac{\lambda}{2}\int_\Omega\frac{u(N+2-2Nu)}{(1-u)^2}dx&=\frac{1}{2}\int_{\partial\Omega}(x\cdot\nu)\left(\frac{\partial u}{\partial\nu}\right)^2 ds\\&\geq\frac{a}{2P(\Omega)}\left(\int_{\partial\Omega}\frac{\partial u}{\partial\nu}ds\right)^2\\&=\frac{a}{2P(\Omega)}\left(-\int_\Omega\Delta udx\right)^2\\&=\frac{a\lambda^2}{2P(\Omega)}\left(\int_\Omega\frac{dx}{(1-u)^2}\right)^2,\end{aligned} \tag{3.10}$$

where we have used the divergence theorem and Hölder's inequality

$$\int_{\partial\Omega}\frac{\partial u}{\partial\nu}ds\leq\left(\int_{\partial\Omega}\left(-\frac{\partial u}{\partial\nu}\right)^2 ds\right)^{1/2}\left(\int_{\partial\Omega}ds\right)^{1/2}.$$

Since

$$\begin{aligned}\int_\Omega\frac{u(N+2-2Nu)}{(1-u)^2}dx&=\int_\Omega\left[-2N\left(u-\frac{N+2}{4N}\right)^2+\frac{(N+2)^2}{8N}\right]\frac{1}{(1-u)^2}dx\\&\leq\frac{(N+2)^2}{8N}\int_\Omega\frac{dx}{(1-u)^2},\end{aligned}$$

we deduce from (3.10) that

$$\frac{(N+2)^2}{8N}\geq\frac{a\lambda}{P(\Omega)}\int_\Omega\frac{dx}{(1-u)^2}\geq\frac{a\lambda|\Omega|}{P(\Omega)},$$

which implies the upper bound (3.8) for λ^*.

Finally, for the special case where $\Omega=B_1(0)\subset R^N$, we have $a=1$ and $\frac{P(B_1(0))}{\omega_N}=N$ and hence the bound $\lambda^*(B_1(0))\leq\bar\lambda_3=\frac{(N+2)^2}{8}$. □

3.3. Numerical estimates for λ^*. In the computations below we shall consider two choices for the domain Ω,

$$\Omega:[-1/2,1/2]\quad\text{(slab)},\qquad\Omega:x^2+y^2\leq 1\quad\text{(unit disk)}. \tag{3.11}$$

Simple calculations yield that

$$\begin{aligned}&\mu_\Omega=\pi^2,&&\phi_\Omega=\frac{\pi}{2}\sin\left[\pi\left(x+\frac{1}{2}\right)\right]\quad\text{(slab)},\\&\mu_\Omega=z_0^2\approx 5.783,&&\phi_\Omega=\frac{z_0}{J_1(z_0)}J_0(z_0|x|)\quad\text{(unit disk)}.\end{aligned} \tag{3.12}$$

TABLE 3.1
Numerical values for pull-in voltage λ^* with the bounds given in Theorem 1.1, where $f(x)$ is the exponential permittivity profile.

Ω	α	$\underline{\lambda}$	λ^*	$\bar{\lambda}_1$	$\bar{\lambda}_2$
Slab	0	1.185	1.401	1.462	3.290
Slab	1.0	1.185	1.733	1.878	4.023
Slab	3.0	1.185	2.637	3.095	5.965
Slab	6.0	1.185	4.848	6.553	10.50
Unit disk	0	0.593	0.789	0.857	1.928
Unit disk	0.5	0.593	1.153	1.413	2.706
Unit disk	1.0	0.593	1.661	2.329	3.746
Unit disk	3.0	0.593	6.091	17.21	11.86

TABLE 3.2
Numerical values for pull-in voltage λ^* with the bounds given in Theorem 1.1, where $f(x)$ is the power-law permittivity profile.

Ω	α	$\lambda_c(\alpha)$	λ^*	$\bar{\lambda}_1$	$\bar{\lambda}_2$
Slab	0	1.185	1.401	1.462	3.290
Slab	1.0	3.556	4.388	∞	9.044
Slab	3.0	11.851	15.189	∞	28.247
Slab	6.0	33.185	43.087	∞	76.608
Unit disk	0	0.593	0.789	0.857	1.928
Unit disk	1.0	1.333	1.775	∞	3.019
Unit disk	5.0	7.259	9.676	∞	15.82
Unit disk	20	71.70	95.66	∞	161.54

Here J_0 and J_1 are Bessel functions of the first kind, and $z_0 \approx 2.4048$ is the first zero of $J_0(z)$. The bounds $\bar{\lambda}_1$ and $\bar{\lambda}_2$ can be evaluated by substituting (3.12) into (3.6) and (3.7). Notice that $\bar{\lambda}_2$ is, in general, determined only up to a numerical quadrature. Using Newton's method and COLSYS [1], one can also solve the boundary value problem $(S)_\lambda$ and numerically calculate λ^* as the saddle-node point for the following two choices of the permittivity profile:

$$\begin{array}{lllll} \text{slab}: & f(x)=|2x|^\alpha \quad \text{(power-law)}; & f(x)=e^{\alpha(x^2-1/4)} & \text{(exponential)}, \\ \text{unit disk}: & f(x)=|x|^\alpha \quad \text{(power-law)}; & f(x)=e^{\alpha(|x|^2-1)} & \text{(exponential)}, \end{array}$$

where $\alpha \geq 0$. Table 3.1 contains numerical values for λ^* in the case of exponential profiles, while Table 3.2 deals with power-law profiles. What is remarkable is that $\bar{\lambda}_1$ and $\bar{\lambda}_2$ are not comparable even when f is bounded away from 0 and that neither one of them provides the optimal value for λ^*. This leads us to conjecture that there should be a better estimate for λ^*, one involving the distribution of f in Ω, as opposed to the infimum or its average against the first eigenfunction ϕ_Ω.

4. The branch of minimal solutions. The branch of minimal solutions corresponds to the lowest branch in the bifurcation diagram, the one connecting the origin point $\lambda = 0$ to the first fold at $\lambda = \lambda^*$. In this section, we analyze further the properties of this branch. To do so, we consider for each solution u of $(S)_\lambda$ the operator $L_{u,\lambda} = -\Delta - \frac{2\lambda f}{(1-u)^3}$ associated with the linearized problem around u. We denote by $\mu_1(\lambda, u)$ the smallest eigenvalue of $L_{u,\lambda}$, that is, the one corresponding to the following Dirichlet eigenvalue problem:

$$-\Delta\phi - \frac{2\lambda f(x)}{(1-u)^3}\phi = \mu\phi, \quad x \in \Omega; \qquad \phi = 0, \qquad x \in \partial\Omega. \tag{4.1}$$

In other words,

$$\mu_1(\lambda,u)=\inf_{\phi\in H_0^1(\Omega)}\frac{\int_\Omega\left\{|\nabla\phi|^2-2\lambda f(1-u)^{-3}\phi^2\right\}dx}{\int_\Omega\phi^2dx}.$$

4.1. Spectral properties of minimal solutions. We start with the following crucial lemma, which shows among other things that semistable solutions are necessarily minimal solutions.

LEMMA 4.1. *Let f be a function satisfying* (1.2) *on a bounded domain Ω in R^N, and let $\lambda^*:=\lambda^*(\Omega,f)$. Suppose that u is a positive solution of $(S)_\lambda$, and consider any (classical) supersolution v of $(S)_\lambda$, that is,*

$$-\Delta v\ge\frac{\lambda f(x)}{(1-v)^2}\quad\text{and}\quad 0\le v(x)<1,\quad x\in\Omega;\quad v=0,\ \ x\in\partial\Omega. \tag{4.2}$$

If $\mu_1(\lambda,u)>0$, then $v\ge u$ on Ω, and if $\mu_1(\lambda,u)=0$, then $v\equiv u$ on Ω.

Proof. For a given λ and $x\in\Omega$, use the fact that $f(x)\ge 0$ and that $t\to\frac{\lambda f(x)}{(1-t)^2}$ is convex on $(0,1)$ to obtain

$$-\Delta(u+\tau(v-u))-\frac{\lambda f(x)}{[1-(u+\tau(v-u))]^2}\ge 0,\quad x\in\Omega, \tag{4.3}$$

for $\tau\in[0,1]$. Note that (4.3) is an identity at $\tau=0$, which means that the first derivative of the left-hand side for (4.3) with respect to τ is nonnegative at $\tau=0$; i.e.,

$$-\Delta(v-u)-\frac{2\lambda f(x)}{(1-u)^3}(v-u)\ge 0,\quad x\in\Omega;\quad v-u=0,\quad x\in\partial\Omega. \tag{4.4}$$

Thus, the maximal principle implies that if $\mu_1(\lambda,u)>0$, we have $v\ge u$ on Ω, while if $\mu_1(\lambda,u)=0$, we have

$$-\Delta(v-u)-\frac{2\lambda f(x)}{(1-u)^3}(v-u)=0,\quad x\in\Omega. \tag{4.5}$$

The second derivative of the left-hand side for (4.3) with respect to τ is therefore nonnegative at $\tau=0$; i.e.,

$$-\frac{6\lambda f(x)}{(1-u)^4}(v-u)^2\ge 0,\quad x\in\Omega. \tag{4.6}$$

From (4.6) we deduce that $v\equiv u$ in $\Omega\setminus\Omega_0$, where $\Omega_0=\{x\in\Omega:\ f(x)=0\text{ for }x\in\Omega\}$. On the other hand, (4.5) reduces to $-\Delta(v-u)=0$ for $x\in\Omega_0$, while $v-u=0$ on $\partial\Omega_0$, which implies $v\equiv u$ on Ω_0. Hence if $\mu_1(\lambda,u)=0$, then $v\equiv u$ on Ω. □

THEOREM 4.2. *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω in R^N, and consider the branch $\lambda\to u_\lambda$ of minimal solutions on $(0,\lambda^*)$. Then the following hold:*

1. *For each $x\in\Omega$, the function $\lambda\to u_\lambda(x)$ is differentiable and strictly increasing on $(0,\lambda^*)$.*
2. *For each $\lambda\in(0,\lambda^*)$, u_λ is a stable solution and $\lambda\to\mu_{1,\lambda}:=\mu_1(\lambda,u_\lambda)$ is decreasing on $(0,\lambda^*)$.*

Proof. Consider $\lambda_1<\lambda_2<\lambda^*$, their corresponding minimal positive solutions u_{λ_1} and u_{λ_2}, and let u^* be a positive solution for $(S)_{\lambda_2}$. For the monotone increasing series $\{u_n(\lambda_1;x)\}$ defined in (2.9), we then have $u^*>u_0(\lambda_1;x)\equiv 0$, and if $u_{n-1}(\lambda_1;x)\le u^*$

in Ω, then

$$-\Delta(u^*-u_n) = f(x)\left[\frac{\lambda_2}{(1-u^*)^2}-\frac{\lambda_1}{(1-u_{n-1})^2}\right] \geq 0, \quad x \in \Omega\,; \quad u^*-u_n = 0, \quad x \in \partial\Omega.$$

So we have $u_n(\lambda_1;x) \leq u^*$ in Ω. Therefore, $u_{\lambda_1} = \lim_{n\to\infty} u_n(\lambda_1;x) \leq u^*$ in Ω, and in particular $u_{\lambda_1} \leq u_{\lambda_2}$ in Ω. Therefore, $\frac{du_\lambda(x)}{d\lambda} \geq 0$ for all $x \in \Omega$.

That $\lambda \to \mu_{1,\lambda}$ is decreasing follows easily from the variational characterization of $\mu_{1,\lambda}$, the monotonicity of $\lambda \to u_\lambda$, as well as the monotonicity of $(1-u)^{-3}$ with respect to u.

Now we define $\lambda^{**} = \sup\{\lambda;\ u_\lambda \text{ is a stable solution for } (S)_\lambda\}$. It is clear that $\lambda^{**} \leq \lambda^*$, and to show equality it suffices to prove that there is no minimal solution for $(S)_\mu$ with $\mu > \lambda^{**}$. For that, suppose w is a minimal solution of $(S)_{\lambda^{**}+\delta}$ with $\delta > 0$; then we would have, for $\lambda \leq \lambda^{**}$,

$$-\Delta w = \frac{(\lambda^{**}+\delta)f(x)}{(1-w)^2} \geq \frac{\lambda f(x)}{(1-w)^2}, \quad x \in \Omega.$$

Since for $0 < \lambda < \lambda^{**}$ the minimal solutions u_λ are stable, it follows from Lemma 4.1 that $1 > w \geq u_\lambda$ for all $0 < \lambda < \lambda^{**}$. Consequently, $\underline{u} = \lim_{\lambda \nearrow \lambda^{**}} u_\lambda$ exists in $C^1(\Omega)$ and is a solution for $(S)_{\lambda^{**}}$. Now from the definition of λ^{**}, we necessarily have $\mu_{1,\lambda^{**}} = 0$; hence by again applying Lemma 4.1, we obtain that $w \equiv \underline{u}$ and $\delta = 0$ on Ω, which is a contradiction, and hence $\lambda^{**} = \lambda^*$.

Since each u_λ is stable, then by setting $F(\lambda, u_\lambda) := -\Delta - \frac{\lambda f}{(1-u_\lambda)^2}$, we get that $F_{u_\lambda}(\lambda, u_\lambda)$ is invertible for $0 < \lambda < \lambda^*$. It then follows from the implicit function theorem that $u_\lambda(x)$ is differentiable with respect to λ.

Finally, by differentiating $(S)_\lambda$ with respect to λ, and since $\lambda \to u_\lambda(x)$ is non-decreasing, we get

$$-\Delta\frac{du_\lambda}{d\lambda} - \frac{2\lambda f(x)}{(1-u_\lambda)^3}\frac{du_\lambda}{d\lambda} = \frac{f(x)}{(1-u_\lambda)^2} \geq 0, \quad x \in \Omega\,; \quad \frac{du_\lambda}{d\lambda} \geq 0, \quad x \in \partial\Omega.$$

Applying the strong maximum principle, we conclude that $\frac{du_\lambda}{d\lambda} > 0$ on Ω for all $0 < \lambda < \lambda^*$. □

Remark 4.1. Lemma 3 of [8] yields $\mu_1(1,0)$ as an upper bound for λ^{**}—at least in the case where $\inf_\Omega f > 0$ on Ω. Since $\lambda^{**} = \lambda^*$, this gives another upper bound for λ^* in our setting. Note, however, that in the case where $f \equiv 1$, we have $\mu_1(1,0) = \frac{\mu_\Omega}{2}$, while the estimate in Theorem 1.1 gives $\frac{4\mu_\Omega}{27}$ for an upper bound.

4.2. Energy estimates and regularity. We start with the following easy observations.

LEMMA 4.3. *Let f be a function satisfying* (1.2) *on a bounded domain Ω in R^N.*

1. *Any (weak) solution u in $H_0^1(\Omega)$ of $(S)_\lambda$ then satisfies $\int_\Omega \frac{f}{(1-u)^2}dx < \infty$.*
2. *If $\inf_\Omega f > 0$ and $N \geq 3$, then any solution u such that $f/(1-u) \in L^{3N/2}$ is a classical solution.*

Proof. (1) Since $u \in H_0^1(\Omega)$ is a positive solution of $(S)_\lambda$, we have

$$\int_\Omega \frac{\lambda f}{(1-u)^2} - \int_\Omega \frac{\lambda f}{1-u} = \int_\Omega \frac{\lambda u f}{(1-u)^2} = \int_\Omega |\nabla u|^2 =: C < +\infty,$$

which implies that

$$\int_\Omega \frac{\lambda f}{(1-u)^2} \le C + \int_\Omega \frac{\lambda f}{1-u} \le C + \int_\Omega \left[C\varepsilon \frac{\lambda f}{(1-u)^2} + \frac{C}{\varepsilon} f \right] \le C + C\varepsilon \int_\Omega \frac{\lambda f}{(1-u)^2}$$

with $\varepsilon > 0$. Therefore, by choosing $\varepsilon > 0$ small enough, we conclude that $\int_\Omega \frac{f}{(1-u)^2} < \infty$.

(2) Suppose u is a weak solution such that $\frac{f(x)}{(1-u)^3} \in L^p(\Omega)$, which means that $\frac{f(x)}{(1-u)^2} \in L^{3p/2}(\Omega)$. By Sobolev's theorem we can already deduce that $u \in C^{0,\alpha}$ with $\alpha = 2 - \frac{2N}{3p}$. To get more regularity, it suffices to show that $u < 1$ on Ω, but then if not, we consider $x_0 \in \bar{\Omega}$ such that $u(x_0) = \|u\|_{C(\bar{\Omega})} = 1$; then we have $|1 - u(x)| = |u(x_0) - u(x)| \le C|x_0 - x|^\alpha$ on $\bar{\Omega}$. This inequality shows that if $p \ge \frac{N}{2}$, then we have

$$\infty > \int_\Omega \left(\frac{f(x)}{(1-u)^3} \right)^p dx \ge C' \int_\Omega |x - x_0|^{-3p\alpha} dx = C' \int_\Omega |x - x_0|^{-N} dx = \infty,$$

a contradiction, which implies that we must have $\|u\|_{C(\bar{\Omega})} < 1$. □

Note that the above argument cannot be applied to the case where $f(x) \ge 0$ vanishes on Ω, and therefore we have to use the iterative scheme outlined in the next theorem.

THEOREM 4.4. *Let f be a function satisfying (1.2) on a bounded domain Ω in R^N. Then for any constant $C > 0$ there exists $0 < K(C, N) < 1$ such that a positive weak solution u of $(S)_\lambda$ $(0 < \lambda < \lambda^*)$ is a classical solution and $\| u \|_{C(\Omega)} \le K(C, N)$, provided that*

1. *$N = 1$ and $\| \frac{f}{(1-u)^3} \|_{L^1(\Omega)} \le C$,*
2. *$N \ge 2$ and $\| \frac{f}{(1-u)^3} \|_{L^{N/2}(\Omega)} \le C$.*

Proof. We prove this lemma by considering the following three cases separately.

(1) If $N = 1$, then for any $I > 0$ we write using the Sobolev inequality with constant $K(1) > 0$,

$$\begin{aligned}
&K(1) \parallel (1-u)^{-1} - 1 \parallel_{L^\infty}^2 \\
&\quad \le \int_\Omega \left| \nabla [(1-u)^{-1} - 1] \right|^2 \\
&\quad = \frac{\lambda}{3} \int_\Omega f(1-u)^{-2} [(1-u)^{-3} - 1] \\
&\quad \le CI + C \int_{\{(1-u)^{-3} \ge I\}} 8f(1-u)^{-2} \\
&\qquad + C \int_{\{(1-u)^{-3} \ge I\}} f \left[\frac{1}{(1-u)^3} + \frac{2}{(1-u)^2} + \frac{4}{1-u} \right] \left[(1-u)^{-1} - 1 \right]^2 \\
&\quad \le CI + C + C \parallel (1-u)^{-1} - 1 \parallel_{L^\infty(\{(1-u)^{-3} \ge I\})}^2 \int_{\{(1-u)^{-3} \ge I\}} \frac{f}{(1-u)^3} \\
&\quad \le CI + C + C\varepsilon(I) \parallel (1-u)^{-1} - 1 \parallel_{L^\infty}^2,
\end{aligned} \tag{4.7}$$

with $\varepsilon(I) = \int_{\{(1-u)^{-3}\geq I\}} \frac{f}{(1-u)^3}$, where Lemma 4.3(1) is applied in the second inequality. From the assumption $f/(1-u)^3 \in L^1(\Omega)$, we have $\varepsilon(I) \to 0$ as $I \to \infty$. We now choose I such that $\varepsilon(I) \leq \frac{K(1)}{2C}$, so that the above estimates implier that $\| (1-u)^{-1} - 1 \|_{L^\infty} < C$. Standard regularity theory for elliptic problems now implies that $1/(1-u) \in C^{2,\alpha}(\Omega)$. Therefore, u is classical, and there exists a constant $K(C,N)$ which can be taken strictly less than 1 such that $\| u \|_{C(\Omega)} \leq K(C,N) < 1$.

(2) Assuming $N = 2$, we need to show that

$$(1-u)^{-1} \in L^p(\Omega) \quad \text{for any } p > 1. \tag{4.8}$$

Fix $p > 1$ and let us introduce $T_k u = \min\{u, 1-k\}$, the truncated function of u at level $1-k$, $0 < k < 1$. For k small, we take $(1-T_k u)^{-1} - 1 \in H_0^1(\Omega)$ as a test function for $(S)_\lambda$:

$$\int_\Omega \frac{|\nabla T_k u|^2}{(1-T_k u)^2} = \int_\Omega \frac{\lambda f(x)}{(1-u)^2}((1-T_k u)^{-1} - 1) \leq \int_\Omega \frac{\lambda f(x)}{|1-u|^3} \leq C < +\infty. \tag{4.9}$$

Since now $\log\left(\frac{1}{1-T_k u}\right) \in H_0^1(\Omega)$, by the Moser–Trudinger inequality and (4.9) we get that for any $p > 1$

$$\int_\Omega (1-T_k u)^{-p} \leq C_1 exp\left(\frac{p^2}{16\pi}\int_\Omega \left|\nabla \log\left(\frac{1}{1-T_k u}\right)\right|^2\right) \leq C_2,$$

where C_1, C_2 denote positive constants depending only on p and C. Taking the limit as $k \to 0$ and using that $u \leq 1$, we get the validity of (4.8).

(3) The case when $N > 2$ is more elaborate, and we first show that $(1-u)^{-1} \in L^q(\Omega)$ for all $q \in (1,\infty)$. Since $u \in H_0^1(\Omega)$ is a solution of $(S)_\lambda$, we already have $\int_\Omega \frac{f}{(1-u)^2} < C$. Now we proceed by iteration to show that if $\int_\Omega \frac{f}{(1-u)^{2+2\theta}} < C$ for some $\theta \geq 0$, then $\int_\Omega \frac{1}{(1-u)^{2^*(1+\theta)}} < C$.

Indeed, for any constant $\theta \geq 0$ and $\ell > 0$ we choose a test function $\phi = [(1-u)^{-3} - 1]\min\{(1-u)^{-2\theta}, \ell^2\}$. By applying this test function to both sides of $(S)_\lambda$, we have

$$\begin{aligned}
&\lambda \int_\Omega f(1-u)^{-2}[(1-u)^{-3} - 1]\min\{(1-u)^{-2\theta}, \ell^2\} \\
&= \int_\Omega \nabla u \cdot \nabla\big[\big((1-u)^{-3} - 1\big)\min\{(1-u)^{-2\theta}, \ell^2\}\big] \\
&= 3\int_\Omega |\nabla u|^2 (1-u)^{-4}\min\{(1-u)^{-2\theta}, \ell^2\} \\
&\quad + 2\theta \int_{\{(1-u)^{-\theta} \leq \ell\}} |\nabla u|^2 (1-u)^{-2\theta-1}[(1-u)^{-3} - 1].
\end{aligned} \tag{4.10}$$

We now suppose $\int_\Omega \frac{f}{(1-u)^{2+2\theta}} < C$. We then obtain from (4.10) and the fact that

$\frac{1}{(1-u)^5} \le C_I \frac{1}{(1-u)^3}(\frac{1}{1-u}-1)^2$ whenever $(1-u)^{-3} \ge I > 1$ that

$$
\begin{aligned}
&\int_\Omega \left|\nabla[\big((1-u)^{-1}-1\big)\min\{(1-u)^{-\theta},\ell\}]\right|^2 \\
&= 2\int_\Omega |\nabla u|^2(1-u)^{-4}\min\{(1-u)^{-2\theta},\ell^2\} \\
&\quad + 2\theta^2\int_{\{(1-u)^{-\theta}\le\ell\}} |\nabla u|^2(1-u)^{-2\theta-1}\left[\frac{1}{(1-u)^3}+\frac{1}{1-u}-\frac{2}{(1-u)^2}\right] \\
&\le C\lambda\int_\Omega f(1-u)^{-2}[(1-u)^{-3}-1]\min\{(1-u)^{-2\theta},\ell^2\} \\
&\le CI + C\int_{\{(1-u)^{-3}\ge I\}} f(1-u)^{-3}\big[(1-u)^{-1}-1\big]^2\min\{(1-u)^{-2\theta},\ell^2\} \\
&\le CI + C\left[\int_{\{(1-u)^{-3}\ge I\}}\left(\frac{f}{(1-u)^3}\right)^{\frac{N}{2}}\right]^{\frac{2}{N}} \\
&\qquad \times\left[\int_{\{(1-u)^{-3}\ge I\}}\big(\big[(1-u)^{-1}-1\big]\min\{(1-u)^{-\theta},\ell\}\big)^{\frac{2N}{N-2}}\right]^{\frac{N-2}{N}} \\
&\le CI + C\varepsilon(I)\int_\Omega \left|\nabla[\big((1-u)^{-1}-1\big)\min\{(1-u)^{-\theta},\ell\}]\right|^2
\end{aligned}
\tag{4.11}
$$

with $\varepsilon(I) = \big[\int_{\{(1-u)^{-3}\ge I\}}\big(\frac{f}{(1-u)^3}\big)^{\frac{N}{2}}\big]^{\frac{2}{N}}$.

From the assumption $f/(1-u)^3 \in L^{\frac{N}{2}}(\Omega)$ we have $\varepsilon(I)\to 0$ as $I\to\infty$. We now choose I such that $\varepsilon(I) = \frac{1}{2C}$, and the above estimates imply that

$$
\int_{\{(1-u)^{-\theta}\le\ell\}} \left|\nabla[(1-u)^{-\theta-1}-(1-u)^{-\theta}]\right|^2 \le CI,
$$

where the bound is uniform with respect to ℓ. This estimate leads to

$$
\begin{aligned}
&\frac{1}{(\theta+1)^2}\int_{\{(1-u)^{-\theta}\le\ell\}} \left|\nabla[(1-u)^{-\theta-1}]\right|^2 = \int_{\{(1-u)^{-\theta}\le\ell\}} (1-u)^{-2\theta-4}\left|\nabla u\right|^2 \\
&\le CI + C\int_{\{(1-u)^{-\theta}\le\ell\}} (1-u)^{-2\theta-3}\left|\nabla u\right|^2 \\
&\le CI + \int_{\{(1-u)^{-\theta}\le\ell\}} \big[C\varepsilon(1-u)^{-2\theta-4} + C/\varepsilon\big]\left|\nabla u\right|^2 \\
&\le CI + C/\varepsilon + C\varepsilon\int_{\{(1-u)^{-\theta}\le\ell\}} (1-u)^{-2\theta-4}\left|\nabla u\right|^2
\end{aligned}
$$

with $\varepsilon > 0$. This means that for $\varepsilon > 0$ sufficiently small

$$
\int_{\{(1-u)^{-\theta}\le\ell\}} \left|\nabla(1-u)^{-\theta-1}\right|^2 = \int_{\{(1-u)^{-\theta}\le\ell\}} (\theta+1)^2(1-u)^{-2\theta-4}\left|\nabla u\right|^2 < C.
$$

Thus we can let $\ell\to\infty$, and we get that $(1-u)^{-\theta-1} \in H^1(\Omega) \hookrightarrow L^{2^*}(\Omega)$, which means that $\int_\Omega \frac{1}{(1-u)^{2^*(1+\theta)}} < C$.

By iterating the above argument for $\theta_i+1=\frac{N}{N-2}(\theta_{i-1}+1)$ for $i\geq 1$ and starting with $\theta_0=0$, we find that $1/(1-u)\in L^q(\Omega)$ for all $q\in(1,\infty)$.

Standard regularity theory for elliptic problems applies again to give that $1/(1-u)\in C^{2,\alpha}(\Omega)$. Therefore, u is a classical solution, and there exists a constant $0<K(C,N)<1$ such that $\|u\|_{C(\Omega)}\leq K(C,N)<1$. □

THEOREM 4.5. *For any dimension $1\leq N\leq 7$ there exists a constant $0<C(N)<1$ independent of λ such that for any $0<\lambda<\lambda^*$ the minimal solution u_λ satisfies $\|u_\lambda\|_{C(\Omega)}\leq C(N)$.*

Consequently, $u^*=\lim_{\lambda\uparrow\lambda^*}u_\lambda$ exists in the topology of $C^{2,\alpha}(\bar\Omega)$ with $0<\alpha<1$. It is the unique classical solution for $(S)_{\lambda^*}$ and satisfies $\mu_{1,\lambda^*}(u^*)=0$.

This result—which yields Theorem 1.3(2)—will follow from the following uniform energy estimate on the minimal solutions u_λ.

PROPOSITION 4.6. *There exists a constant $C(p)>0$ such that for $\lambda\in(0,\lambda^*)$ the minimal solution u_λ satisfies $\|\frac{f}{(1-u_\lambda)^3}\|_{L^p(\Omega)}\leq C(p)$ as long as $p<1+\frac{4}{3}+2\sqrt{\frac{2}{3}}$.*

Proof. Since minimal solutions are stable, we have

$$\lambda\int_\Omega\frac{2f(x)}{(1-u_\lambda)^3}w^2dx\leq-\int_\Omega w\Delta w dx=\int_\Omega|\nabla w|^2dx \tag{4.12}$$

for all $0<\lambda<\lambda^*$ and nonnegative $w\in H_0^1(\bar\Omega)$. Setting

$$w=(1-u_\lambda)^i-1>0,\quad\text{where } -2-\sqrt{6}<i<0, \tag{4.13}$$

then (4.12) becomes

$$i^2\int_\Omega(1-u_\lambda)^{2i-2}|\nabla u_\lambda|^2dx\geq\lambda\int_\Omega\frac{2[1-(1-u_\lambda)^i]^2f(x)}{(1-u_\lambda)^3}dx. \tag{4.14}$$

On the other hand, multiplying $(S)_\lambda$ by $\frac{i^2}{1-2i}[(1-u_\lambda)^{2i-1}-1]$ and applying integration by parts yields that

$$i^2\int_\Omega(1-u_\lambda)^{2i-2}|\nabla u_\lambda|^2dx=\lambda\frac{i^2}{2i-1}\int_\Omega\frac{[1-(1-u_\lambda)^{2i-1}]f(x)}{(1-u_\lambda)^2}dx. \tag{4.15}$$

Hence (4.14) and (4.15) reduce to

$$\begin{aligned}&\frac{\lambda\,i^2}{2i-1}\int_\Omega\frac{f(x)}{(1-u_\lambda)^2}dx-2\lambda\int_\Omega\frac{f(x)}{(1-u_\lambda)^3}dx+4\lambda\int_\Omega\frac{f(x)}{(1-u_\lambda)^{3-i}}dx\\&\quad\geq\lambda\left(2+\frac{i^2}{2i-1}\right)\int_\Omega\frac{f(x)}{(1-u_\lambda)^{3-2i}}dx.\end{aligned} \tag{4.16}$$

From the choice of i in (4.13) we have $2+\frac{i^2}{2i-1}>0$. So (4.16) and Holder's inequality imply that

$$\begin{aligned}\int_\Omega\frac{f(x)}{(1-u_\lambda)^{3-2i}}dx&\leq C\int_\Omega\frac{f(x)}{(1-u_\lambda)^{3-i}}dx\\&\leq C\left(\int_\Omega\left|\frac{f^{\frac{3-i}{3-2i}}}{(1-u_\lambda)^{3-i}}\right|^{\frac{3-2i}{3-i}}dx\right)^{\frac{3-i}{3-2i}}\cdot\left(\int_\Omega\left|f^{\frac{-i}{3-2i}}\right|^{\frac{3-2i}{-i}}dx\right)^{\frac{-i}{3-2i}}\\&\leq C\left(\int_\Omega\frac{f(x)}{(1-u_\lambda)^{3-2i}}dx\right)^{\frac{3-i}{3-2i}}.\end{aligned} \tag{4.17}$$

It follows that $\int_\Omega \frac{f(x)}{(1-u_\lambda)^{3-2i}}dx \le C$, and therefore we have

$$\int_\Omega \left|\frac{f(x)}{(1-u_\lambda)^3}\right|^{\frac{3-2i}{3}} dx = \int_\Omega f^{\frac{-2i}{3}} \cdot \frac{f}{(1-u_\lambda)^{3-2i}}dx \le C\int_\Omega \frac{f}{(1-u_\lambda)^{3-2i}}dx \le C. \tag{4.18}$$

In view of (4.13) we finally get that $\| \frac{f(x)}{(1-u_\lambda)^3} \|_{L^p} \le C$, where $p = \frac{3-2i}{3} \le 1 + \frac{4}{3} + 2\sqrt{\frac{2}{3}}$. □

Proof of Theorem 4.5. The existence of u^* as a classical solution follows from Proposition 4.6 and Theorem 4.4, as long as $\frac{N}{2} < 1 + \frac{4}{3} + 2\sqrt{\frac{2}{3}}$, which happens when $N \le 7$.

Since $\mu_{1,\lambda} > 0$ on the minimal branch for any $\lambda < \lambda^*$, we have the limit $\mu_{1,\lambda^*} \ge 0$. If now $\mu_{1,\lambda^*} > 0$, the implicit function theorem could be applied to the operator $L_{u_{\lambda^*},\lambda^*}$ and would allow the continuation of the minimal branch $\lambda \mapsto u_\lambda$ of classical solutions beyond λ^*, which is a contradiction, and hence $\mu_{1,\lambda^*} = 0$. The uniqueness in the class of classical solutions then follows from Lemma 4.1.

5. Uniqueness and multiplicity of solutions. We first note that in view of the monotonicity in λ and the uniform boundedness of the first branch of solutions, the extremal function defined by $u^*(x) = \lim_{\lambda\uparrow\lambda^*} u_\lambda(x)$ always exists and can always be considered as a solution for $(S)_{\lambda^*}$ in a generalized sense. Now if there exists $0 < C < 1$ such that $\| u_\lambda \|_{C(\Omega)} \le C$ for each $\lambda < \lambda^*$—just like in the case where $1 \le N < 8$—then we have seen in Theorem 4.5 that u^* is unique among the classical solutions. In the sequel, we tackle the important case when u^* is a weak solution (i.e., in $H_0^1(\Omega)$) of $(S)_{\lambda^*}$ but with the possibility that $\|u^*\|_\infty = 1$.

5.1. Uniqueness of the solution at $\lambda = \lambda^*$. We shall borrow ideas from [3, 5, 16], where the authors deal with the case of regular nonlinearities. However, unlike those papers, where solutions are considered in a very weak sense, we consider here a more focused and much simpler situation. We establish the following useful characterization of the extremal solution.

THEOREM 5.1. *Assume that f is a function satisfying* (1.2) *on a bounded domain Ω in R^N. For $\lambda > 0$, consider $u \in H_0^1(\Omega)$ to be a weak solution of $(S)_\lambda$ (in the $H_0^1(\Omega)$ sense) such that $\| u \|_{L^\infty(\Omega)} = 1$. Then the following assertions are equivalent:*

1. *$\mu_{1,\lambda} \ge 0$; that is, u satisfies $\int_\Omega |\nabla\phi|^2 \ge \int_\Omega \frac{2\lambda f(x)}{(1-u)^3}\phi^2$ for all $\phi \in H_0^1(\Omega)$.*
2. *$\lambda = \lambda^*$ and $u = u^*$.*

We need the following uniqueness result.

PROPOSITION 5.2. *Let $f \in C(\bar\Omega)$ be a nonnegative function. Let u_1, u_2 be two $H_0^1(\Omega)$-weak solutions of $(S)_\lambda$ so that $\mu_{1,\lambda}(u_i) \ge 0$ for $i = 1, 2$. Then $u_1 = u_2$ almost everywhere (a.e.) in Ω.*

Proof. For any $\theta \in [0, 1]$ and $\phi \in H_0^1(\Omega)$, $\phi \ge 0$, we have that

$$\begin{aligned} I_{\theta,\phi} :&= \int_\Omega \nabla\big(\theta u_1 + (1-\theta)u_2\big)\nabla\phi - \int_\Omega \frac{\lambda f(x)}{\big(1-\theta u_1 - (1-\theta)u_2\big)^2}\phi \\ &= \lambda\int_\Omega f(x)\left(\frac{\theta}{(1-u_1)^2} + \frac{1-\theta}{(1-u_2)^2} - \frac{1}{\big(1-\theta u_1 - (1-\theta)u_2\big)^2}\right)\phi \ge 0 \end{aligned}$$

due to the convexity of $u \to 1/(1-u)^2$. Since $I_{0,\phi} = I_{1,\phi} = 0$, the derivative of $I_{\theta,\phi}$ at $\theta = 0, 1$ provides

$$\int_\Omega \nabla\big(u_1 - u_2\big)\nabla\phi - \int_\Omega \frac{2\lambda f(x)}{\big(1-u_2\big)^3}(u_1-u_2)\phi \geq 0,$$

$$\int_\Omega \nabla\big(u_1 - u_2\big)\nabla\phi - \int_\Omega \frac{2\lambda f(x)}{\big(1-u_1\big)^3}(u_1-u_2)\phi \leq 0,$$

for any $\phi \in H_0^1(\Omega)$, $\phi \geq 0$. Testing the first inequality on $\phi = (u_1-u_2)^-$ and the second one on $(u_1-u_2)^+$, we get that

$$\int_\Omega \left[|\nabla(u_1-u_2)^-|^2 - \frac{2\lambda f(x)}{(1-u_2)^3}\big((u_1-u_2)^-\big)^2\right] \leq 0,$$

$$\int_\Omega \left[|\nabla(u_1-u_2)^+|^2 - \frac{2\lambda f(x)}{(1-u_1)^3}\big((u_1-u_2)^+\big)^2\right] \leq 0.$$

Since $\mu_{1,\lambda}(u_1) \geq 0$, we have

(1) either $\mu_{1,\lambda}(u_1) > 0$ and then $u_1 \leq u_2$ a.e.,

(2) or $\mu_{1,\lambda}(u_1) = 0$, which then gives $\int_\Omega \nabla(u_1-u_2)\nabla\bar\phi - \int_\Omega \frac{2\lambda f(x)}{(1-u_1)^3}(u_1-u_2)\bar\phi = 0$, where $\bar\phi = (u_1-u_2)^+$. Since $I_{\theta,\bar\phi} \geq 0$ for any $\theta \in [0,1]$ and $I_{1,\bar\phi} = \partial_\theta I_{1,\bar\phi} = 0$, we get that $\partial^2_{\theta\theta} I_{1,\bar\phi} = -\int_\Omega \frac{6\lambda f(x)}{(1-u_1)^4}((u_1-u_2)^+)^3 \geq 0$. Letting $Z_0 = \{x \in \Omega : f(x) = 0\}$, we clearly have $(u_1-u_2)^+ = 0$ a.e. in $\Omega \setminus Z_0$, and from the above we obtain $\int_\Omega |\nabla(u_1-u_2)^+|^2 = 0$. Hence, $u_1 \leq u_2$ a.e. in Ω. The same argument applies to prove the reversed inequality, and the proof is complete. □

Since $\| u_\lambda \| < 1$ for any $\lambda \in (0, \lambda^*)$, we need—in order to prove Theorem 5.1—to show only that $(S)_\lambda$ does not have any $H_0^1(\Omega)$-weak solution for $\lambda > \lambda^*$. By the definition of λ^*, this is already true for classical solutions. We shall now extend this property to the class of weak solutions by means of the following result.

PROPOSITION 5.3. *If w is a $H_0^1(\Omega)$-weak solution of $(S)_\lambda$, then for any $\varepsilon \in (0,1)$ there exists a classical solution w_ε of $(S)_{\lambda(1-\varepsilon)}$.*

Proof. First we prove that for any $\psi \in C^2([0,1])$ concave function so that $\psi(0) = 0$ we have that

$$\int_\Omega \nabla\psi(w)\nabla\varphi \geq \int_\Omega \frac{\lambda f}{(1-w)^2}\dot\psi(w)\varphi \tag{5.1}$$

for any $\varphi \in H_0^1(\Omega)$, $\varphi \geq 0$. Indeed, by the concavity of ψ we get

$$\begin{aligned}\int_\Omega \nabla\psi(w)\nabla\varphi = \int_\Omega \dot\psi(w)\nabla w\nabla\varphi &= \int_\Omega \nabla w\nabla\left(\dot\psi(w)\varphi\right) - \int_\Omega \ddot\psi(w)\varphi|\nabla w|^2 \\ &\geq \int_\Omega \frac{\lambda f(x)}{(1-w)^2}\dot\psi(w)\varphi\end{aligned}$$

for any $\varphi \in C_0^\infty(\Omega)$, $\varphi \geq 0$, and by density we get (5.1) for any $\varphi \in H_0^1(\Omega)$, $\varphi \geq 0$.

Now let $\varepsilon \in (0,1)$, and define $\psi_\varepsilon(w) := 1 - \big(\varepsilon + (1-\varepsilon)(1-w)^3\big)^{\frac{1}{3}}$ for $0 \leq w \leq 1$. Since $\psi_\varepsilon \in C^2([0,1])$ is a concave function, $\psi_\varepsilon(0) = 0$, and since $\dot\psi_\varepsilon(w) = (1-\varepsilon)$

$\frac{g(\psi_\varepsilon(w))}{g(w)}$, where $g(s) := (1-s)^{-2}$, we obtain from (5.1) that for any $\varphi \in H_0^1(\Omega)$, $\varphi \geq 0$,

$$\int_\Omega \nabla\psi_\varepsilon(w)\nabla\varphi \geq \int_\Omega \frac{\lambda f(x)}{(1-w)^2}\dot{\psi}_\varepsilon(w)\varphi = \lambda(1-\varepsilon)\int_\Omega f(x)g\big(\psi_\varepsilon(w)\big)\varphi$$
$$= \int_\Omega \frac{\lambda(1-\varepsilon)f(x)}{(1-\psi_\varepsilon(w))^2}\varphi.$$

Hence, $\psi_\varepsilon(w)$ is a $H_0^1(\Omega)$-weak supersolution of $(S)_{\lambda(1-\varepsilon)}$ so that $0 \leq \psi_\varepsilon(w) \leq 1-\varepsilon^{\frac{1}{3}} < 1$. Since 0 is a subsolution for any $\lambda > 0$, we get the existence of a $H_0^1(\Omega)$-weak solution w_ε of $(S)_{\lambda(1-\varepsilon)}$ so that $0 \leq w_\varepsilon \leq 1-\varepsilon^{\frac{1}{3}}$. By standard elliptic regularity theory, w_ε is a classical solution of $(S)_{\lambda(1-\varepsilon)}$. □

5.2. Uniqueness of low energy solutions for small voltage. We now focus on the uniqueness when λ is small. We first define nonminimal solutions for $(S)_\lambda$ as follows.

DEFINITION 5.4. *A solution $0 \leq u < 1$ is said to be a nonminimal positive solution of $(S)_\lambda$ if there exist another positive solution v of $(S)_\lambda$ and a point $x \in \Omega$ such that $u(x) > v(x)$.*

LEMMA 5.5. *Suppose u is a nonminimal solution of $(S)_\lambda$ with $\lambda \in (0, \lambda^*)$. Then $\mu_1(\lambda, u) < 0$, and the function $w = u - u_\lambda$ is in the negative space of $L_{u,\lambda} = -\Delta - \frac{2\lambda f(x)}{(1-u)^3}$.*

Proof. For a fixed $\lambda \in (0, \lambda^*)$, let u_λ be the minimal solution of $(S)_\lambda$. We have $w = u - u_\lambda \geq 0$ in Ω, and $-\Delta w - \frac{\lambda(2-u-u_\lambda)f}{(1-u)^2(1-u_\lambda)^2}w = 0$ in Ω. Hence the strong maximum principle yields that $u_\lambda < u$ in Ω.

Let $\Omega_0 = \{x \in \Omega : f(x) = 0\}$ and $\Omega/\Omega_0 = \{x \in \Omega : f(x) > 0\}$. Direct calculations give that

$$-\Delta(u-u_\lambda) - \frac{2\lambda f(u-u_\lambda)}{(1-u)^3} = \lambda f\left[\frac{1}{(1-u)^2} - \frac{1}{(1-u_\lambda)^2} - \frac{2(u-u_\lambda)}{(1-u)^3}\right]$$
$$= 0, \ \ x \in \Omega_0\,; \quad < 0, \ x \in \Omega/\Omega_0. \tag{5.2}$$

From this we get

$$\langle L_{u,\lambda}w, w\rangle = \lambda\int_{\Omega/\Omega_0} f(x)(u-u_\lambda)$$
$$\cdot\left[\frac{1}{(1-u)^2} - \frac{1}{(1-u_\lambda)^2} - \frac{2}{(1-u)^3}(u-u_\lambda)\right]dx < 0. \tag{5.3}$$

Now we are able to prove the following uniqueness result.

THEOREM 5.6. *For every $M > 0$ there exists $0 < \lambda_1^*(M) < \lambda^*$ such that for $\lambda \in (0, \lambda_1^*(M))$ the equation $(S)_\lambda$ has a unique solution v satisfying the following:*

1. $\|\frac{f}{(1-v)^3}\|_1 \leq M$ *and the dimension* $N = 1$,
2. $\|\frac{f}{(1-v)^3}\|_{1+\epsilon} \leq M$ *for some* $\epsilon > 0$ *and* $N = 2$,
3. $\|\frac{f}{(1-v)^3}\|_{N/2} \leq M$ *and* $N > 2$.

Proof. For any fixed $\lambda \in (0, \lambda^*)$, let u_λ be the minimal solution of $(S)_\lambda$, and suppose $(S)_\lambda$ has a nonminimal solution u. The preceding lemma then gives

$$\int_\Omega |\nabla(u-u_\lambda)|^2 dx < \int_\Omega \frac{2\lambda(u-u_\lambda)^2 f(x)}{(1-u)^3}dx.$$

This implies in the case where $N > 2$ that

$$C(N)\left(\int_\Omega |u-u_\lambda|^{\frac{2N}{N-2}}dx\right)^{\frac{N-2}{N}} < \lambda\int_\Omega \frac{2f(x)}{|1-u|^3}|u-u_\lambda|^2dx$$
$$\le 2\lambda\left(\int_\Omega \left|\frac{f}{|1-u|^3}\right|^{\frac{N}{2}}\right)^{\frac{2}{N}}\left(\int_\Omega |u-u_\lambda|^{\frac{2N}{N-2}}\right)^{\frac{N-2}{N}}$$
$$\le 2\lambda M^{\frac{2}{N}}\left(\int_\Omega (u-u_\lambda)^{\frac{2N}{N-2}}dx\right)^{\frac{N-2}{N}},$$

which is a contradiction if $\lambda < \frac{C(N)}{2M^{\frac{2}{N}}}$ unless $u \equiv u_\lambda$. If $N = 1$, then we write

$$C(1)\|(u-u_\lambda)\|_\infty^2 < \lambda\int_\Omega \frac{2f(x)}{(1-u)^3}(u-u_\lambda)^2dx \le 2\lambda\|(u-u_\lambda)\|_\infty^2\int_\Omega \frac{f}{(1-u)^3}dx,$$

and the proof follows. A similar proof holds for dimension $N = 2$. □

Remark 5.1. The above gives uniqueness for small λ among all solutions that either stay away from 1 or those that approach it slowly. We do not know whether, if λ is small enough, any positive solution v of $(S)_\lambda$ satisfies $\int_\Omega (1-v)^{-\frac{3N}{2}}dx \le M$ for some uniform bound M independent of λ. Numerical computations do show that we may have uniqueness for small λ—at least for radially symmetric solutions—as long as $N \ge 2$.

5.3. Second solutions around the bifurcation point. Our next result is quite standard.

LEMMA 5.7. *Suppose there exists* $0 < C < 1$ *such that* $\|u_\lambda\|_{C(\bar\Omega)} \le C$ *for each* $\lambda < \lambda^*$. *Then there exists* $\delta > 0$ *such that the solutions of* $(S)_\lambda$ *near* $(\lambda^*, u_{\lambda^*})$ *form a curve* $\rho(s) = \{(\bar\lambda(s), v(s)) : |s| < \delta\}$, *and the pair* $(\bar\lambda(s), v(s))$ *satisfies*

$$\bar\lambda(0) = \lambda^*,\ \bar\lambda'(0) = 0,\ \bar\lambda''(0) < 0, \quad \textit{and} \quad v(0) = u_{\lambda^*},\ v'(0)(x) > 0 \textit{ in } \Omega. \tag{5.4}$$

In particular, if $1 \le N \le 7$, *then for* λ *close enough to* λ^* *there exists a unique second branch* U_λ *of solutions for* $(S)_\lambda$, *bifurcating from* u^*, *such that* $\mu_{1,\lambda}(U_\lambda) < 0$ *while* $\mu_{2,\lambda}(U_\lambda) > 0$.

Proof. The proof is similar to a related result of Crandall and Rabinowitz (cf. [7, 8]) so we will be brief. First, the assumed upper bound on u_λ in C^1 and standard regularity theory shows that if $f \in C(\bar\Omega)$, then $\|u_\lambda\|_{C^{2,\alpha}(\bar\Omega)} \le C < 1$ for some $0 < \alpha < 1$ (while if $f \in L^\infty$, then $\|u_\lambda\|_{C^{1,\alpha}(\bar\Omega)} \le C < 1$). It follows that $\{(\lambda, u_\lambda)\}$ is precompact in the space $R \times C^{2,\alpha}$, and hence we have a limiting point $(\lambda^*, u_{\lambda^*})$, as desired. Since $\frac{\lambda^* f(x)}{(1-u_{\lambda^*})^2}$ is nonnegative, Theorem 3.2 of [7] characterizes the solution set of $(S)_\lambda$ near $(\lambda^*, u_{\lambda^*})$: $\bar\lambda(0) = \lambda^*$, $\bar\lambda'(0) = 0$, $v(0) = u_{\lambda^*}$, and $v'(0) > 0$ in Ω. The same computation as in Theorem 4.8 in [7] gives that $\bar\lambda''(0) < 0$. In particular, if $1 \le N \le 7$, then our Theorem 4.5 gives the compactness of $u^* = u_{\lambda^*}$, and the theory of Crandall and Rabinowitz in [8] then implies that, for λ close enough to λ^*, there exists a unique second branch U_λ of solutions for $(S)_\lambda$, bifurcating from u^*, such that $\mu_{1,\lambda}(U_\lambda) < 0$ while $\mu_{2,\lambda}(U_\lambda) > 0$. □

Remark 5.2. A version of these results will be established variationally in the companion paper [9]. Indeed, we shall give there a variational characterization for both the stable and unstable solutions u_λ, U_λ in the following sense: For $1 \le N \le 7$, there exists $\delta > 0$ such that for any $\lambda \in (\lambda^* - \delta, \lambda^*)$ the minimal solution u_λ is a local minimum for some regularized energy functional $J_{\varepsilon,\lambda}$ on the space $H_0^1(\Omega)$, while the second solution U_λ is a mountain pass for the functional $J_{\varepsilon,\lambda}$.

6. Radially symmetric case and power-law permittivity profiles. In this section, we discuss issues of uniqueness and multiplicity of solutions for $(S)_\lambda$ when Ω is a symmetric domain and when f is a radially symmetric permittivity profile. Here, one can again define the corresponding pull-in voltage $\lambda_r^*(\Omega, f)$ requiring the solutions to be radially symmetric, that is,

$$\lambda_r^*(\Omega, f) = \sup\{\lambda;\ (S)_\lambda \text{ has a radially symmetric solution}\}.$$

PROPOSITION 6.1. *Let Ω be a symmetric domain, and let f be a nonnegative bounded radially symmetric permittivity profile on Ω; then the minimal solutions of $(S)_\lambda$ are necessarily radially symmetric, and consequently $\lambda_r^*(\Omega, f) = \lambda^*(\Omega, f)$. Moreover, if Ω is a ball, then any radial solution of $(S)_\lambda$ attains its maximum at 0.*

Proof. It is clear that $\lambda_r^*(\Omega, f) \le \lambda^*(\Omega, f)$, and the reverse will be proved if we establish that every minimal solution of $(S)_\lambda$ with $0 < \lambda < \lambda^*(\Omega, f)$ is radially symmetric. This is a straightforward application of the recursive scheme defined in Theorem 2.2, which gives a radially symmetric function at each step, and therefore the resulting limiting function—which is the minimal solution—is radially symmetric.

For any radially symmetric $u(r)$ of $(S)_\lambda$ defined in the ball of radius R, we have $u_r(0) = 0$ and

$$-u_{rr} - \frac{N-1}{r}u_r = \frac{\lambda f}{(1-u)^2} \qquad \text{in } (0, R).$$

Multiplying by r^{N-1}, we get that $-\frac{d(r^{N-1}u_r)}{dr} = \frac{\lambda f r^{N-1}}{(1-u)^2} \ge 0$, and therefore $u_r < 0$ in $(0, R)$ since $u_r(0) = 0$. This shows that $u(r)$ attains its maximum at 0. ∎

The bifurcation diagrams shown in the introduction actually reflect the radially symmetric situation, and our emphasis in this section is on whether there is a better chance to analyze mathematically the higher branches of solutions in this case. Now some of the classical work of Joseph and Lundgren [14] and many that followed can be adapted to this situation when the permittivity profile is constant. However, the case of a power-law permittivity profile $f(x) = |x|^\alpha$ defined in a unit ball already presents a much richer situation. We now present various analytical and numerical evidence for various conjectures relating to this case, some of which are established rigorously in [9].

Power-law permittivity profiles. Consider the domain Ω to be a unit ball $B_1(0) \subset R^N$ $(N \ge 1)$, and let $f(x) = |x|^\alpha$ $(\alpha \ge 0)$. We analyze in this case the branches of radially symmetric solutions of $(S)_\lambda$ for $\lambda \in (0, \lambda^*]$. In this case, $(S)_\lambda$ reduces to

$$-u_{rr} - \frac{N-1}{r}u_r = \frac{\lambda r^\alpha}{(1-u)^2}, \quad 0 < r \le 1, \quad u'(0) = 0, \quad u(1) = 0. \tag{6.1}$$

Here $r = |x|$ and $0 < u = u(r) < 1$ for $0 < r < 1$. Looking first for a solution of the form $u(r) = 1 - \beta w(P)$ with $P = \gamma r$, where $\gamma,\ \beta > 0$, equation (6.1) implies that $\gamma^2\beta(w'' + \frac{N-1}{P}w') = \frac{\lambda P^\alpha}{\beta^2\gamma^\alpha}\frac{1}{w^2}$. We set $w(0) = 1$ and $\lambda = \gamma^{2+\alpha}\beta^3$. This yields the following initial value problem:

$$w'' + \frac{N-1}{P}w' = \frac{P^\alpha}{w^2}, \quad P > 0\,; \qquad w'(0) = 0,\ w(0) = 1. \tag{6.2}$$

Since $u(1) = 0$ we have $\beta = 1/w(\gamma)$, where $w(\gamma)$ is a solution of (6.2). Therefore, we conclude that

$$u(0) = 1 - \frac{1}{w(\gamma)}, \qquad \lambda = \frac{\gamma^{2+\alpha}}{w^3(\gamma)}. \tag{6.3}$$

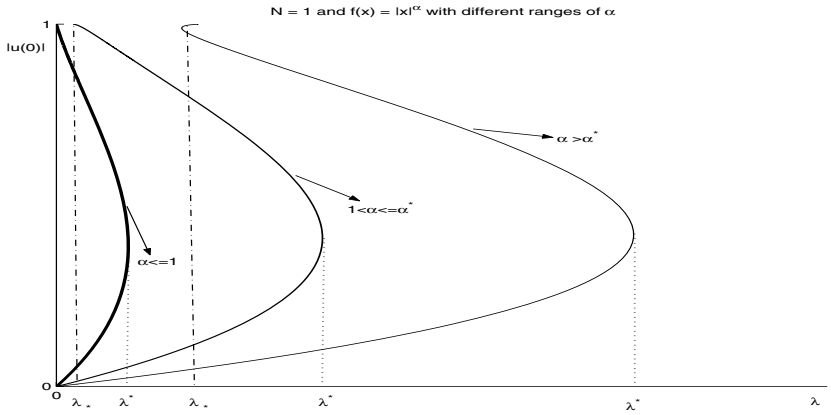


FIG. 6.1. *Plots of $u(0)$ versus λ for the power-law permittivity profile $f(x) = |x|^\alpha$ ($\alpha \geq 0$) defined in the slab domain ($N = 1$). The numerics point to a constant $\alpha^* > 1$ (analytically given in* (6.10)*) such that the bifurcation diagrams are greatly different for different ranges of α: $0 \leq \alpha \leq 1$, $1 < \alpha \leq \alpha^*$, and $\alpha > \alpha^*$.*

As was done in [19], one can numerically integrate the initial value problem (6.2) and use the results to compute the complete bifurcation diagram for (6.1). We show such a computation of $u(0)$ versus λ defined in (6.3) for the slab domain ($N = 1$) in Figure 6.1. In this case, one observes from the numerical results that when $N = 1$ and $0 \leq \alpha \leq 1$, there exist exactly two solutions for $(S)_\lambda$ whenever $\lambda \in (0, \lambda^*)$. On the other hand, the situation becomes more complex for $\alpha > 1$ as $u(0) \to 1$. This leads to the question of determining the asymptotic behavior of $w(P)$ as $P \to \infty$. Towards this end, we proceed as follows.

Setting $\eta = \log P$ and $w(P) = P^B V(\eta) > 0$ for some positive constant B, we obtain from (6.2) that

$$P^{B-2}V'' + (2B + N - 2)P^{B-2}V' + B(B + N - 2)P^{B-2}V = \frac{P^{\alpha-2B}}{V^2}. \tag{6.4}$$

Choosing $B - 2 = \alpha - 2B$ so that $B = (2 + \alpha)/3$, we get that

$$V'' + \frac{3N + 2\alpha - 2}{3}V' + \frac{(2 + \alpha)(3N + \alpha - 4)}{9}V = \frac{1}{V^2}. \tag{6.5}$$

We can already identify from this equation the following regime.

Case 1. Assume that

$$N = 1 \quad \text{and} \quad 0 \leq \alpha \leq 1. \tag{6.6}$$

In this case, there is no equilibrium point for (6.5), which means that the bifurcation diagram vanishes at $\lambda = 0$, from which one infers that in this case there exist exactly two solutions for $\lambda \in (0, \lambda^*)$ and just one for $\lambda = \lambda^*$.

Case 2. N and α satisfy either one of the following conditions:

$$\begin{array}{l} N = 1 \quad \text{and} \quad \alpha > 1, \\ N \geq 2. \end{array} \tag{6.7}$$

There exists then an equilibrium point V_e of (6.5) which must be positive and satisfies

$$V_e^3 = \frac{9}{(2 + \alpha)(3N + \alpha - 4)} > 0. \tag{6.8}$$

Linearizing around this equilibrium point by writing $V = V_e + Ce^{\sigma\eta},\ 0 < C \ll 1$, we obtain that $\sigma^2 + \frac{3N+2\alpha-2}{3}\sigma + \frac{(2+\alpha)(3N+\alpha-4)}{3} = 0$. This reduces to

$$\begin{aligned} &\sigma_\pm = -\frac{3N+2\alpha-2}{6} \pm \frac{\sqrt{\triangle}}{6}, \quad \text{where} \\ &\triangle = -8\alpha^2 - (24N-16)\alpha + (9N^2 - 84N + 100). \end{aligned} \tag{6.9}$$

We note that $\sigma_\pm < 0$ whenever $\triangle \geq 0$. Now define

$$\alpha^* = -\frac{1}{2} + \frac{1}{2}\sqrt{\frac{27}{2}}, \qquad \alpha^{**} = \frac{4 - 6N + 3\sqrt{6}(N-2)}{4} \quad (N \geq 8). \tag{6.10}$$

Next, we discuss different ranges of N and α in terms of their effect on the sign of $\triangle$.

Case 2.A. N and α satisfy either one of the following:

$$\begin{aligned} &N = 1 \quad \text{with } 1 < \alpha \leq \alpha^*, \\ &N \geq 8 \quad \text{with } 0 \leq \alpha \leq \alpha^{**}. \end{aligned} \tag{6.11}$$

In this case, we have $\triangle \geq 0$ and $V \sim (\frac{9}{(2+\alpha)(3N+\alpha-4)})^{\frac{1}{3}} + C_1 e^{-\frac{3N+2\alpha-2-\sqrt{\triangle}}{6}\eta} + \cdots$ as $\eta \to +\infty$. Further, we conclude that $w \sim P^{\frac{2+\alpha}{3}}(\frac{9}{(2+\alpha)(3N+\alpha-4)})^{\frac{1}{3}} + C_1 P^{-\frac{N-2}{2}+\frac{\sqrt{\triangle}}{6}} + \cdots$ as $P \to +\infty$. In both cases, the branch monotonically approaches the value 1 as $\eta \to +\infty$. Moreover, since $\lambda = \gamma^{2+\alpha}/w^3(\gamma)$, we have

$$\lambda \sim \lambda_* = \frac{(2+\alpha)(3N+\alpha-4)}{9} \qquad \text{as } \gamma \to \infty, \tag{6.12}$$

which is another important critical threshold for the voltage.

In the case (6.11) illustrated by Figure 6.1, we have $\lambda_* < \lambda^*$, and the number of solutions increases but remains finite as λ approaches λ_*. On the other hand, in the case of (6.11) illustrated by Figure 6.2(bottom), we have $\lambda_* = \lambda^*$, and there seems to be only one branch of solutions.

Case 2.B. N and α satisfy any one of the following three conditions:

$$\begin{aligned} &N = 1 \quad \text{with } \alpha > \alpha^*, \\ &2 \leq N \leq 7 \quad \text{with } \alpha \geq 0, \\ &N \geq 8 \quad \text{with } \alpha > \alpha^{**}. \end{aligned} \tag{6.13}$$

In this case, we have $\triangle < 0$ and

$$V \sim \left(\frac{9}{(2+\alpha)(3N+\alpha-4)}\right)^{\frac{1}{3}} + C_1 e^{-\frac{3N+2\alpha-2}{6}\eta} cos\left(\frac{\sqrt{-\triangle}}{6}\eta + C_2\right) + \cdots \quad \text{as} \quad \eta \to +\infty.$$

We also have

$$w \sim P^{\frac{2+\alpha}{3}}\left(\frac{9}{(2+\alpha)(3N+\alpha-4)}\right)^{\frac{1}{3}} + C_1 P^{-\frac{N-2}{2}} cos\left(\frac{\sqrt{-\triangle}}{6} logP + C_2\right) + \cdots \quad \text{as } P \to +\infty, \tag{6.14}$$

and from the fact that $\lambda = \gamma^{2+\alpha}/w^3(\gamma)$ we get again that $\lambda \sim \lambda_* = \frac{(2+\alpha)(3N+\alpha-4)}{9}$ as $\gamma \to \infty$. Note the oscillatory behavior of $w(P)$ in (6.14) for large P, which means that

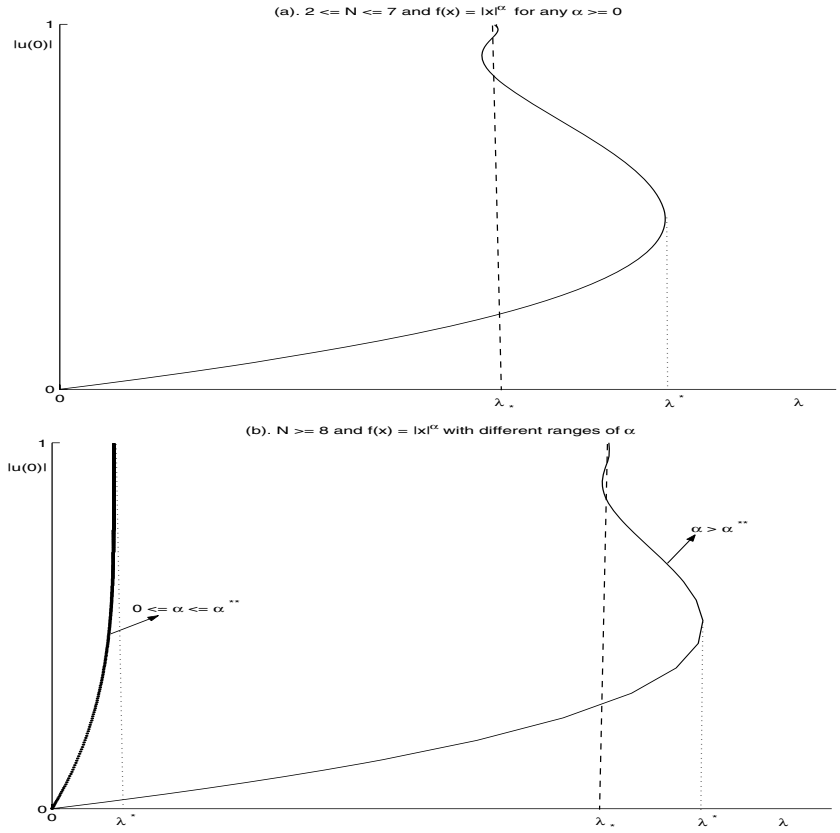


FIG. 6.2. *Top figure: Plots of $u(0)$ versus λ for the power-law permittivity profile $f(x) = |x|^\alpha$ ($\alpha \geq 0$) defined in the unit ball $B_1(0) \subset R^N$ with $2 \leq N \leq 7$. In this case, $u(0)$ oscillates around the value λ_* defined in* (6.12)*, and u^* is regular. Bottom figure: Plots of $u(0)$ versus λ for the power-law permittivity profile $f(x) = |x|^\alpha$ ($\alpha \geq 0$) defined in the unit ball $B_1(0) \subset R^N$ with $N \geq 8$. The characters of the bifurcation diagrams depend on different ranges of α: when $0 \leq \alpha \leq \alpha^{**}$, there exists a unique solution for $(S)_\lambda$ with $\lambda \in (0, \lambda^*)$ and u^* is singular; when $\alpha > \alpha^{**}$, $u(0)$ oscillates around the value λ_* defined in* (6.12) *and u^* is regular.*

$u(0)$ is expected to oscillate around the value $\lambda_* = \frac{(2+\alpha)(3N+\alpha-4)}{9}$ as $P \to \infty$. The diagrams below point to the existence of a sequence $\{\lambda_i\}$ satisfying $\lambda_0 = 0$, $\lambda_1 = \lambda^*$, $\lambda_{2k} \nearrow \lambda_*$, $\lambda_{2k-1} \searrow \lambda_*$ as $k \to \infty$ and such that exactly $2k+1$ solutions for $(S)_\lambda$ exist when $\lambda \in (\lambda_{2k}, \lambda_{2k+2})$, while there are exactly $2k$ solutions when $\lambda \in (\lambda_{2k+1}, \lambda_{2k-1})$. Furthermore, $(S)_\lambda$ has infinitely multiple solutions at $\lambda = \lambda_*$.

The three cases (6.13) considered here for N and α are illustrated by the diagrams in Figures 6.1, 6.2(top), and 6.2(bottom), respectively.

Acknowledgments. We are grateful to Michael. J. Ward for introducing us to the PDE models for electrostatic MEMS devices, and to Louis Nirenberg for leading us to the pioneering work of Joseph and Lundgren and the related papers of Crandall and Rabinowitz. Special thanks go to Pierpaolo Esposito for his thorough reading of the manuscript, which led to many improvements, and to Xavier Cabre for pointing out the related paper of Mignot and Puel, and for providing his own recent preprints on the subject.

REFERENCES

[1] U. Ascher, R. Christiansen, and R. Russell, *Collocation software for boundary value ODE's*, Math. Comp., 33 (1979), pp. 659–679.

[2] C. Bandle, *Isoperimetric inequalities and applications*, in Monographs and Studies in Mathematics, Pitman, Boston, 1980, pp. 184–196.

[3] H. Brezis, T. Cazenave, Y. Martel, and A. Ramiandrisoa, *Blow up for $u_t - \Delta u = g(u)$ revisited*, Adv. Differential Equations, 1 (1996), pp. 73–90.

[4] C.-M. Brauner and B. Nicolaenko, *Sur une classe de problèmes elliptiques non linéaires*, C.R. Acad. Sci. Paris Sér. A, (1978), pp. 1007–1010.

[5] H. Brezis and J. L. Vazquez, *Blow-up solutions of some nonlinear elliptic problems*, Rev. Mat. Univ. Compl. Madrid, 10 (1997), pp. 443–469.

[6] X. Cabré, *Extremal solutions and instantaneous complete blow-up for elliptic and parabolic problems*, preprint, 2005.

[7] M. G. Crandall and P. H. Rabinowitz, *Bifurcation, perturbation of simple eigenvalues and linearized stability*, Arch. Ration. Mech. Anal., 52 (1973), pp. 161–180.

[8] M. G. Crandall and P. H. Rabinowitz, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Ration. Mech. Anal., 58 (1975), pp. 207–218.

[9] P. Esposito, N. Ghoussoub, and Y. Guo, *Compactness along the branch of semi-stable and unstable solutions for an elliptic problem with a singular nonlinearity*, Comm. Pure Appl. Math., (2006), to appear.

[10] G. Flores, G. A. Mercado, and J. A. Pelesko, *Dynamics and touchdown in electrostatic MEMS*, in Proceedings of the 2003 International Conference on *MEMS*, Nano, and Smart Systems (ICMENS 2003) Banff, AB, 2003, IEEE Computer Society Press, Piscataway, NJ, pp. 182–187.

[11] N. Ghoussoub and Y. Guo, *On the partial differential equations of electrostatic MEMS devices* II*: Dynamic case*, submitted, 2006.

[12] Y. Guo, Z. Pan, and M. J. Ward, *Touchdown and pull-in voltage behavior of a MEMS device with varying dielectric properties*, SIAM J. Appl. Math., 66 (2005), pp. 309–338.

[13] A. Haraux and F. B. Weissler, *Non-uniqueness for a semilinear initial value problem*, Indiana Univ. Math. J., 31 (1982), pp. 167–189.

[14] D. D. Joseph and T. S. Lundgren, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Ration. Mech. Anal., 49 (1973), pp. 241–268.

[15] J. P. Keener and H. B. Keller, *Positive solutions of convex nonlinear eigenvalue problems*, J. Differential Equations, 16 (1974), pp. 103–125.

[16] Y. Martel, *Uniqueness of weak extremal solutions of nonlinear elliptic problems*, Houston J. Math., 23 (1997), pp. 161–168.

[17] F. Mignot and J. P. Puel, *Sur une classe de problèmes non linéaires avec non linéarité positive, croissante, convexe*, Comm. Partial Differential Equations, 5 (1980), pp. 791–836.

[18] H. C. Nathanson, W. E. Newell, R. A. Wickstrom, and J. R. Davis, *The resonant gate transistor*, IEEE Trans. Elect. Devices, 14 (1967), pp. 117–133.

[19] J. A. Pelesko, *Mathematical modeling of electrostatic MEMS with tailored dielectric properties*, SIAM J. Appl. Math., 62 (2002), pp. 888–908.

[20] J. A. Pelesko and D. H. Bernstein, *Modeling MEMS and NEMS*, Chapman and Hall/CRC Press, Boca Raton, FL, 2002.

[21] I. Stakgold, *Green's Functions and Boundary Value Problems*, Wiley, New York, 1998.

[22] G. I. Taylor, *The coalescence of closely spaced drops when they are at different electric potentials*, Proc. Roy. Soc. A, 306 (1968), pp. 423–434.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1450–1473

ASYMPTOTICS OF THE POISSON PROBLEM IN DOMAINS WITH CURVED ROUGH BOUNDARIES*

ALEXANDRE L. MADUREIRA† AND FRÉDÉRIC VALENTIN†

Abstract. Effective boundary conditions (wall laws) are commonly employed to approximate PDEs in domains with rough boundaries, but it is neither easy to design such laws nor to estimate the related approximation error. A two-scale asymptotic expansion based on a domain decomposition result is used here to mitigate such difficulties, and as an application we consider the Poisson equation. The proposed scheme considers rough curved boundaries and allows a complete asymptotic expansion for the solution, highlighting the influence of the boundary curvature. The derivation and estimation of high order effective conditions is a corollary of such development. Sharp estimates for first and second order wall law approximations are considered for different Sobolev norms and show superior convergence rates in the interior of the domain. A numerical test illustrates several of the results obtained here.

Key words. Poisson equation, rough boundary, effective boundary conditions, asymptotic expansion, wall laws, curved domain

AMS subject classifications. 58J37, 35J05, 35J25

DOI. 10.1137/050633895

1. Introduction.

In several applications, it is necessary to solve PDEs in domains with boundaries that are rough. Analytic solutions are rarely available, and direct numerical computations are usually out of reach since the rapidly varying wrinkles and the domain have different length scales. The traditional remedy is to pose special boundary conditions on a mollified domain to capture the geometrical influence of the wrinkles. The development of such conditions is cumbersome in general, and modeling error estimates can be out of reach. The aim of this paper is to investigate and explicate such issues.

Problems posed on domains with rough boundary pervade many fields of research. In aerodynamics, aircrafts and space shuttles are often covered with tiles, hence their walls have an array of periodic gaps [22].

Similarly, small air injecting nozzles are periodically introduced over wings of aircrafts to decrease the drag [6]. Another interesting example in the fluid mechanics is the flow field around golf balls, in which the wrinkles associated to the curvature decrease the gap between the air-pressure behind and in front of the ball. Finally, in hemodynamics, the cell surfaces of the endothelium modifies the wall shear stress produced by the blood flow, and realistic computer simulations must take this effect into account [32].

To avoid discretizing such intricate boundaries, practitioners start resorting to *wall laws*, which are effective boundary conditions that try to emulate the effect of the wrinkles without actually resolving them. Ingenious methods were developed, some in ad hoc fashion, but many of them based on firm mathematical ground.

*Received by the editors June 16, 2005; accepted for publication (in revised form) July 26, 2006; published electronically January 12, 2007.

http://www.siam.org/journals/sima/38-5/63389.html

†Departamento de Matemática Aplicada, Laboratório Nacional de Computação Científica, Av. Getúlio Vargas 333, CEP 25651-070 Petrópolis - RJ, Brazil (alm@lncc.br, valentin@lncc.br). The first author was partially supported by the CNPq/Brazil Project 306104/2004-0, and the second author by the CNPq/Brazil Project 300348/2003-7.

Nonetheless, even when mathematics played a significant role and error estimates were found, some qualitative aspects of the resulting models that were observed numerically were missed by the theory. For instance, close to the wrinkles, the exact solution wiggles, where the model solution does not. Hence the approximation there is precarious in the H^1 norm, but fine in the L^2 norm. However, far from the boundary, where the solution is "smooth," a better approximation occurs, even when derivatives are considered. Previously, some authors considered some of these effects, but it seems hard to generalize their results to other operators, or to second order, curvature dependent approximations.

There are several papers devoted to finding good wall laws as well as the corresponding modeling errors estimates. Most of the articles fit in the framework of two-scale asymptotic expansions [14, 23, 29, 30]. For fluids, [9, 11] deal with the Stokes equations, and [5, 7, 13, 24] focus on the steady and unsteady incompressible Navier–Stokes equations. An interesting alternative way to derive effective boundary conditions is based on domain decomposition strategies as introduced in [3] for the Laplace operator, and extended to other operators [4, 7]; see also [31] for a survey and [10, 27, 28] for related techniques and problems.

The previous references considered the wrinkles to be laid upon a flat line or surface, or considered first order approximations only. In [1] first and second order models for diffraction of an electromagnetic wave by a cylindrical curved grating were considered, and some H^1 norm estimates were obtained. The references [2, 20, 21] also developed wall laws for wave scattering problems.

We extend here the results of [25], where we considered first and second order wall laws for general curved boundaries. We estimate the modeling errors in the L^2 and H^1 norms, both on the whole domain and in its interior, confirming several numerical predictions. Our mathematical framework mixes two-scale asymptotic expansions and domain decomposition ideas. Using such procedure, wall laws of arbitrary order come by naturally, and we derive first and second order effective boundary conditions. Local boundary fitted coordinates expose how the exact solution depends on the curvature of the boundary (in a sense that we make clear in what follows). This is crucial to develop high order models, which depend on the curvature. We believe that our approach is quite general and can handle more sophisticated operators.

We now outline the contents of this paper. In the next section, we introduce basic definitions and highlight the main ingredients of the approach. Section 3 presents wall laws of different orders, along with error estimates and a summary of effective conditions. Section 4 contains the development of the asymptotic expansion, and the details necessary to define the boundary layer terms are in section 5. The errors associated with the asymptotic expansion are considered in section 6. Finally, in section 7 we validate the models numerically.

We now briefly introduce and explain some basic notation that we use throughout this paper. As usual, if D is an open set, then ∂D denotes its boundary, $\overline{D}$ its closure, $L^2(D)$ is the set of square integrable functions in D, and $H^s(D)$ is the corresponding Sobolev space of order s, for a real number s. We denote the norms of those spaces by $\|\cdot\|_{L^2(D)}$ and $\|\cdot\|_{H^s(D)}$. Also, the symbol $\cdot|_D$ denotes the restriction of a function to the domain D. Without loss of generality, we have chosen to work in two dimensions. Nonetheless, all that follows can be generalized to the three-dimensional case. Bold fonts indicate two-dimensional vectors, and the symbol $\partial_{\boldsymbol{n}}$ indicates the (outward) normal derivative with respect to the domain Ω_s. Similarly, ∂_x denotes the derivative with respect to the variable x, etc. We denote by c a generic constant (not necessarily

the same in all occurrences) which is independent of ε, but may depend on Ω_s and Sobolev norms of f.

2. Definitions and main results. We denote the domain of interest by $\Omega^\varepsilon \subset \mathbb{R}^2$, which is open, bounded, and ε-dependent. Here, ε indicates the length scale of the roughness element. It is convenient to consider $\Omega^\varepsilon = \Omega_s \cup \Omega_r^\varepsilon \cup \Gamma$, where the limit domain Ω_s is open and ε-independent, the open set Ω_r^ε depends on ε with $\Omega_s \cap \Omega_r^\varepsilon = \emptyset$, and the interface $\Gamma = \partial\Omega_s \cap \partial\Omega_r^\varepsilon$. The precise definition of these subdomains follow.

We assume that Ω_s has its boundary $\partial\Omega_s$ constituted of two disjoint parts, a smooth inner boundary Γ, and a Lipschitz-continuous outer boundary. We arc length parameterize the smooth curve Γ by an ε-independent function $\boldsymbol{\psi} : \mathbb{R} \to \mathbb{R}^2$, which is periodic with period L, and injective in $(0, L)$. In other words, Γ is a simple closed curve with length L, and it bounds a region of the plane, which we call its interior. We orient Γ in such a way that it is positively oriented, i.e., going along the direction of increasing parameter, the interior of the curve stays on the left. We assume that $\varepsilon = L/N$, for some positive integer N.

The domain Ω_r^ε has Γ as its outer boundary and Γ_r^ε as its inner boundary. The curve Γ_r^ε, which is also closed, is defined as a perturbation of Γ, and is parameterized by

$$\boldsymbol{\psi}^\varepsilon(\theta) = \boldsymbol{\psi}(\theta) + \varepsilon\big[d_0 - \psi_r(\varepsilon^{-1}\theta)\big]\,\boldsymbol{n}(\theta),$$

where $\boldsymbol{n}$ is the normal vector pointing towards the interior of Γ, and $d_0 > 1$ is such that $d_0\varepsilon$ is smaller than the minimum radius of curvature of Γ. The function $\psi_r : \mathbb{R} \to \mathbb{R}$ is independent of ε, Lipschitz-continuous with $\psi_r(0) = 0$, and periodic with period 1. Without loss of generality, we assume that $\|\psi_r\|_{L^\infty(\mathbb{R})} = 1$. Formally,

$$\Omega_r^\varepsilon \,=\, \{\, \mathbf{x} = \boldsymbol{\psi}(\theta) + (\varepsilon\, d_0 - \rho)\, \boldsymbol{n}(\theta) \,:\, \theta \in [0, L),\, \rho \in (\varepsilon\, \psi_r(\varepsilon^{-1}\theta), \varepsilon\, d_0)\,\}.$$

Hence, Ω^ε has its boundary constituted of two parts, a rough inner boundary Γ_r^ε, and a Lipschitz-continuous outer boundary that is independent of ε and does not intersect Γ. Note that Γ splits the original domain Ω^ε into two regions Ω_s and Ω_r^ε, one ε-independent, and the other containing the wrinkles. The subdomain Ω_r^ε is the set of points between Γ and Γ_r^ε, and Ω_s is the ε-independent domain comprehended between Γ and the outer boundary of Ω^ε. This is the set of all points at least "slightly away" from the wrinkles. See Figure 1. Finally, we denote a typical point in it by $\mathbf{x} = (x_1, x_2)$.

We consider the problem

$$\begin{aligned} -\Delta\, u^\varepsilon &= f \quad \text{in } \Omega^\varepsilon, \\ u^\varepsilon &= 0 \quad \text{on } \partial\Omega^\varepsilon, \end{aligned} \tag{1}$$

where f has support in Ω_s.

It is clear that the solution u^ε depends in a nontrivial way on the small parameter ε. It is our goal to unfold this dependence and show how to develop models for (1). It is possible to expand u^ε in a formal power series with respect to ε. This expansion is far from trivial since it has to take into account effects from the wrinkles as well as from the curvature. To focus on the main steps of our approach, we start by presenting the first few terms of this expansion, and only in Ω_s. The details of the asymptotics are considered in section 4.

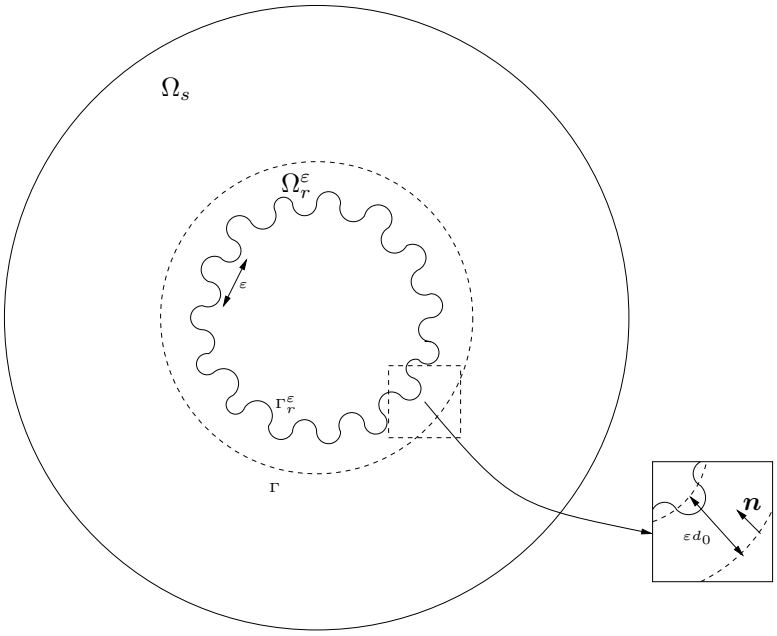


FIG. 1. *The domain* Ω^ε.

The asymptotic expansion of u^ε in Ω_s is a formal combination of an ε-independent part and a highly oscillatory part which decays exponentially to zero away from Γ,

$$u^\varepsilon \sim u^0 + \varepsilon\, u^1 + \varepsilon\, W^{1,0} + \dots \quad \text{in } \Omega_s. \tag{2}$$

While u^0, u^1 are ε-independent, the oscillatory function $W^{1,0}$ depends on ε, but only in a trivial manner.

It is natural to define the first term of the asymptotic such that

$$\begin{aligned} -\Delta\, u^0 &= f \quad \text{in } \Omega_s, \\ u^0 &= 0 \quad \text{on } \partial\Omega_s. \end{aligned} \tag{3}$$

To continue the description of the expansion, it is necessary to introduce a *cell problem*. This is no different from other singularly perturbed problems, perhaps elliptic PDEs with highly oscillatory coefficients being the most notorious. Such cell problems are a essential part in up-scaling procedures and brings information related to the small scale geometry into the large scale behavior of the solution.

In the present case, the cell problem is defined in the semi-infinite strip Ω_r, which "contains" the geometry of the wrinkles,

$$\Omega_r = \{\, (\hat\theta, \hat\rho) \in \mathbb{R}^2 \,:\, \hat\theta \in (0,1),\, \hat\rho \in (\psi_r(\hat\theta), +\infty)\,\},$$

i.e., Ω_r occupies the region delimited by straight lateral boundaries at $\hat\theta = 0$ and $\hat\theta = 1$, and by the lower boundary $\Gamma_r = \{\, (\hat\theta, \psi_r(\hat\theta)) \,:\, \hat\theta \in (0,1)\,\}$; see Figure 2.

We define $C^\infty_{per}(\Omega_r)$ by restricting to Ω_r the functions in $C^\infty(\mathbb{R}^2)$ which are one-periodic with respect to $\hat\theta$. Let $H^1_{per}(\Omega_r)$ be the closure of $C^\infty_{per}(\Omega_r)$ with respect to the $H^1(\Omega_r)$ norm. We also introduce the space of exponentially decaying functions

$$S(\Omega_r) = \{ w \in H^1_{per}(\Omega_r) \,:\, w\, e^{\alpha\hat\rho} \in H^1(\Omega_r) \text{ for some } \alpha > 0 \}\,.$$

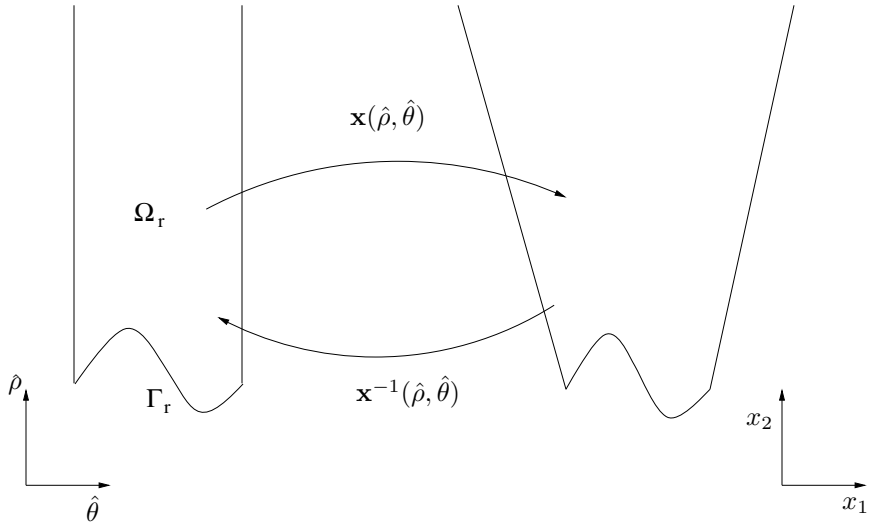


FIG. 2. *The cell domain.*

The following result guarantees that certain Poisson problems posed in Ω_r are well posed, and the solutions have nice properties. The reference [5] deals with related questions for the Stokes operator.

LEMMA 1. *Let $F \in S^*(\Omega_r)$, the dual space of $S(\Omega_r)$. Then there is a unique solution $w \in H^1_{loc}(\Omega_r)$ that is one-periodic with respect to $\hat{\theta}$, and such that $\nabla w \in L^2(\Omega_r)$, and*

$$\begin{aligned} (\partial_{\hat{\theta}\hat{\theta}} + \partial_{\hat{\rho}\hat{\rho}})w &= F \quad \text{in } \Omega_r, \\ w &= \hat{\rho} \quad \text{on } \Gamma_r. \end{aligned} \tag{4}$$

Moreover, there exists a unique constant z such that $w - z \in S(\Omega_r)$, and, if $F \equiv 0$,

$$z \leq \|\psi_r\|_{L^\infty(\mathbb{R})}.$$

Proof. A simple modification of the beautiful arguments of [8, Lemma 4.4] guarantees well posedness and yields a proof of the decaying behavior of the solution towards a constant. Assume now that w is harmonic, and for $t \geq \|\psi_r\|_{L^\infty(\mathbb{R})}$, let $\gamma_t = (0,1) \times \{t\}$. Then Green's identity yields that $\int_{\gamma_t} \partial_{\boldsymbol{n}} w$ is constant with respect to t. Letting $t \to \infty$, we have that actually $\int_{\gamma_t} \partial_{\boldsymbol{n}} w = 0$ for all $t \geq \|\psi_r\|_{L^\infty(\mathbb{R})}$. Using again Green's identity in $S_{t,\tilde{t}} = (0,1) \times (t,\tilde{t})$, for $\tilde{t} > t \geq \|\psi_r\|_{L^\infty(\mathbb{R})}$, we gather that

$$\int_{\partial S_{t,\tilde{t}}} w\, \partial_{\boldsymbol{n}}\, \hat{\rho} = \int_{\partial S_{t,\tilde{t}}} \hat{\rho}\, \partial_{\boldsymbol{n}}\, w = 0.$$

Thus $\int_{\gamma_t} w = \int_{\gamma_{\tilde{t}}} w$, and letting $\tilde{t} \to \infty$, we see that $z = \int_{\gamma_t} w$. Then $z \leq \|w\|_{L^\infty(\gamma_t)}$, and we conclude from the maximum principle [17] that $z \leq \|\psi_r\|_{L^\infty(\mathbb{R})}$. ∎

Remark 1. Note that $S^*(\Omega_r)$ contains, for instance, functions that grow at most algebraically with respect to $\hat{\rho}$.

We define $w^{0,0} \in S(\Omega_r)$, and the constant $z^{0,0}$ as the solution of

$$\begin{aligned} (\partial_{\hat{\theta}\hat{\theta}} + \partial_{\hat{\rho}\hat{\rho}})\, w^{0,0} &= 0 \quad \text{in } \Omega_r, \\ w^{0,0} &= \hat{\rho} - z^{0,0} \quad \text{on } \Gamma_r. \end{aligned} \tag{5}$$

It follows immediately from Lemma 1 that (5) is well defined. Both $z^{0,0}$ and $w^{0,0}$ are related to the boundary layers that naturally appear in the original problem.

To incorporate the influence of the cell problem into the asymptotic expansion (2), we introduce boundary fitted coordinates (θ, ρ) for points "close enough" to Γ; see [12]. Let ρ_0 be a positive number smaller than the minimum radius of curvature of Γ. For a given $\theta \in [0, L)$ and $\rho \in (\varepsilon\, \psi_r(\varepsilon^{-1}\theta), \varepsilon\, d_0 + \rho_0)$, we have

$$\mathbf{x}(\theta, \rho) = \boldsymbol{\psi}(\theta) + (\varepsilon\, d_0 - \rho)\, \boldsymbol{n}(\theta) \in \Omega^\varepsilon.$$

Note that $|\rho - \varepsilon\, d_0| = \operatorname{dist}(\mathbf{x}, \Gamma)$ is the distance between $\mathbf{x}$ and Γ and that the above map defines a local diffeomorphism. The change of coordinates $\mathbf{x} \to (\rho, \theta)$ is not well defined globally in Ω^ε, but only for points with distance from Γ smaller than the minimum radius of curvature of Γ.

In such a new system of coordinates, we simply write the normal derivative of u^0 at a point $\mathbf{x} \in \Gamma$ as $\partial_{\boldsymbol{n}}\, u^0(\theta, \varepsilon\, d_0)$, where θ is such that $\mathbf{x} = \boldsymbol{\psi}(\theta)$. We set

$$W^{1,0}(\theta, \rho) = \Upsilon(\varepsilon d_0 + \rho)\, w^{0,0}(\varepsilon^{-1}\theta, \varepsilon^{-1}\rho)\, \partial_{\boldsymbol{n}}\, u^0(\theta, \varepsilon\, d_0) \tag{6}$$

in the formal expansion (2), where $\Upsilon(\cdot)$ is a smooth ε-independent cutoff function, such that $\Upsilon(\rho)$ equals one if ρ is smaller than a fixed number smaller than ρ_0, and vanishes for $\rho \geq \rho_0$. For instance, we may set Υ identically equal to one in $(-\infty, \rho_0/3]$ and vanishing in $[\rho_0, +\infty)$. The following estimates follow from standard regularity results and scaling arguments (see also Lemma 3):

$$\|W^{1,0}\|_{L^2(\Omega^\varepsilon)} \leq c\, \varepsilon^{1/2}, \qquad \|W^{1,0}\|_{H^1(\Omega^\varepsilon)} \leq c\, \varepsilon^{-1/2}. \tag{7}$$

Finally, let

$$\begin{gathered} -\Delta\, u^1 = 0 \quad \text{in } \Omega_s, \\ u^1 = \left(-d_0 + z^{0,0}\right)\, \partial_{\boldsymbol{n}}\, u^0 \quad \text{on } \Gamma, \qquad u^1 = 0 \quad \text{on } \partial\Omega_s \backslash \Gamma. \end{gathered} \tag{8}$$

Albeit (2) is formal, we show below (Theorem 4) that if

$$e = u^\varepsilon - u^0 - \varepsilon\, u^1 - \varepsilon\, W^{1,0},$$

then there exists an ε-independent constant c such that

$$\|e\|_{H^1(\Omega_s)} \leq c\, \varepsilon^{3/2}. \tag{9}$$

Several other estimates follow from a combination of (9), the triangle inequality, and (7). For instance, we easily find that

$$\|u^\varepsilon - u^0 - \varepsilon\, u^1\|_{H^1(\Omega_s)} \leq \|e\|_{H^1(\Omega_s)} + \|\varepsilon\, W^{1,0}\|_{H^1(\Omega_s)} \leq c\, \varepsilon^{1/2}, \tag{10}$$

$$\|u^\varepsilon - u^0 - \varepsilon\, u^1\|_{L^2(\Omega_s)} \leq \|e\|_{H^1(\Omega_s)} + \|\varepsilon\, W^{1,0}\|_{L^2(\Omega_s)} \leq c\, \varepsilon^{3/2}. \tag{11}$$

The culprit for the low convergence rates in some of the estimates above are the boundary layers. Hence, *interior estimates*, i.e., estimates that bound the errors in domains that are away from the boundary ought to show better rates. It is possible to obtain such estimates by adding a higher order boundary layer term similar to (6) to the expansion. The new term, which we denote by $\breve{\mathrm{W}}$, behaves like $W^{1,0}$, i.e., decays exponentially fast to zero with ρ/ε, and

$$\|\breve{W}\|_{L^2(\Omega^\varepsilon)} \leq c\, \varepsilon^{1/2}, \qquad \|\breve{W}\|_{H^1(\Omega^\varepsilon)} \leq c\, \varepsilon^{-1/2}.$$

TABLE 1
Relative error convergence rates for a zeroth order model.

quantity	$L^2(\Omega_s)$ error	$L^2(\Omega_s^{int})$ error
u	$O(\varepsilon)$	$O(\varepsilon)$
∇u	$O(\varepsilon^{1/2})$	$O(\varepsilon)$

For the sake of simplicity, we do not describe $\check{W}$ now. Such a function is defined after two different cell problems are solved; see (19) and (20). It is now possible to derive a better estimate in the H^1 norm:

$$\|u^\varepsilon - u^0 - \varepsilon\, u^1 - \varepsilon\, W^{1,0} - \varepsilon^2\, \check{W}\|_{H^1(\Omega_s)} \le c\,\varepsilon^2. \tag{12}$$

Finally, let $\Omega_s^{int} \subset \Omega_s$ be such that $\overline{\Omega_s^{int}} \cap \overline{\Gamma} = \emptyset$. Then

$$\begin{aligned} &\|u^\varepsilon - u^0 - \varepsilon\, u^1\|_{H^1(\Omega_s^{int})} \\ &\le \|u^\varepsilon - u^0 - \varepsilon\, u^1 - \varepsilon\, W^{1,0} - \varepsilon^2\, \check{W}\|_{H^1(\Omega_s)} + \|\varepsilon\, W^{1,0}\|_{H^1(\Omega_s^{int})} + \|\varepsilon^2\, \check{W}\|_{H^1(\Omega_s^{int})} \le c\,\varepsilon^2. \end{aligned} \tag{13}$$

Note in (13) that the exponential decay of both $W^{1,0}$ and $\check{W}$ guarantees that their $H^1(\Omega_s^{int})$ norms are also exponentially small and hence bounded by $c\,\varepsilon^2$.

3. Derivation of wall laws.

3.1. Zeroth order wall law. A first attempt to approximate u^ε would use u^0. It immediately follows from (10), (11) and regularity estimates for u^1 that

$$\begin{aligned} \|u^\varepsilon - u^0\|_{H^1(\Omega_s)} &\le \|u^\varepsilon - u^0 - \varepsilon\, u^1\|_{H^1(\Omega_s)} + \|\varepsilon\, u^1\|_{H^1(\Omega_s)} \le c\,\varepsilon^{1/2}, \\ \|u^\varepsilon - u^0\|_{L^2(\Omega_s)} &\le \|u^\varepsilon - u^0 - \varepsilon\, u^1\|_{L^2(\Omega_s)} + \|\varepsilon\, u^1\|_{L^2(\Omega_s)} \le c\,\varepsilon. \\ \|u^\varepsilon - u^0\|_{L^2(\Omega_s^{int})} &\le \|u^\varepsilon - u^0 - \varepsilon\, u^1\|_{H^1(\Omega_s^{int})} + \|\varepsilon\, u^1\|_{H^1(\Omega_s)} \le c\,\varepsilon. \end{aligned}$$

The $O(\varepsilon^{1/2})$ error in the $H^1(\Omega_s)$ norm is due to the inability of this approximation to capture the oscillatory behavior of the solution close to the wrinkles. This explains the better performance in the $L^2(\Omega_s)$ and interior norms. Table 1 presents various relative error estimates with respect to ε, including interior estimates.

3.2. First order wall law. Inspired by (2), (10)–(13), we would like to approximate u^ε by the first terms of its asymptotic expansion, but without solving the PDEs that define these terms. A first step in this direction is to consider only the functions that actually have influence in the interior of the domain, i.e., we assume

$$u^\varepsilon \approx u^0 + \varepsilon\, u^1. \tag{14}$$

Thus, over Γ, from (3), (8), and (14),

$$u^\varepsilon \approx \varepsilon\,(-d_0 + z^{0,0})\,\partial_{\boldsymbol{n}}\, u^0, \qquad \partial_{\boldsymbol{n}}\, u^\varepsilon \approx \partial_{\boldsymbol{n}}\, u^0 + \varepsilon\, \partial_{\boldsymbol{n}}\, u^1. \tag{15}$$

So

$$u^\varepsilon + \varepsilon\,(d_0 - z^{0,0})\,\partial_{\boldsymbol{n}}\, u^\varepsilon \approx \varepsilon^2\,(d_0 - z^{0,0})\,\partial_{\boldsymbol{n}}\, u^1,$$

TABLE 2
Relative error convergence rates for a first order model.

quantity	$L^2(\Omega_s)$ error	$L^2(\Omega_s^{int})$ norm error
u	$O(\varepsilon^{3/2})$	$O(\varepsilon^2)$
∇u	$O(\varepsilon^{1/2})$	$O(\varepsilon^2)$

on Γ, and this amount can be small enough for certain applications. We define then $\bar{u} \in H^1(\Omega_s)$ approximating u^ε in Ω_s by

$$\begin{aligned} -\Delta\,\bar{u} &= f \quad \text{in } \Omega_s, \\ \bar{u} + \varepsilon\,(d_0 - z^{0,0})\,\partial_{\boldsymbol{n}}\,\bar{u} &= 0 \quad \text{on } \Gamma, \qquad \bar{u} = 0 \quad \text{on } \partial\Omega_s\backslash\Gamma. \end{aligned} \tag{16}$$

It follows from Lemma 1 that $z^{0,0} \leq \|\psi_r\|_{L^\infty(\mathbb{R})}$, and since $\|\psi_r\|_{L^\infty(\mathbb{R})} = 1 < d_0$, the difference $d_0 - z^{0,0}$ is positive. Thus (16) is well posed for all positive ε.

To estimate the modeling error, we first note that if $\bar{e} = \bar{u} - u^0 - \varepsilon\,u^1$, then

$$\begin{aligned} -\Delta\,\bar{e} &= 0 \quad \text{in } \Omega_s, \\ \bar{e} + \varepsilon\,(d_0 - z^{0,0})\,\partial_{\boldsymbol{n}}\,\bar{e} &= -\varepsilon^2\,(d_0 - z^{0,0})\,\partial_{\boldsymbol{n}}\,u^1 \quad \text{on } \Gamma, \qquad \bar{e} = 0 \quad \text{on } \partial\Omega_s\backslash\Gamma. \end{aligned} \tag{17}$$

It follows from regularity estimates [16, Theorem 4.24] that there exists an ε-independent constant c such that

$$\|\bar{e}\|_{H^1(\Omega_s)} \leq c\,\varepsilon^2.$$

The modeling error estimates are then as follows:

$$\|u^\varepsilon - \bar{u}\|_{H^1(\Omega_s)} \leq \|u^\varepsilon - u^0 - \varepsilon\,u^1\|_{H^1(\Omega_s)} + \|\bar{u} - u^0 - \varepsilon\,u^1\|_{H^1(\Omega_s)} \leq c\,\varepsilon^{1/2},$$

where we used the triangle inequality and (10).

Analogously, using (11), (13), we obtain L^2 and interior estimates

$$\|u^\varepsilon - \bar{u}\|_{L^2(\Omega_s)} \leq c\,\varepsilon^{3/2} \qquad \|u^\varepsilon - \bar{u}\|_{H^1(\Omega_s^{int})} \leq c\,\varepsilon^2.$$

We summarize the convergence results in Table 2.

3.3. Second order wall law. The derivation of higher order approximations to u^ε follows the same *modus operandi* as in the previous subsection. We first consider only the terms that have influence away from Γ_r^ε and assume that

$$u^\varepsilon \approx u^0 + \varepsilon\,u^1 + \varepsilon^2\,u^2. \tag{18}$$

To define the term u^2 above we introduce two new cell problems, seeking $w^{1,0}$ and $w^{1,1}$ in $S(\Omega_r)$, and the constants $z^{1,0}$ and $z^{1,1}$ satisfying

$$-(\partial_{\hat{\theta}\hat{\theta}} + \partial_{\hat{\rho}\hat{\rho}})\,w^{1,0} = \chi - \partial_{\hat{\rho}}\,w^{0,0} + 2\,\hat{\rho}\,\partial_{\hat{\theta}\hat{\theta}}\,w^{0,0} \quad \text{in } \Omega_r, \qquad w^{1,0} = -z^{1,0} \quad \text{on } \Gamma_r, \tag{19}$$

$$-(\partial_{\hat{\theta}\hat{\theta}} + \partial_{\hat{\rho}\hat{\rho}})\,w^{1,1} = 2\,\partial_{\hat{\theta}}\,w^{0,0} \quad \text{in } \Omega_r, \qquad w^{1,1} = z^{1,1} \quad \text{on } \Gamma_r, \tag{20}$$

where $\chi(\hat{\rho}) = 1$ if $\hat{\rho} < d_0$, and $\chi(\hat{\rho}) = 0$ if $\hat{\rho} \geq d_0$. The previous cell problems are well posed, as Lemma 1 guarantees. The expression of $\check{W}$ mentioned on page 1455 is as follows:

$$\begin{aligned}\check{W}(\theta,\rho) = \Upsilon(\varepsilon d_0+\rho)\big[&w^{1,0}(\varepsilon^{-1}\theta,\varepsilon^{-1}\rho)\kappa(\theta)\,\partial_{\boldsymbol{n}}\,u^0(\theta,\varepsilon\,d_0)\\ &+ w^{1,1}(\varepsilon^{-1}\theta,\varepsilon^{-1}\rho)\partial_\theta\,\partial_{\boldsymbol{n}}\,u^0(\theta,\varepsilon\,d_0) + w^{0,0}(\varepsilon^{-1}\theta,\varepsilon^{-1}\rho)\,\partial_{\boldsymbol{n}}\,u^1(\theta,\varepsilon\,d_0)\big],\end{aligned}$$

where $\kappa(\theta)$ is the curvature of Γ at the point $\boldsymbol{\psi}(\theta)$.

Next, we define u^2 by

$$\begin{gathered}-\Delta\,u^2 = 0 \quad \text{in } \Omega_s,\\ u^2 = (-d_0+z^{0,0})\,\partial_{\boldsymbol{n}}\,u^1 + z^{1,0}\kappa\,\partial_{\boldsymbol{n}}\,u^0 + z^{1,1}\partial_\theta\,\partial_{\boldsymbol{n}}\,u^0 \quad \text{on } \Gamma, \qquad u^2 = 0 \quad \text{on } \partial\Omega_s\backslash\Gamma,\end{gathered}$$

and the estimates below follow:

$$\|u^\varepsilon - u^0 - \varepsilon\,u^1 - \varepsilon^2\,u^2\|_{H^1(\Omega_s)} \leq c\,\varepsilon^{1/2}, \tag{21}$$

$$\|u^\varepsilon - u^0 - \varepsilon\,u^1 - \varepsilon^2\,u^2\|_{L^2(\Omega_s)} \leq c\,\varepsilon^{3/2}, \tag{22}$$

$$\|u^\varepsilon - u^0 - \varepsilon\,u^1 - \varepsilon^2\,u^2\|_{H^1(\Omega_s^{int})} \leq c\,\varepsilon^{3}. \tag{23}$$

If (18) holds, then

$$\begin{aligned}u^\varepsilon &\approx \varepsilon\,(-d_0+z^{0,0}+\varepsilon\,z^{1,0}\kappa)\,\partial_{\boldsymbol{n}}\,u^0 + \varepsilon^2\,(-d_0+z^{0,0})\,\partial_{\boldsymbol{n}}\,u^1 + \varepsilon^2\,z^{1,1}\partial_\theta\,\partial_{\boldsymbol{n}}\,u^0\\ &= \varepsilon\,(-d_0+z^{0,0}+\varepsilon\,z^{1,0}\kappa)\,\partial_{\boldsymbol{n}}\,u^0 + \varepsilon^2\,(-d_0+z^{0,0})\,\partial_{\boldsymbol{n}}\,u^1 + \varepsilon^2\,\frac{z^{1,1}}{-d_0+z^{0,0}}\partial_\theta u^1\end{aligned}$$

over Γ, where we used from (8) that $\partial_\theta u^1 = (-d_0+z^{0,0})\partial_\theta\,\partial_{\boldsymbol{n}}\,u^0$ to obtain the last equality. We also have

$$\partial_{\boldsymbol{n}}\,u^\varepsilon \approx \partial_{\boldsymbol{n}}\,u^0 + \varepsilon\,\partial_{\boldsymbol{n}}\,u^1 + \varepsilon^2\,\partial_{\boldsymbol{n}}\,u^2, \qquad \partial_\theta u^\varepsilon \approx \partial_\theta u^0 + \varepsilon\,\partial_\theta u^1 + \varepsilon^2\,\partial_\theta u^2,$$

over Γ. Hence,

$$\begin{aligned}u^\varepsilon + (\varepsilon\,d_0 - \varepsilon\,z^{0,0} - \varepsilon^2\,z^{1,0}\,\kappa)\,\partial_{\boldsymbol{n}}\,u^\varepsilon - \varepsilon^2\,\frac{z^{1,1}}{-d_0+z^{0,0}}\,\partial_\theta u^\varepsilon &\approx -\varepsilon^3\,z^{1,0}\,\kappa\,\partial_{\boldsymbol{n}}\,u^1\\ - \varepsilon^3(-d_0+z^{0,0}+\varepsilon\,z^{1,0}\,\kappa)\,\partial_{\boldsymbol{n}}\,u^2 - \varepsilon^3\,\frac{z^{1,1}}{-d_0+z^{0,0}}\,\partial_\theta u^1 - \varepsilon^4\,\frac{z^{1,1}}{-d_0+z^{0,0}}\,\partial_\theta u^2.&\end{aligned}$$

So we define $\bar{\bar{u}} \in H^1(\Omega_s)$ approximating u^ε in Ω_s by

$$\begin{gathered}-\Delta\,\bar{\bar{u}} = f \quad \text{in } \Omega_s,\\ \bar{\bar{u}} + (\varepsilon\,d_0 - \varepsilon\,z^{0,0} - \varepsilon^2\,z^{1,0}\,\kappa)\,\partial_{\boldsymbol{n}}\,\bar{\bar{u}} - \varepsilon^2\,\frac{z^{1,1}}{-d_0+z^{0,0}}\,\partial_\theta\bar{\bar{u}} = 0 \quad \text{on } \Gamma,\\ \bar{\bar{u}} = 0 \quad \text{on } \partial\Omega_s\backslash\Gamma.\end{gathered} \tag{24}$$

Since Γ is a closed curve,

$$\int_\Gamma \bar{\bar{u}}\partial_\theta\bar{\bar{u}}\,d\theta = 0,$$

TABLE 3
Relative error convergence rates for a second order model.

quantity	$L^2(\Omega_s)$ error	$L^2(\Omega_s^{int})$ norm error
u	$O(\varepsilon^{3/2})$	$O(\varepsilon^3)$
∇u	$O(\varepsilon^{1/2})$	$O(\varepsilon^3)$

and for ε small enough the well posedness of (24) follows from Lemma 1 and the Lax–Milgram's lemma. To estimate the modeling error we first define $\bar{\bar{e}} = \bar{\bar{u}} - u^0 - \varepsilon\, u^1 - \varepsilon^2\, u^2$. Thus

$$
\begin{aligned}
&-\Delta\, \bar{\bar{e}} = 0 \quad \text{in } \Omega_s,\\
&\bar{\bar{e}} + (\varepsilon\, d_0 - \varepsilon\, z^{0,0} - \varepsilon^2\, z^{1,0}\kappa)\ \partial_{\boldsymbol{n}}\, \bar{\bar{e}} - \varepsilon^2\, \frac{z^{1,1}}{-d_0 + z^{0,0}}\, \partial_\theta \bar{\bar{e}} = -\varepsilon^3 z^{1,0}\, \kappa\ \partial_{\boldsymbol{n}}\, u^1\\
&-\varepsilon^3\, (-d_0 + z^{0,0} + \varepsilon\, z^{1,0}\, \kappa)\ \partial_{\boldsymbol{n}}\, u^2 - \varepsilon^3\, \frac{z^{1,1}}{-d_0 + z^{0,0}}\, \partial_\theta u^1 - \varepsilon^4\, \frac{z^{1,1}}{-d_0 + z^{0,0}}\, \partial_\theta u^2 \quad \text{on } \Gamma,\\
&\bar{\bar{e}} = 0 \quad \text{on } \partial\Omega_s \backslash \Gamma.
\end{aligned}
$$

Regularity estimates [16, Theorem 4.24] guarantee the existence of an ε-independent constant c such that

$$\|\bar{\bar{e}}\|_{H^1(\Omega_s)} \leq c\, \varepsilon^3.$$

Using the triangle inequality and (21), it is possible to estimate the $H^1(\Omega_s)$ norm modeling error,

$$\|u^\varepsilon - \bar{\bar{u}}\|_{H^1(\Omega_s)} \leq \|u^\varepsilon - u^0 - \varepsilon\, u^1 - \varepsilon^2\, u^2\|_{H^1(\Omega_s)} + \|\bar{\bar{u}} - u^0 - \varepsilon\, u^1 - \varepsilon^2\, u^2\|_{H^1(\Omega_s)} \leq c\, \varepsilon^{1/2}.$$

Analogously, using (22), (23), we obtain L^2 and interior estimates,

$$\|u^\varepsilon - \bar{\bar{u}}\|_{L^2(\Omega_s)} \leq c\, \varepsilon^{3/2} \qquad \|u^\varepsilon - \bar{\bar{u}}\|_{H^1(\Omega_s^{int})} \leq c\, \varepsilon^3.$$

These results are displayed in Table 3.

3.4. Summary: The proposed effective problems. The first order boundary value problem in Ω_s is the following: find $\bar{u} \in H^1(\Omega_s)$ such that

$$
\begin{aligned}
&-\Delta\, \bar{u} = f \quad \text{in } \Omega_s,\\
&\partial_{\boldsymbol{n}}\, \bar{u} = -\frac{1}{\varepsilon\, (d_0 - z^{0,0})}\, \bar{u} \quad \text{on } \Gamma, \qquad \bar{u} = 0 \quad \text{on } \partial\Omega_s \backslash \Gamma,
\end{aligned}
\tag{25}
$$

where $z^{0,0}$ is obtained from (5). For error estimates, see Table 2.

The second order boundary value problem in Ω_s is: find $\bar{\bar{u}} \in H^1(\Omega_s)$ such that

$$
\begin{aligned}
&-\Delta\, \bar{\bar{u}} = f \quad \text{in } \Omega_s,\\
&\partial_{\boldsymbol{n}}\, \bar{\bar{u}} = -C_1^\varepsilon\, \bar{\bar{u}} + C_2^\varepsilon\, \partial_\theta \bar{\bar{u}} \quad \text{on } \Gamma, \qquad \bar{\bar{u}} = 0 \quad \text{on } \partial\Omega_s \backslash \Gamma,
\end{aligned}
\tag{26}
$$

with

$$C_1^\varepsilon = \frac{1}{\varepsilon\, (d_0 - z^{0,0} - \varepsilon\, z^{1,0}\, \kappa)}, \quad C_2^\varepsilon = \frac{\varepsilon\, z^{1,1}}{(d_0 - z^{0,0} - \varepsilon\, z^{1,0}\, \kappa)(z^{0,0} - d_0)},$$

and where $z^{0,0}$ is computed from (5), and $z^{1,0}$, $z^{1,1}$ from (19), (20). For error estimates, see Table 3.

4. Asymptotic expansion definition. We now find and justify the terms presented previously. Consider a formal asymptotic expansion in the general form

$$u^\varepsilon \sim u^0 + \varepsilon\, u^1 + \varepsilon^2\, u^2 + \cdots + W^{BL}(\varepsilon) \quad \text{in } \Omega_s. \tag{27}$$

Here, $W^{BL}(\varepsilon)$ corresponds to the oscillatory part of the solution, which dies away exponentially fast with the distance to the boundary. Our procedure to find out the terms in the expansion uses a domain decomposition result that we state below.

It is convenient to introduce the jump function $[\![\cdot]\!]$ that assigns the *absolute value* of the jump over the interface Γ.

LEMMA 2. *Let Ω^ε, Ω_s, Ω_r^ε, and Γ be as above. Then there exists an ε-independent constant c such that*

$$\begin{aligned}\|e\|_{H^1(\Omega_r^\varepsilon)} + \|e\|_{H^1(\Omega_s)} \\ \leq c\left(\|\Delta e\|_{L^2(\Omega_r^\varepsilon)} + \|\Delta e\|_{L^2(\Omega_s)} + \|[\![e]\!]\|_{H^{1/2}(\Gamma)} + \|[\![\partial_{\boldsymbol{n}} e]\!]\|_{H^{-1/2}(\Gamma)}\right)\end{aligned} \tag{28}$$

whenever $e|_{\Omega_r^\varepsilon} \in H^1(\Omega_r^\varepsilon)$, $\Delta e|_{\Omega_r^\varepsilon} \in L^2(\Omega_r^\varepsilon)$, and $e|_{\Omega_s} \in H^1(\Omega_s)$, $\Delta e|_{\Omega_s} \in L^2(\Omega_s)$, with $e = 0$ on $\partial\Omega_r^\varepsilon \backslash \Gamma \cup \partial\Omega_s \backslash \Gamma$.

Proof. We first define

$$e^- = e|_{\Omega_r^\varepsilon}, \qquad e^+ = e|_{\Omega_s}.$$

It follows from Green's identity that

$$\begin{aligned}\int_{\Omega_r^\varepsilon} |\nabla e^-|^2\, d\mathbf{x} &= -\int_{\Omega_r^\varepsilon} e^-\, \Delta e^-\, d\mathbf{x} - < e^-, \partial_{\boldsymbol{n}} e^- >_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)}, \\ \int_{\Omega_s} |\nabla e^+|^2\, d\mathbf{x} &= -\int_{\Omega_s} e^+\, \Delta e^+\, d\mathbf{x} + < e^+, \partial_{\boldsymbol{n}} e^+ >_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)},\end{aligned}$$

where $< \cdot, \cdot >_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)}$ indicates the duality pairing between $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$. Combining both identities and then adding and subtracting the quantity $< e^-, \partial_{\boldsymbol{n}} e^+ >_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)}$, we gather that

$$\begin{aligned}|e^-|^2_{H^1(\Omega_r^\varepsilon)} + |e^+|^2_{H^1(\Omega_s)} = -\int_{\Omega_r^\varepsilon} e^-\, \Delta e^-\, d\mathbf{x} - \int_{\Omega_s} e^+\, \Delta e^+\, d\mathbf{x} \\ + < e^-, \partial_{\boldsymbol{n}} e^+ - \partial_{\boldsymbol{n}} e^- >_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)} + < e^+ - e^-, \partial_{\boldsymbol{n}} e^+ >_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)}.\end{aligned}$$

When estimating the above quantities, a delicate question is how the constants depend on the domains. For $u \in H^{1/2}(\Gamma)$ and $v \in H^{-1/2}(\Gamma)$, the inequality

$$< u, v >_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)} \leq \|u\|_{H^{1/2}(\Gamma)} \|v\|_{H^{-1/2}(\Gamma)}$$

comes by naturally by inducing the operator norm in $H^{-1/2}(\Gamma)$. Thus, with the aid of the Cauchy–Schwarz inequality it follows that

$$\begin{aligned}|e^-|^2_{H^1(\Omega_r^\varepsilon)} + |e^+|^2_{H^1(\Omega_s)} \leq \|e^-\|_{L^2(\Omega_r^\varepsilon)} \|\Delta e^-\|_{L^2(\Omega_r^\varepsilon)} + \|e^+\|_{L^2(\Omega_s)} \|\Delta e^+\|_{L^2(\Omega_s)} \\ + \|e^-\|_{H^{1/2}(\Gamma)} \|[\![\partial_{\boldsymbol{n}} e]\!]\|_{H^{-1/2}(\Gamma)} + \|[\![e]\!]\|_{H^{1/2}(\Gamma)} \|\partial_{\boldsymbol{n}} e^+\|_{H^{-1/2}(\Gamma)}.\end{aligned}$$

Next, let Ω' be the interior of Γ, i.e., the open domain circumvented by Γ. Hence $\Omega_r^\varepsilon \subset \Omega'$, and since the trace of e vanishes on Γ_r^ε, the extension (by zero) operator

$$P : \{v \in H^1(\Omega_r^\varepsilon) :\ v = 0 \text{ on } \Gamma_r^\varepsilon\} \to H^1(\Omega')$$

given by $Pv = v$ in Ω_r^ε and $Pv = 0$, otherwise, is an isometry [26, 19]. By construction, P preserves L^2 norms as well and we gather the trace and Poincaré inequalities,

$$\|v\|_{H^{1/2}(\Gamma)} \le c\|Pv\|_{H^1(\Omega')} = c\|v\|_{H^1(\Omega_r^\varepsilon)} \tag{29}$$

$$\|v\|_{L^2(\Omega_r^\varepsilon)} = \|Pv\|_{L^2(\Omega')} \le c|Pv|_{H^1(\Omega')} = c|v|_{H^1(\Omega_r^\varepsilon)} \tag{30}$$

for all $v \in H^1(\Omega_r^\varepsilon)$, where the constant c is independent of ε.

Using now the trace inequality $\|v\|_{H^{1/2}(\Gamma)} \le c\|v\|_{H^1(\Omega_s)}$ for all $v \in H^1(\Omega_s)$ and (29), it follows that

$$\begin{aligned} |e^-|^2_{H^1(\Omega_r^\varepsilon)} + |e^+|^2_{H^1(\Omega_s)} \le{}& (\|\Delta e^-\|_{L^2(\Omega_r^\varepsilon)} + \|\Delta e^+\|_{L^2(\Omega_s)})\|e\|_{L^2(\Omega^\varepsilon)} \\ &+ \|e^-\|_{H^1(\Omega_r^\varepsilon)}\|[\![\partial_{\boldsymbol{n}} e]\!]\|_{H^{-1/2}(\Gamma)} + c\|[\![e]\!]\|_{H^{1/2}(\Gamma)}\|e^+\|_{H^1(\Omega_s)}. \end{aligned} \tag{31}$$

To conclude the proof, it is enough to use in (31) the Poincaré inequality in Ω_r^ε given by (30) and also in Ω_s. □

We shall apply Lemma 2 repeatedly with e being the difference between u^ε and a truncated asymptotic expansion. Hence, to make such a difference as small as possible, we ought to minimize the L^2 norm of Δe in Ω_r^ε and Ω_s and control the jumps of both e and $\partial_{\boldsymbol{n}} e$ over Γ.

A natural choice for the first term of the asymptotic of u^ε is u^0 given by (3), plus the condition $u^0 = 0$ in $\overline{\Omega_r^\varepsilon}$. Applying Lemma 2 with $e = u^\varepsilon - u^0$, we see that the source of error is the normal derivative jump $[\![\partial_{\boldsymbol{n}} u^0]\!]$. We remedy this by adding $\varepsilon\,\zeta^1$ to the asymptotic, where

$$\zeta^1(\mathbf{x}) = \begin{cases} -\varepsilon^{-1}\rho\ \partial_{\boldsymbol{n}} u^0(\theta, d_0\,\varepsilon) & \text{in } \Omega_r^\varepsilon, \\ 0 & \text{in } \Omega_s. \end{cases}$$

The function ζ^1 defined as above satisfies the following properties:

1. $\zeta^1 \equiv 0$ outside Ω_r^ε,
2. $[\![\varepsilon\,\partial_{\boldsymbol{n}}\,\zeta^1]\!] \equiv |\,\partial_{\boldsymbol{n}}\,u^0|$ on Γ,
3. $\zeta^1 = \psi_r(\varepsilon^{-1}\theta)\ \partial_{\boldsymbol{n}}\,u^0(\theta, d_0\,\varepsilon)$ on Γ_r^ε.

So, in general, the correction of the jump of the normal derivative on Γ violates the zero Dirichlet condition at Γ_r^ε.

Proceeding with the computations, we have to add a boundary corrector to compensate for the value of ζ^1 on Γ_r^ε. This is nontrivial since a "typical" boundary corrector does not decay to zero; see Lemma 1 and (5). A similar, but actually simpler situation occurs for the asymptotics of plates [18]. Thus, we add a boundary corrector that is the sum of two functions and is given by $\varepsilon\left[W^1(\varepsilon) + \chi_r^\varepsilon Z^1(\varepsilon)\right]$. One part, corresponding to $W^1(\varepsilon)$, decays exponentially fast to zero with $\varepsilon^{-1}\rho$ and undulates with $\varepsilon^{-1}\theta$. The other part, corresponding to $Z^1(\varepsilon)$, depends only on θ and is nonzero only in Ω_r^ε. Hence,

$$-\Delta\, W^1(\varepsilon) = \chi_r^\varepsilon\,\Delta\left[\zeta^1 + Z^1(\varepsilon)\right] \quad \text{in } \Omega^\varepsilon, \tag{32}$$

$$W^1(\varepsilon) = -\zeta^1 - Z^1(\varepsilon) \quad \text{on } \Gamma_r^\varepsilon, \tag{33}$$

and the characteristic function of Ω_r^ε is given by χ_r^ε, where

$$\chi_r^\varepsilon(\rho) = \begin{cases} 1 & \text{if } \rho < \varepsilon d_0, \\ 0 & \text{otherwise.} \end{cases}$$

At first sight, finding $W^1(\varepsilon)$ and $Z^1(\varepsilon)$ satisfying (32), (33) seems (at least!) as hard as solving the original problem (1). Nevertheless, it is possible to make use of the periodicity of the wrinkles, and formally recast (32), (33) as a sequence of ε-independent problems which are easier to solve. We write

$$W^1(\varepsilon) \sim W^{1,0} + \varepsilon\, W^{1,1} + \varepsilon^2\, W^{1,2} + \cdots, \tag{34}$$

$$Z^1(\varepsilon) \sim Z^{1,0} + \varepsilon\, Z^{1,1} + \varepsilon^2\, Z^{1,2} + \cdots. \tag{35}$$

We shall impose on Γ_r^ε that $W^{1,0} = -\zeta^1 - Z^{1,0}$ and that $W^{1,j} = -Z^{1,j}$ for $j \neq 0$. We postpone the precise definition of these terms for now, but add, formally, the term $\varepsilon\, \chi_r^\varepsilon(\mathbf{x}) Z^{1,0}(\theta) + \varepsilon\, W^{1,0}(\theta, \varepsilon^{-1}\theta, \varepsilon^{-1}\rho)$ to the asymptotic. The remaining terms of the expansion for $W^1(\varepsilon)$, $Z^1(\varepsilon)$ shall be added as we continue to develop the expansion.

So far the asymptotic reads as

$$\begin{cases} \varepsilon\, \zeta^1 + \varepsilon\, W^{1,0} + \varepsilon\, Z^{1,0} & \text{in } \Omega_r^\varepsilon, \\ u^0 + \varepsilon\, W^{1,0} & \text{in } \Omega_s. \end{cases} \tag{36}$$

Note that now the normal derivative of the difference between u^ε and the expression in (36) has zero jump on Γ, but the difference itself has nontrivial jump equal to $-d_0\, \varepsilon\, \partial_{\boldsymbol{n}}\, u^0 + \varepsilon\, Z^{1,0}$ on Γ. Such error no longer depends on the fast variable and it can be corrected adding to the asymptotic expansion a new term that depends only on the slow variable.

We continue to define the terms of the expansion, this time trying to cancel out the error due to the jump of the expression in (36) on Γ. Consider u^1 the solution of

$$\begin{gathered} -\Delta\, u^1 = 0 \quad \text{in } \Omega_s, \\ u^1 = -d_0\, \partial_{\boldsymbol{n}}\, u^0 + Z^{1,0} \quad \text{on } \Gamma, \qquad u^1 = 0 \quad \text{on } \partial\Omega_s \backslash \Gamma, \qquad u^1 = 0 \quad \text{on } \Omega_r^\varepsilon. \end{gathered} \tag{37}$$

Remark 2. Although (37) looks different from (8), it is not. In fact, $Z^{1,0} = z^{0,0}\, \partial_{\boldsymbol{n}}\, u^0$, but that will become clear later.

Adding $\varepsilon\, u^1$ to the expansion corrects the previous error, but results in a jump in the normal derivative across Γ. Mimicking what we did before, we add $\varepsilon^2\, \zeta^2$ to the expansion, where

$$\zeta^2(\mathbf{x}) = -\varepsilon^{-1} \rho\, \chi_r^\varepsilon(\rho)\, \partial_{\boldsymbol{n}}\, u^1(\theta, d_0\varepsilon).$$

Ideally the next contribution would be $\varepsilon^2 \big[W^2(\varepsilon) + \chi_r^\varepsilon Z^2(\varepsilon)\big]$, where

$$-\Delta\, W^2(\varepsilon) = \chi_r^\varepsilon\, \Delta \big[\zeta^2 + Z^2(\varepsilon)\big] \quad \text{in } \Omega^\varepsilon, \tag{38}$$

$$W^2(\varepsilon) = -\zeta^2 - Z^2(\varepsilon) \quad \text{on } \Gamma_r^\varepsilon. \tag{39}$$

As in (34), (35),

$$W^2(\varepsilon) \sim W^{2,0} + \varepsilon\, W^{2,1} + \varepsilon^2\, W^{2,2} + \cdots, \tag{40}$$

$$Z^2(\varepsilon) \sim Z^{2,0} + \varepsilon\, Z^{2,1} + \varepsilon^2\, Z^{2,2} + \cdots. \tag{41}$$

On Γ_r^ε we shall have $W^{2,0} = -\zeta^2 - Z^{2,0}$, and $W^{2,j} = -Z^{2,j}$ for $j \neq 0$. Then, we simply add $\varepsilon^2 \big[W^{1,1} + W^{2,0} + \chi_r^\varepsilon \big(Z^{1,1} + Z^{2,0}\big)\big]$ to our asymptotic expansion. Note that terms in ε^2 corresponding to the expansions for $W^1(\varepsilon)$, $Z^1(\varepsilon)$ are included now.

At this point, the asymptotic reads as

$$(42)\quad \begin{cases} \varepsilon\,\zeta^1+\varepsilon^2\,\zeta^2+\varepsilon\,W^{1,0}+\varepsilon\,Z^{1,0}+\varepsilon^2\left(W^{1,1}+W^{2,0}\right)+\varepsilon^2\left(Z^{1,1}+Z^{2,0}\right) & \text{in } \Omega_r^\varepsilon,\\ u^0+\varepsilon\,u^1+\varepsilon\,W^{1,0}+\varepsilon^2\left(W^{1,1}+W^{2,0}\right) & \text{in } \Omega_s. \end{cases}$$

The expansion pattern should be clear by now, and the successive terms are defined in similar manner. In general, after the kth step, the asymptotic expansion reads as

$$(43)\quad \begin{cases} \zeta_{k,\varepsilon}+W_{k,\varepsilon}^{BL}+Z_{k,\varepsilon} & \text{in } \Omega_r^\varepsilon,\\ u_{k-1,\varepsilon}^{smooth}+W_{k,\varepsilon}^{BL} & \text{in } \Omega_s, \end{cases}$$

where $\zeta_{k,\varepsilon}(\theta,\rho)=\varepsilon\,\zeta^1+\cdots+\varepsilon^k\,\zeta^k$, and $\zeta^i=-\varepsilon^{-1}\rho\,\chi_r^\varepsilon\;\partial_{\boldsymbol{n}}\,u^{i-1}(\theta,d_0\varepsilon)$. Also,

$$W_{k,\varepsilon}^{BL}=\varepsilon\,W^{1,0}+\varepsilon^2\left(W^{1,1}+W^{2,0}\right)+\cdots+\varepsilon^k\left(W^{1,k-1}+W^{2,k-2}+\cdots+W^{k,0}\right),$$
$$Z_{k,\varepsilon}=\chi_r^\varepsilon\left[\varepsilon\,Z^{1,0}+\varepsilon^2\left(Z^{1,1}+Z^{2,0}\right)+\cdots+\varepsilon^k\left(Z^{1,k-1}+Z^{2,k-2}+\cdots+Z^{k,0}\right)\right].$$

Here, although we did not fully define these functions yet, $W^{i,j}$, $Z^{i,j}$ depend only on u^{i-1}, $W^{i,j-1},\cdots,W^{i,0}$ and $Z^{i,j-1},\cdots,Z^{i,0}$. Also, we shall have on Γ_r^ε that

$$W^{i,0}=-\zeta^i-Z^{i,0},\qquad W^{i,j}=-Z^{i,j}\quad\text{for } j\neq 0.$$

Finally, $u_{k-1,\varepsilon}^{smooth}=u^0+\varepsilon\,u^1+\cdots+\varepsilon^{k-1}\,u^{k-1}$, where u^0 is as in (3), and for i positive,

$$\begin{gathered} -\Delta\,u^i=0\quad\text{in } \Omega_s,\\ u^i=-d_0\;\partial_{\boldsymbol{n}}\,u^{i-1}+Z^{1,i-1}+Z^{2,i-2}+\cdots+Z^{i,0}\quad\text{on } \Gamma,\qquad u^i=0\quad\text{on } \partial\Omega_s\backslash\Gamma,\\ u^i=0\quad\text{in } \Omega_r^\varepsilon. \end{gathered}$$

5. The boundary corrector problem. We now analyze the boundary corrector problem, as (32), (33) and (38), (39), in more detail. The presence of the curvature makes this problem cumbersome and a lot of insight can be gained by studying the zero curvature case first; see [5] and references therein.

Consider the problem

$$(44)\quad -\Delta\,w(\varepsilon)=\chi_r^\varepsilon\;\Delta\left[-\varepsilon^{-1}\rho\,\phi(\theta)+z(\varepsilon)\right]\quad\text{in } \Omega^\varepsilon,$$
$$(45)\quad w(\varepsilon)=\varepsilon^{-1}\rho\,\phi(\theta)-z(\varepsilon)\quad\text{on } \Gamma_r^\varepsilon.$$

Here, ϕ is a given function of θ only. The function $z(\varepsilon)$ is unknown a priori, but it is introduced to guarantee that $w(\varepsilon)$ decays exponentially to zero with ρ. It is desirable to have $z(\varepsilon)$ as simple as possible, and it suffices to assume $z(\varepsilon)$ independent of ρ.

Although ϕ is not necessarily periodic, we try to make use of the periodicity of the wrinkles, and recast the corrector problem as a sequence of problems in periodic domains. Using the stretched coordinates $(\hat\theta,\hat\rho)=(\varepsilon^{-1}\theta,\varepsilon^{-1}\rho)$, we seek solutions that are product of functions in the stretched coordinates with functions of θ only. With this in mind, we write the Laplacian of a function in the form $v(\mathbf{x})=h(\hat\theta,\hat\rho)\,g(\theta)$ as

$$\begin{aligned}(46)\quad -\Delta\,v=&-\varepsilon^{-2}\left(\partial_{\hat\theta\hat\theta}h+\partial_{\hat\rho\hat\rho}h\right)g+\varepsilon^{-1}\left(\kappa\,\partial_{\hat\rho}\,h-2\,\kappa\,\hat\rho\,\partial_{\hat\theta\hat\theta}\,h\right)g-\varepsilon^{-1}2\,\partial_{\hat\theta}\,h\,g'\\ &-\sum_{j=0}^{\infty}\varepsilon^j\,\hat\rho^j\left[\left(\;\hat\rho\;\;a_1^{j+1}\,\partial_{\hat\rho}\,h+a_3^{j+1}\,\hat\rho\,\partial_{\hat\theta}\,h+a_2^{j+2}\,\hat\rho^2\,\partial_{\hat\theta\hat\theta}\,h\right)g+\left(a_3^j\,h+2\,a_2^{j+1}\,\hat\rho\,\partial_{\hat\theta}\,h\right)g'\right.\\ &\left.+\,a_2^j\,h\,g''\right],\end{aligned}$$

where

$$a_1^j = -[\kappa(\theta)]^{j+1}, \quad a_2^j = (j+1)\,[\kappa(\theta)]^j, \quad a_3^j = \frac{j(j+1)}{2}\,[\kappa(\theta)]^{j-1}\,\kappa'(\theta),$$

and we recall that κ is the curvature of Γ.

From (46), and using that $z(\varepsilon)$ is independent of ρ,

$$\begin{aligned} -\Delta[\varepsilon^{-1}\rho\,\phi(\theta)] &= \varepsilon^{-1}\kappa\,\phi - \sum_{j=0}^{\infty}\varepsilon^j\,\hat{\rho}^j\,\left[\hat{\rho}\,a_1^{j+1}\,\phi + \hat{\rho}(a_3^j\,\phi' + a_2^j\,\phi'')\right], \\ \Delta\,z(\varepsilon) &= \sum_{j=0}^{\infty}\varepsilon^j\,\hat{\rho}^j\,\left[a_3^j\,\partial_\theta z(\varepsilon) + a_2^j\,\partial_{\theta\theta}\,z(\varepsilon)\right]. \end{aligned} \tag{47}$$

Assuming the expansion

$$z(\varepsilon) \sim z^0 + \varepsilon\,z^1 + \varepsilon^2\,z^2 + \cdots \tag{48}$$

and using the formal identity

$$\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\varepsilon^{i+j}\,c_j\,d_i = \sum_{j=0}^{\infty}\sum_{k=0}^{j}\varepsilon^j\,c_k\,d_{j-k}, \tag{49}$$

we gather from (47), (48), (49) the identity

$$\begin{aligned} \Delta\left[-\varepsilon^{-1}\rho\,\phi + z(\varepsilon)\right] = \varepsilon^{-1}\,\kappa\,\phi + \sum_{j=0}^{\infty}\varepsilon^j\,\Big\{ &-\hat{\rho}^j\,\left[\hat{\rho}\,a_1^{j+1}\,\phi + \hat{\rho}\,(a_3^j\,\phi' + a_2^j\,\phi'')\right] \\ &+ \sum_{k=0}^{j}\hat{\rho}^k\,\left(a_3^k\,\partial_\theta z^{j-k} + a_2^k\,\partial_{\theta\theta}\,z^{j-k}\right)\Big\}. \end{aligned} \tag{50}$$

Assuming the expansion

$$w(\varepsilon) \sim w^0 + \varepsilon\,w^1 + \varepsilon^2\,w^2 + \cdots \tag{51}$$

and that $w^0(\theta,\rho) = w^{0,0}(\hat{\theta},\hat{\rho})\,\phi(\theta)$ and $z^0(\theta) = z^{0,0}\,\phi(\theta)$, where $z^{0,0}$ is a constant, we gather that (44), (45), (46), and (50) lead to (5).

It is clear that w^0, z^0 do not satisfy (44) exactly, but only the highest order (with power ε^{-2}) term. In fact, if ϕ is smooth, it follows from usual regularity estimates and a scaling argument that

$$\left\|\Delta\big(\Upsilon w^0\big) - \varepsilon^{-1}\chi_r^\varepsilon\;\Delta\big(\rho\,\phi - z^0\big)\right\|_{L^2(\Omega^\varepsilon)} \leq c\,\varepsilon^{-1/2}.$$

Here, as in page 1455, we need the cutoff function Υ, since w^0 is not well defined all over Ω^ε.

The remainder shall be corrected by the equations defining w^1, w^2, etc. Note also that defining w^0 as a product between $w^{0,0}$ and ϕ allows us to impose periodic boundary conditions in the PDE defining $w^{0,0}$. This trick reduces the original boundary corrector problem (44), (45) to a much easier to solve *sequence of cell problems*.

Continuing the procedure with the aid of (46), we set

$$w^1(\theta,\rho) = w^{1,0}(\hat{\theta},\hat{\rho})\,\kappa(\theta)\,\phi(\theta) + w^{1,1}(\hat{\theta},\hat{\rho})\,\phi'(\theta),$$

and $z^1(\theta) = z^{1,0}\,\kappa(\theta)\,\phi(\theta) + z^{1,1}\,\phi'(\theta)$, where $w^{1,0}$ and $w^{1,1} \in S(\Omega_r)$, and $z^{1,0}$ and $z^{1,1}$ are constants such that (19), (20) holds.

Now, $w^0 + \varepsilon\, w^1$ is a better approximate solution to (44) since,

$$\big\|\Delta\big[\Upsilon\big(w^0 + \varepsilon\, w^1\big)\big] - \varepsilon^{-1}\chi_r^\varepsilon\ \Delta\big(\rho\,\phi - z^0\big)\big\|_{L^2(\Omega^\varepsilon)} \le c\,\varepsilon^{1/2}.$$

It is easy to see that the right-hand sides of the equations become more involved as we proceed. The crucial point is to note that in the above cases, the equations do not involve the nonperiodic terms ϕ, κ or their derivatives.

Proceeding in a similar manner, we define w^2, w^3, etc., and

$$\big\|\Delta\big[\Upsilon\big(w^0 + \varepsilon\, w^1 + \cdots + \varepsilon^k\, w^k\big)\big] - \varepsilon^{-1}\chi_r^\varepsilon\ \Delta\big(\rho\,\phi - z^0\big)\big\|_{L^2(\Omega^\varepsilon)} \le c\,\varepsilon^{k-1/2}. \tag{52}$$

Finally, it follows from our computations that

$$w \sim w^{0,0}\,\phi + \varepsilon\,(w^{1,0}\,\kappa\,\phi + w^{1,1}\,\phi') + \varepsilon^2 \cdots, \tag{53}$$

$$z \sim z^{0,0}\,\phi + \varepsilon\,(z^{1,0}\,\kappa\,\phi + z^{1,1}\,\phi') + \varepsilon^2 \cdots. \tag{54}$$

In terms of the expansions for the boundary corrector for our original problem, see (32)–(35), we define

$$\begin{aligned}
W^{1,0}(\mathbf{x}) &= \Upsilon(\rho)\, w^{0,0}(\hat\theta,\hat\rho)\ \partial_{\boldsymbol{n}}\, u^0(\theta, d_0\,\varepsilon),\\
W^{1,1}(\mathbf{x}) &= \Upsilon(\rho)\,[w^{1,0}(\hat\theta,\hat\rho)\,\kappa(\theta)\ \partial_{\boldsymbol{n}}\, u^0(\theta, d_0\,\varepsilon) + w^{1,1}\,(\hat\theta,\hat\rho)\,\partial_\theta\,\partial_{\boldsymbol{n}}\, u^0(\theta, d_0\,\varepsilon)],\\
Z^{1,0}(\theta) &= z^{0,0}\ \partial_{\boldsymbol{n}}\, u^0(\theta, d_0\,\varepsilon),\\
Z^{1,1}(\theta) &= z^{1,0}\,\kappa(\theta)\ \partial_{\boldsymbol{n}}\, u^0(\theta, d_0\,\varepsilon) + z^{1,1}\,\partial_\theta\,\partial_{\boldsymbol{n}}\, u^0(\theta, d_0\,\varepsilon),
\end{aligned} \tag{55}$$

etc. Similarly, from (38)–(41), we define

$$\begin{aligned}
W^{2,0}(\mathbf{x}) &= \Upsilon(\rho)\, w^{0,0}(\hat\rho,\hat\theta)\ \partial_{\boldsymbol{n}}\, u^1(\theta, d_0\,\varepsilon),\\
W^{2,1}(\mathbf{x}) &= \Upsilon(\rho)\,[w^{1,0}(\hat\theta,\hat\rho)\,\kappa(\theta)\ \partial_{\boldsymbol{n}}\, u^1(\theta, d_0\,\varepsilon) + w^{1,1}\,(\hat\theta,\hat\rho)\,\partial_\theta\ \partial_{\boldsymbol{n}}\, u^1(\theta, d_0\,\varepsilon)],\\
Z^{2,0}(\theta) &= z^{0,0}\ \partial_{\boldsymbol{n}}\, u^1(\theta, d_0\,\varepsilon),\\
Z^{2,1}(\theta) &= z^{1,0}\,\kappa(\theta)\ \partial_{\boldsymbol{n}}\, u^1(\theta, d_0\,\varepsilon) + z^{1,1}\,\partial_\theta\,\partial_{\boldsymbol{n}}\, u^1(\theta, d_0\,\varepsilon),
\end{aligned}$$

and so on.

6. Convergence estimate. In this section we estimate the difference between a truncated asymptotic expansion and the exact solution. To bound such difference, some a priori estimates are necessary, thus the regularity of the terms in the expansion is worthy of consideration. The results below are based on standard regularity estimates [15].

The boundary layer terms $w^{i,j}$ solve Poisson problems of the form (4), and we assume first that Ω_r is a convex polygon. Then $w^{0,0} \in H^2(\Omega_r)$. For $i > 1$, it follows from (46) that the right-hand side of the Poisson problem for $w^{i,j}$ depends on a linear combination of $w^{k,l}$, $\partial_{\hat\rho} w^{k,l}$, $\partial_{\hat\theta} w^{k,l}$, $\partial_{\hat\theta\hat\theta} w^{k,l}$, where $k < i$. Thus, $w^{i,j} \in H^2(\Omega_r)$. Similarly, if Ω_r is a nonconvex polygon with largest angle equal to ω, then $w^{i,j} \in H^s(\Omega_r)$ for all $s < 1 + \pi/\omega$. Note that the regularity results above depend *only* on the geometry of Ω_r and not on f. On the other hand, the regularity of $W^{i,j}$ depends on f, since it also depends on the u^i; see, e.g., (55).

Concerning the regularity of u^i, we rely on smoothness of Ω_s to conclude from (3) that $\|u_0\|_{H^{m+2}(\Omega_s)} \le c\|f\|_{H^m(\Omega_s)}$ for all real m, where c depends only on Ω_s. Since

u^i is harmonic for i positive, its regularity is determined by its Dirichlet boundary condition on Γ. Using (46), (50), we gather that the boundary condition for u^i depends, among other more regular terms, on $\partial_\theta^j \, \partial_{\boldsymbol{n}} \, u^{i-j-1}$, for $j = 0, \ldots, i-1$. Thus, an induction argument leads to the existence of a constant c depending only on Ω_s and m such that $\|u^i\|_{H^{m+2-i}(\Omega_s)} \leq c\|f\|_{H^m(\Omega_s)}$, for all real m.

Standard scaling arguments lead to the result below.

LEMMA 3. *Assume that f is a smooth function with support in Ω_s. Then, for every integers i, j, there exists a constant c such that*

$$\|W^{i,j}\|_{L^2(\Omega^\varepsilon)} \leq c\,\varepsilon^{1/2}, \qquad \|W^{i,j}\|_{H^1(\Omega^\varepsilon)} \leq c\,\varepsilon^{-1/2},$$
$$\|Z^{i,j}\|_{H^1(\Omega_r^\varepsilon)} + \|\zeta^i\|_{H^1(\Omega_r^\varepsilon)} \leq c\,\varepsilon^{1/2}, \qquad \|u^i\|_{H^1(\Omega_s)} \leq c.$$

The constant c might depend on f, Ω_s, and Ω_r, but it is independent of ε.

Considering now the truncated expansion as in (43), we define the error

$$e_k = \begin{cases} u^\varepsilon - \zeta_{k,\varepsilon} - W_{k,\varepsilon}^{BL} - Z_{k,\varepsilon} & \text{in } \Omega_r^\varepsilon, \\ u^\varepsilon - u_{k-1,\varepsilon}^{smooth} - W_{k,\varepsilon}^{BL} & \text{in } \Omega_s. \end{cases} \tag{56}$$

Aiming to use Lemma 2, we first note that e_k vanishes on $\partial\Omega^\varepsilon$. Also, the jumps across Γ are such that

$$[\![e_k]\!] = 0, \qquad [\![\partial_{\boldsymbol{n}} \, e_k]\!] = \varepsilon^k |\, \partial_{\boldsymbol{n}} \, u^k|. \tag{57}$$

Estimating $\Delta \, e_k$ is nontrivial since e_k is not harmonic in general. Indeed,

$$-\Delta \, e_k = \begin{cases} \Delta \big[\zeta_{k,\varepsilon} + W_{k,\varepsilon}^{BL} + Z_{k,\varepsilon}\big] & \text{in } \Omega_r^\varepsilon, \\ \Delta \, W_{k,\varepsilon}^{BL} & \text{in } \Omega_s. \end{cases}$$

It follows from the construction of $W_{k,\varepsilon}^{BL}$ in section 5, (52), and Lemma 3, that

$$\|\, \Delta \, e_k\|_{L^2(\Omega^\varepsilon)} \leq c\,\varepsilon^{k-1/2}. \tag{58}$$

With the above estimates it is not hard to prove the following result, which shows the rate of convergence in ε of the asymptotic expansion.

THEOREM 4. *For any positive integer k there exists a constant c such that the difference between the truncated asymptotic expansion and the original solution measured in the original domain is bounded as follows:*

$$\|e_k\|_{H^1(\Omega_s)} + \|e_k\|_{H^1(\Omega_r^\varepsilon)} \leq c\,\varepsilon^{k+1/2}. \tag{59}$$

Proof. From Lemmas 2 and 3 and estimates (57), (58), we have that

$$\|e_k\|_{H^1(\Omega_s)} + \|e_k\|_{H^1(\Omega_r^\varepsilon)} \leq c\,\varepsilon^{k-1/2}. \tag{60}$$

Although the above estimate is not sharp, it is not hard to improve it. In fact,

$$\begin{aligned} \|e_k\|_{H^1(\Omega_s)} + \|e_k\|_{H^1(\Omega_r^\varepsilon)} \leq{}& \|e_{k+1}\|_{H^1(\Omega_s)} + \|e_{k+1}\|_{H^1(\Omega_r^\varepsilon)} \\ &+ \|e_{k+1} - e_k\|_{H^1(\Omega_s)} + \|e_{k+1} - e_k\|_{H^1(\Omega_r^\varepsilon)} \leq c\,\varepsilon^{k+1/2}, \end{aligned}$$

where we used (60) and Lemma 3 to obtain the last inequality. ∎

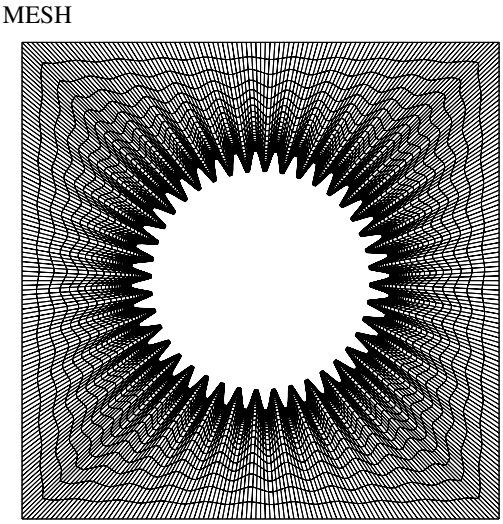


FIG. 3. *Mesh of rough domain.*

7. Numerical validation: Rough cylinder. We consider $\Omega_s \subset \mathbb{R}^2$ as the two-dimensional region having as outer boundary a square of size 4 and as inner boundary a circle of radius 1.15. Formally we have

$$\Omega_s = \{\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2 \,:\, |\mathbf{x}| > 1.15,\, |x_i| < 2,\, i = 1, 2\}.$$

We define the rough domain as having the same square as outer boundary and a "perturbed" circle as the inner boundary. We lay upon a circle of unitary radius 20 periodic wrinkles of height 0.1. Thus $d_0\varepsilon = 0.15$. The test considered is a variation of (1), where $f = 0$, and $u = 1$ at the outer boundary. We obtain an "exact" solution by fully discretizing the rough domain with a refined mesh shown in Figures 3 and 4. We remark that the polygonal appearance of the boundary in Figure 4 is deceiving and results from approximation a smooth domain using polygonal meshes. Equations (25) yield the first order solution, and (26) yield the second order solution. Figures 5 and 6 show the isolines and profiles of the solutions for the first and second order cell problems (5) and (19). Note that we plot only $w^{1,0}$, since $z^{1,1} = 1.0 \times 10^{-5}$ and can be disregarded in the computations. The computed effective constants are $z^{0,0} = 0.77$ and $z^{1,0} = 0.27$. In Figure 7 we plot the level curves for the exact solution and the second order approximation. In Figures 8, 9, and 10 we compare the profiles of the exact solution with the first and second order approximations, at different heights above the wrinkles. It is possible to see that the second order approximation yields the best results, as predicted by the theory.

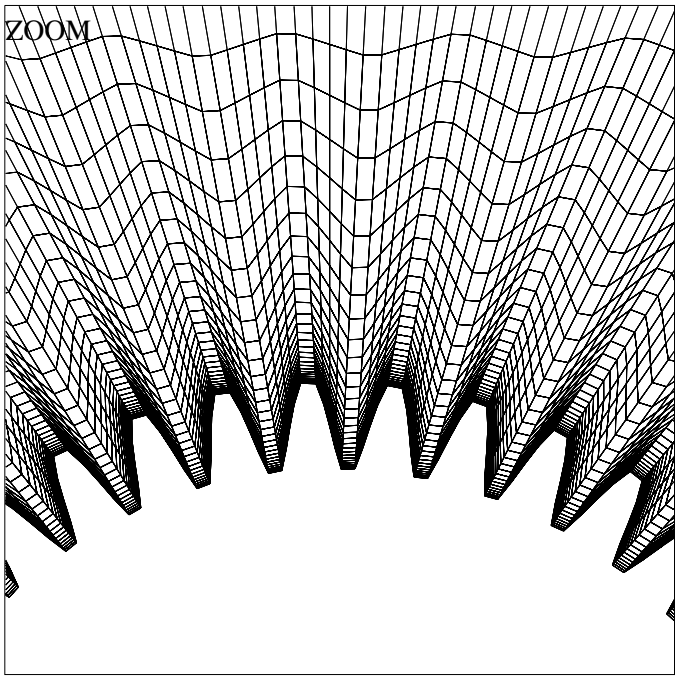


FIG. 4. *Zoom of the wrinkles.*

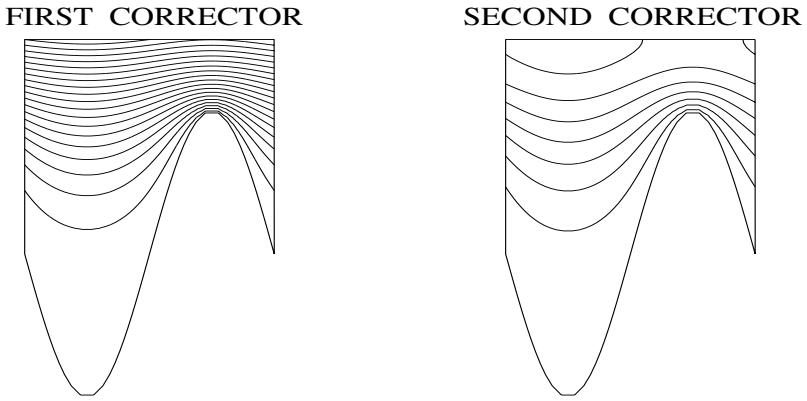


FIG. 5. *Isovalues of corrector* $\hat{\rho} - (w^{0,0} + z^{0,0})$ *(left) and* $w^{1,0} + z^{1,0}$ *(right) corresponding to the first and second order cell problems.*

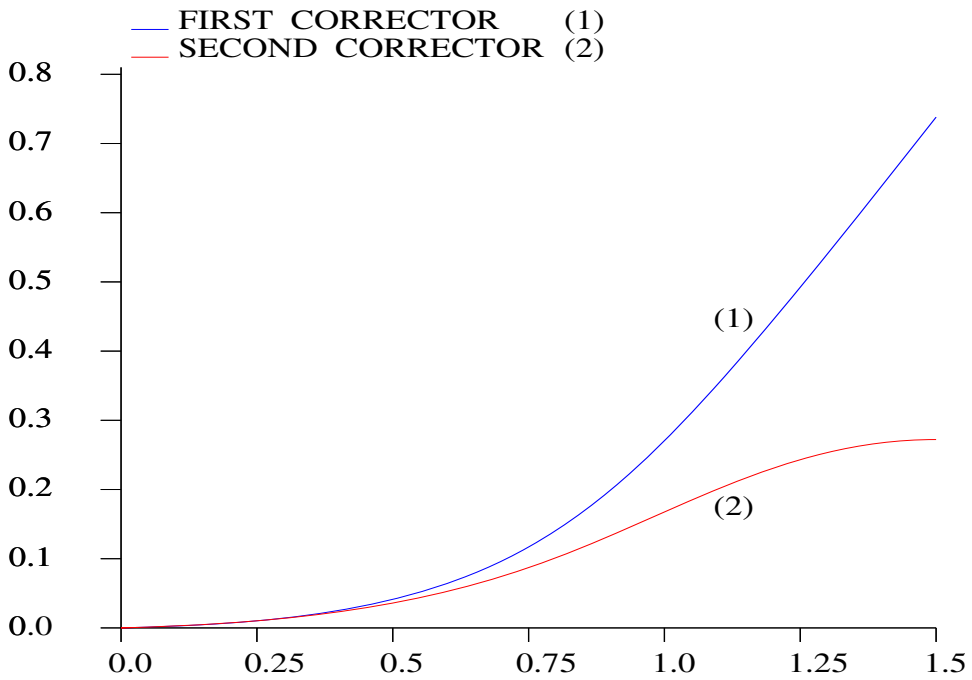


FIG. 6. *Profiles at* $\hat{\theta} = 0.78$ *of* $\hat{\rho} - (w^{0,0} + z^{0,0})$ *and* $w^{1,0} + z^{1,0}$.

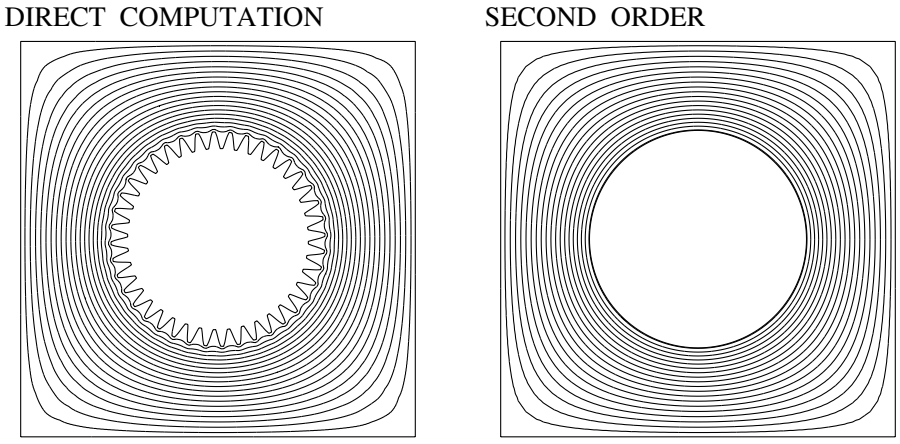


FIG. 7. *Second order approximation solves accurately the original problem.*

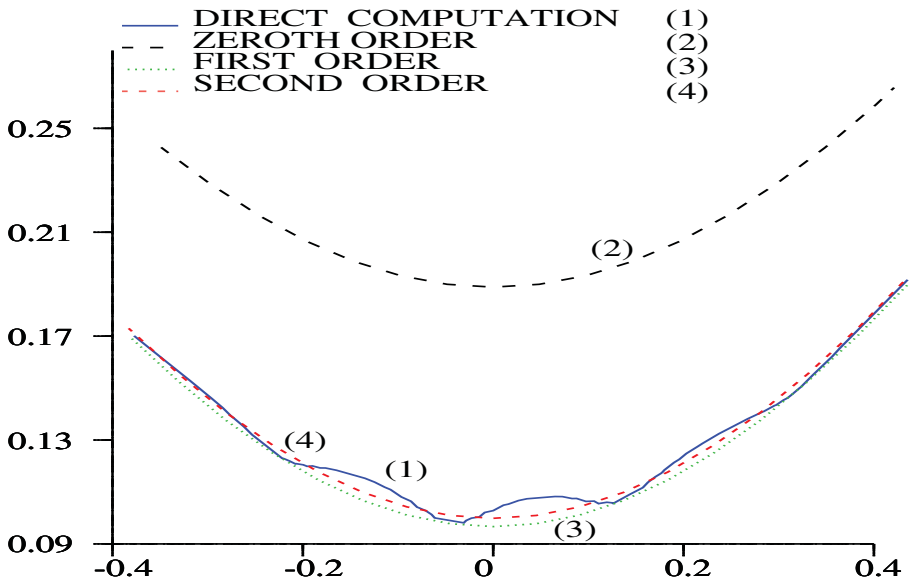


FIG. 8. *Profile of solutions at* $x_1 = 1.15$.

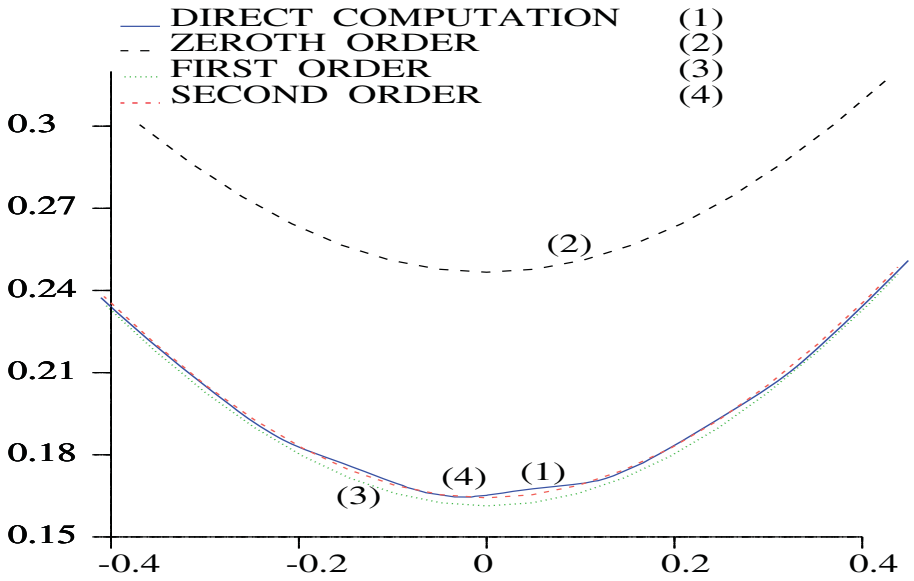


FIG. 9. *Profile of solutions at* $x_1 = 1.2$.

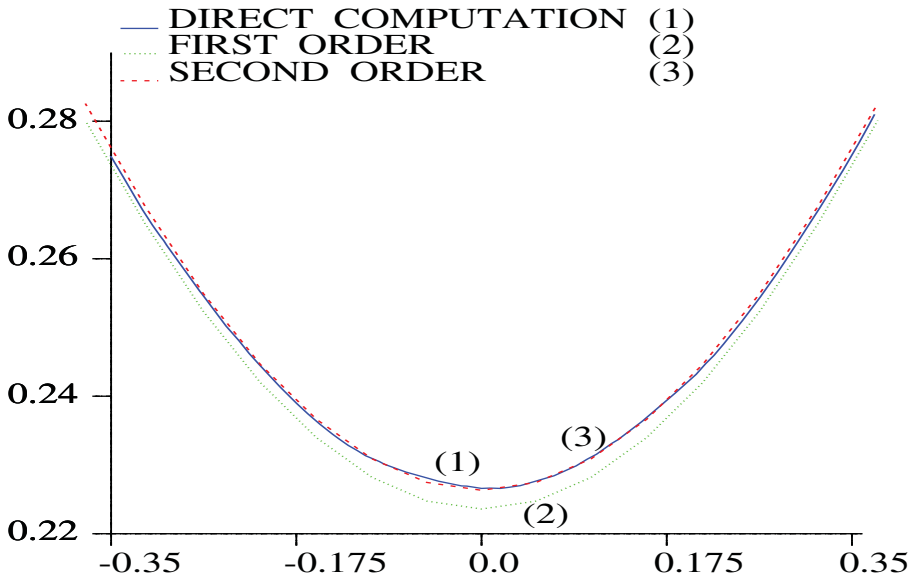


FIG. 10. *Profile of solutions at* $x_1 = 1.25$.

8. Conclusions. We investigated in this paper the problem of developing and estimating wall laws for problems defined in domains with rough and curved boundaries. For the sake of simplicity, the Poisson problem was considered. We developed a general methodology consisting of a two-scale expansion technique based on a domain decomposition result and obtained high order effective boundary conditions. Numerical tests accompanied the several sharp error estimates presented for first and second order approximations. In particular, this work proves that to obtain accurate

numerical results, the curvature must be considered.

Our approach can be carried over to more sophisticated operators and to higher dimensions, yielding then a general procedure to develop and estimate effective boundary conditions.

Acknowledgments. The authors thank the anonymous referees for several corrections and suggestions.

REFERENCES

[1] T. Abboud and H. Ammari, *Diffraction at a curved grating: TM and TE cases, homogenization*, J. Math. Anal. Appl., 202 (1996), pp. 995–1026.

[2] Y. Achdou, *Effet d'un mince revêtement métallisé sur la réflexion d'une onde électromagnetique*, C.R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 217–222.

[3] Y. Achdou and O. Pironneau, *Domain decomposition and wall laws*, C.R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 541–547.

[4] Y. Achdou, O. Pironneau, and F. Valentin, *Etude des Lois de Parois d'ordre* 1 *et* 2 *Pour des Domaines Rugueux par Décomposition de Domaine*, Technical report RR-3326, INRIA, Rocquencourt, France, 1997.

[5] Y. Achdou, O. Pironneau, and F. Valentin, *Effective boundary conditions for laminar flows over periodic rough boundaries*, J. Comput. Physics, 147 (1998), pp. 187–218.

[6] Y. Achdou, O. Pironneau, and F. Valentin, *Shape control versus boundary control*, in Équations aux Dérivées Partielles et Applications - Articles Dédiés à Jacques-Louis Lions, Elsevier, Paris, France, 1998, pp. 1–18.

[7] Y. Achdou, P. Le Tallec, F. Valentin, and O. Pironneau, *Constructing wall laws with domain decomposition or asymptotic expansion techniques*, Comput. Methods Appl. Mech. Engrg., 151 (1998), pp. 215–232.

[8] G. Allaire and M. Amar, *Boundary layer tails in periodic homogenization*, ESAIM Control Optim. Calc. Var., (1999), pp. 209–243.

[9] Y. Amirat and O. Bodart, *Numerical aproximation of laminar flows over rough walls with sharp asperities*, J. Comput. Appl. Math., 164 (2004), pp. 25–38.

[10] Y. Amirat, O. Bodart, U. De Maio, and A. Gaudiello, *Asymptotic approximation of the solution of the Laplace equation in a domain with highly oscillating boundary*, SIAM J. Math. Anal., 35 (2004), pp. 1598–1616.

[11] Y. Amirat, B. Climent, E. Fernandez-Cara, and J. Simon, *The Stokes equations with Fourier boundary conditions on a wall with asperities*, Math. Methods Appl. Sci., 24 (2001), pp. 255–276.

[12] D. N. Arnold and R. S. Falk, *Asymptotic analysis of the boundary layer for the Reissner–Mindlin plate model*, SIAM J. Math. Anal., 27 (1996), pp. 486–514.

[13] G. Barrenechea, P. Le Tallec, and F. Valentin, *New wall laws for the unsteady incompressible Navier-Stokes equations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 177–203.

[14] A. Bensoussan, J. Lions, and G. Papanicoulaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, New York, 1978.

[15] C. Bernardi and Y. Maday, *Spectral methods*, in Handbook of Numerical Analysis, Vol. 5, Handb. Numer. Anal., North-Holland, Amsterdam, 1997, pp. 209–485.

[16] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford University Press, New York, 1999.

[17] J. B. Conway, *Functions of One Complex Variable*, 2nd ed., Graduate Texts in Mathematics 11, Springer-Verlag, New York, 1978.

[18] M. Dauge and I. Gruais, *Asymptotics of arbitrary order for a thin elastic clamped plate.* I. *Optimal error estimates*, Asymptot. Anal., 13 (1996), pp. 167–197.

[19] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Functional and Variational Methods, with the collaboration of Michel Artola, Marc Authier, Philippe Bénilan, Michel Cessenat, Jean Michel Combes, Hélène Lanchon, Bertrand Mercier, Claude Wild, and Claude Zuily, translated from the French by Ian N. Sneddon. Springer-Verlag, Berlin, 1988.

[20] B. Engquist and J.-C. Nédélec, *Effective Boundary Conditions for Acoustic and Electromagnetic Scattering in Thin Layers*, Technical report, CMAP, Palaiseau, France, 1993.

[21] H. Haddar and P. Joly, *Effective boundary conditions for thin ferromagnetic coatings. Asymptotic analysis of the* 1*D model*, Asymptot. Anal., 27 (2001), pp. 127–160.
[22] S. F. Hoerner, *Résistance à L'avancement dans les Fluides*, Gauthier-Villars, Paris, 1965.
[23] A. M. Il′in, *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, Translations of Mathematical Monographs 102, translated from the Russian by V. Minachin [V. V. Minakhin], AMS, Providence, RI, 1992.
[24] W. Jäger and A. Mikelić, *On the roughness-induced effective boundary conditions for an incopmpressible viscous flow*, J. Differential Equations, 170 (2001), pp. 96–102.
[25] A. Madureira and F. Valentin, *Analysis of curvature influence on effective boundary conditions*, C.R. Math. Acad. Sci. Paris, 335 (2002), pp. 499–504.
[26] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson et Cie, Éditeurs, Paris, 1967.
[27] N. Neuss, M. Neuss-Radu, and A. Mikelić, *Effective laws for the Poisson equation on domains with curved oscillating boundaries*, Appl. Anal., 85 (2006), pp. 479–502.
[28] J. Nevard and J. B. Keller, *Homogenization of rough boundaries and interfaces*, SIAM J. Appl. Math., 57 (1997), pp. 1660–1686.
[29] O. A. Oleĭnik, A. S. Shamaev, and G. A. Yosifian, *Mathematical Problems in Elasticity and Homogenization*, Studies in Mathematics and its Applications 26, North-Holland, Amsterdam, 1992.
[30] E. Sanches-Palencia, *Ecoulement d'un Fluide à Travers une Paroi Perforée*, Technical report 57, Colection de la direction des études et recherches d'E.D.F., 1985.
[31] F. Valentin, *Nouvelles Conditions aux Limites Equivalentes pour des Interfaces Rugueuses en Mécanique des Fluides : Développement, Analyse et Mise en Ouvre Numérique*, Ph.D. thesis, Université Paris 6, Paris, 1998.
[32] T. Yamaguchi, *Computational mechanics simulation for clinical cardiovascular medicine*, in Proceedings of the ECCOMAS 2000, CIMNE, Barcelona, Spain, 2000.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1474–1488

OPTIMAL TRACING OF VISCOUS SHOCKS IN SOLUTIONS OF VISCOUS CONSERVATION LAWS*

WEN SHEN† AND MEE REA PARK†

Abstract. This paper contains a qualitative study of a scalar conservation law with viscosity: $u_t + f(u)_x = u_{xx}$. We consider the problem of identifying the location of viscous shocks, thus obtaining an optimal finite dimensional description of solutions to the viscous conservation law. We introduce a nonlinear functional whose minimizers yield the viscous traveling profiles which optimally fit the given solution. We prove that outside an initial time interval and away from times of shock interactions, our functional remains very small, i.e., the solution can be accurately represented by a finite number of viscous traveling waves.

Key words. optimal viscous shock tracing, viscous conservation laws, viscous traveling shocks, data compression, finite dimensional representation

AMS subject classifications. 35L60, 35L65, 35L67

DOI. 10.1137/050642642

1. Introduction.

Consider a scalar conservation law with viscosity

$$u_t + f(u)_x = u_{xx}\,. \tag{1.1}$$

We assume that the flux f is smooth and genuinely nonlinear, so that $f''(u) \geq \kappa > 0$ for every u. Our main interest here is how to identify the emergence of viscous shocks in a solution, and how to optimally trace their locations and strengths.

More generally, one may ask the following question. Assume that a particular solution $u = u(t,x)$ has already been computed. If we are allowed only a finite number of parameters to describe its most relevant features, what is the best way to compress the information? In the literature, the problem of finite dimensional approximation of a dynamical system has been studied mainly by looking at ω-limit sets [T]. Several results, valid for evolution equations of parabolic type, provide estimates on the dimension of an attractor. Of course, this yields a bound on the number of parameters needed to describe the evolution of the system asymptotically as $t \to +\infty$.

In the present paper, the focus is different. Namely, we seek a finite dimensional description which is accurate not only in the asymptotic limit as $t \to +\infty$ but also in the transient regime. For solutions to a scalar, viscous conservation law, this transient behavior is actually the most interesting feature that can be observed. On the other hand, at least in the case of convex flux, the ω-limit set is rather trivial. The asymptotic limit of any solution $t \mapsto u(t,\cdot)$ can be described in terms of the solution of a Riemann problem, i.e., either a single rarefaction or a viscous shock wave. For general theory on hyperbolic conservation laws, we refer to the books [Sm, B, S].

The problem of optimal location of viscous shock profiles was mentioned also in [W]. In this connection, we introduce a scalar functional whose minimizers identify the strengths and locations of viscous shock profiles present in the solution. We also

*Received by the editors October 13, 2005; accepted for publication (in revised form) August 1, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sima/38-5/64264.html
†Department of Mathematics, Penn State University, University Park, State College, PA 16802 (shen_w@math.psu.edu, park_m@math.psu.edu).

prove that outside a set of times with finite measure, at all other times our functional has very small values. In other words, the description of the solution profile $u(t,\cdot)$ in terms of finitely many viscous shocks is accurate, for most times t. The exceptional set consists of an initial time interval and times at which shock interactions occur; see Figure 1.

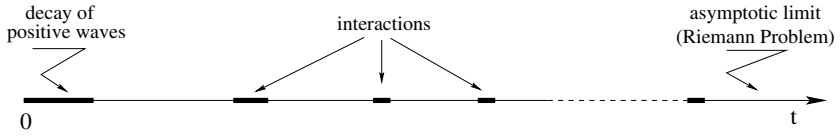


FIG. 1. *The exceptional set of times where the finite dimensional representation is not accurate.*

2. The main result. We consider here the single conservation law with viscosity

$$u_t + f(u)_x = u_{xx}\,. \tag{2.1}$$

We fix $M > 0$ and let $\mathcal{F}_M$ denote the set of all solutions to the Cauchy problem for (2.1) with initial data

$$u(0,x) = \bar u(x) \tag{2.2}$$

satisfying

$$\text{Tot.Var.}\{\bar u\} \leq M\,, \qquad\qquad \|\bar u\|_{\mathbf{L}^\infty} \leq M\,. \tag{2.3}$$

We shall assume that the flux f is $\mathcal{C}^2$ and strictly convex, so that $f''(u) > 0$ for all $u \in \mathbb{R}$. In particular, this implies that there exist constants κ, κ',

$$0 < \kappa \leq f''(u) \leq \kappa' \qquad\qquad \text{for all } u \in [-M, M]\,. \tag{2.4}$$

In essence, what we want to show is the following. Apart from a small set of times $J \subset [0,\infty[$, the profile $u(t,\cdot)$ of any solution of (2.1) can be accurately described in terms of the superposition of finitely many traveling viscous shocks. Indeed, the assumption (2.4) of genuine nonlinearity implies that all rarefaction waves will decay within an initial time interval. Moreover, in regions where the gradient u_x is large and negative, viscous shock profiles will form. These can travel for a long time without much changing their shape, except when they interact with each other. The set J of exceptional times where our description is not accurate will thus include an initial time interval and also the intervals where wave interactions occur. Much of the following analysis aims at making rigorous the above claims.

For every $u^- > u^+$ and $y \in \mathbb{R}$, let $\omega^{(u^\pm, y)}$ be the unique viscous shock profile joining the states u^-, u^+, centered at y. This profile can be found as the unique solution to the ODE

$$\omega' = f(\omega) - \sigma\,\omega - \big[f(u^-) - \sigma u^-\big]\,, \qquad\qquad \sigma = \frac{f(u^-) - f(u^+)}{u^- - u^+}\,, \tag{2.5}$$

satisfying the additional conditions

$$\omega''(y) = 0\,, \qquad\qquad \omega(-\infty) = u^-, \qquad \omega(+\infty) = u^+. \tag{2.6}$$

Notice that the last two identities in (2.6) follow from (2.5) and the convexity of f. Given any solution $u \in \mathcal{F}_M$ of the conservation law, for each $t > 0$ we introduce a description based on optimal location of shock profiles. Fix an integer $N \geq 1$ and let $\omega_i = \omega^{(u_i^{\pm}, y_i)}$ be the ith viscous shock profile we try to fit in. We consider the functional

$$\mathcal{J}\big(u(t),\ \omega_1, \ldots, \omega_N\big) \doteq \sum_{i=1}^{N} \int_{\mathbb{R}} \Big|u(t,x) - \omega_i(x)\Big| \cdot \big|\omega_{i,x}(x)\big|^2 \, dx + \int_{\mathbb{R}} \left|u_x(t,x) - \sum_{i=1}^{N} \omega_{i,x}(x)\right|^2 dx\,. \tag{2.7}$$

Notice that the first integral measures the distance between u and the traveling viscous shock ω_i, multiplied by a weight function $|\omega_{i,x}|^2$ which is vanishingly small away from the center of the ith shock. The second integral measures how well the derivative u_x is approximated by derivatives of traveling shock profiles. See Figure 2 for an illustration of fitting two viscous shocks in a solution.

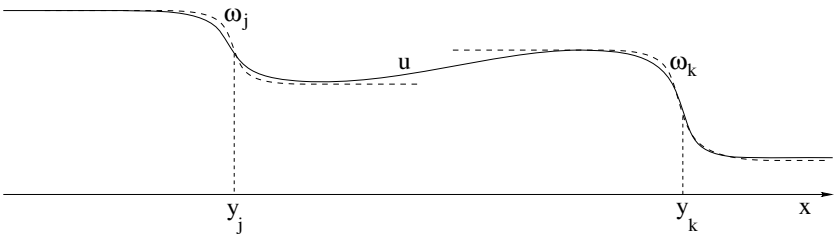


FIG. 2. *Fitting two viscous shocks ω_j, ω_k in a solution.*

If we fix a priori the complexity of our description, i.e., the integer N, how small can we render the integral $\mathcal{J}$? This problem can be formulated as

$$\inf_{\omega_1, \ldots, \omega_N} \mathcal{J}\big(u(t), \omega_1, \ldots, \omega_N\big)\,, \tag{2.8}$$

where the infimum is taken over all N-tuples of traveling shock profiles $\omega_i = \omega^{(u_i^{\pm}, y_i)}$, for some states $u_i^- > u_i^+$ and $y_i \in \mathbb{R}$. Notice that if we choose $\omega_i \equiv 0$ for $i = 1, \ldots, N$ (i.e., all traveling waves of zero amplitude), then the first integral in (2.7) vanishes because trivially $\omega_{i,x} \equiv 0$. However, in this case the second integral equals $\big\|u_x(t,\cdot)\big\|^2_{\mathbf{L}^2}$, which is of order Tot.Var.$(u)^3$ due to regularization and can be large.

To estimate the quantity in (2.8), an intuitive argument goes as follows. Set $\delta = M/N$, where M is given in (2.3). Since the total variation of $u(t)$ is bounded by M, there can be at most N shock profiles of strength $\geq \delta$. Each one of these can be traced accurately. In addition, there may be an arbitrarily large number of smaller shocks, say, of strengths σ_j, $j \geq 1$, with

$$\sigma_j \leq \delta\,, \qquad \sum_j \sigma_j \leq M\,. \tag{2.9}$$

Each shock which is not traced produces an error in the second integral of (2.7) of the order

$$\int \big|\omega_{j,x}(x)\big|^2 \, dx = \mathcal{O}(1) \cdot \sigma_j^3\,. \tag{2.10}$$

Because of (2.9) we thus expect that the minimum of $\mathcal{J}$ is approximately

$$\mathcal{J}_{\min} \approx \mathcal{O}(1) \cdot M\delta^2 = \mathcal{O}(1) \cdot \frac{M^3}{N^2}\,. \tag{2.11}$$

The estimate (2.11) should indeed hold outside an initial time interval, where positive waves will decay, and away from interaction times. Our main results are as follows.

THEOREM 2.1. *Assume $f''(u) \geq \kappa > 0$ for every $u \in \mathbb{R}$. Let $u \in \mathcal{F}_M$ be a solution of the viscous conservation law* (2.1)*, and fix $N \geq 1$. Then, for every $t > 0$, the minimization problem* (2.8) *has at least one solution.*

THEOREM 2.2. *There exist constants α (uniformly valid for all $N \geq 1$ and $u \in \mathcal{F}_M$) and $\beta = \beta_{N,M}$ (depending only on N and M) such that*

$$\mathcal{J}_{\min}\big(u(t)\big) \leq \alpha \cdot \frac{1}{N^2} \tag{2.12}$$

for all $t \in [0, \infty[\,\setminus I^u$ for an exceptional set I^u of times, with meas$(I^u) \leq \beta$.

The remainder of the paper contains a proof of the above two theorems. We remark that Theorem 2.1 states the existence of a minimizer for the scalar function $\mathcal{J} : \mathbb{R}^{3N} \mapsto \mathbb{R}$. Since $\mathcal{J}$ is continuous and positive, the result would be trivial if $\mathcal{J}(y) \to \infty$ as $|y| \to \infty$. However, it is easily seen that this coercivity condition fails. The heart of the proof consists in showing that, if $\{X^{(m)}\}_{m\geq 1}$ is a minimizing sequence with $|X^{(m)}| \to \infty$, then a second minimizing sequence $\tilde{X}^{(m)}$ can be defined (in terms of $X^{(m)}$) whose elements remain uniformly bounded.

The proof of Theorem 2.2 involves a deeper argument. With a solution of the viscous equation (2.1) we associate a curve γ moving in the plane. By results in [BB, BB1, BB2], the total area swept by this curve in its motion is a priori bounded in terms of a monotone decreasing area functional $Q(u)$. We then show that at every time t where the rate of decrease $\frac{d}{dt}Q(u(t))$ is sufficiently small, the inequality (2.12) holds.

We remark that in (2.8), the integer N is fixed. Of course, one could let N vary and look at the minimization problem

$$\min_{N\geq 0} \inf_{\omega_1,\ldots,\omega_N} \Big\{\epsilon\, N + \mathcal{J}\big(u(t),\, \omega_1, \ldots, \omega_N\big)\Big\}\,. \tag{2.13}$$

Here the first term penalizes the complexity of the description, adding a cost for each new viscous profile. The small constant $\epsilon > 0$ acts as a threshold parameter. Small viscous shock waves, whose strength $\|\omega_x\|^2_{\mathbf{L}^2}$ is of order $< \epsilon$, will not be traced. From Theorem 2.1 it immediately follows that the problem (2.13) also admits a global minimizer. This can be interpreted as an optimal description of the solution profile $u(t, \cdot)$ as superposition of traveling viscous shocks.

3. Proof of Theorem 2.1.

Step 1. At any fixed time $t > 0$, the solution $u(t, \cdot)$ of the viscous conservation law (2.1) is a C^1 function with bounded total variation. We shall prove, more generally, that the functional $\mathcal{J}(u;\, \omega_1, \ldots, \omega_N)$ admits a global minimum for every $\mathcal{C}^1$ function $u : \mathbb{R} \mapsto \mathbb{R}$ with bounded variation.

Step 2. Recall that $\omega_k \doteq \omega^{(u_k^{\pm}, y_k)}$. Observing that the traveling wave profiles ω_k as well as their derivatives $\omega_{k,x}$ depend continuously on the scalar parameters $u_k^-,\ u_k^+,\ y_k$ $(k = 1, \ldots, N)$, we have to prove that the continuous scalar function $\mathcal{J}(u; \cdot) : \mathbb{R}^{3N} \mapsto \mathbb{R}$ admits a global minimum.

Since $\mathcal{J} \geq 0$, this function has a nonnegative infimum $\mathcal{J}^{min}$. We can thus construct a minimizing sequence in $\mathbb{R}^{3N}$, converging to J^{min}, say,

$$\left\{ X^{(m)} \doteq \left(y_1^{(m)},\, u_1^{(m)+},\, u_1^{(m)-}, \ldots,\, y_N^{(m)},\, u_N^{(m)+},\, u_N^{(m)-} \right);\ \ m \geq 1 \right\}. \tag{3.1}$$

By possibly taking a subsequence, we can assume that each component of the vector $X^{(m)} \in \mathbb{R}^{3N}$ either converges to a finite limit or else diverges to $\pm\infty$.

Step 3. If

$$\sup_{m\geq 1} \left\{ \left| y_k^{(m)} \right| + \left| u_k^{(m)+} \right| + \left| u_k^{(m)-} \right| \right\} < \infty$$

for each $k = 1, \ldots, N$, then the entire minimizing sequence $\{X^{(m)}\}_{m\geq 1}$ is bounded in $\mathbb{R}^{3N}$. By our previous assumption, it converges to some limit

$$\bar{X} = \left(\bar{y}_1,\, \bar{u}_1^+,\, \bar{u}_1^-, \ldots,\, \bar{y}_N,\, \bar{u}_N^+,\, \bar{u}_N^- \right).$$

By continuity, we thus have $\mathcal{J}(u;\, \bar{X}) = \mathcal{J}^{min}$, proving the existence of a minimizer.

Step 4. In general, however, one cannot guarantee the minimizing sequence to be bounded, because the function $\mathcal{J}(u\,;\,\cdot)$ is not coercive on $\mathbb{R}^{3N}$. We shall thus adopt an alternative strategy. Assume that for some index j,

$$\lim_{m\to\infty} \left\{ \left| y_j^{(m)} \right| + \left| u_j^{(m)+} \right| + \left| u_j^{(m)-} \right| \right\} = \infty\,.$$

Consider the new sequence

$$\tilde{X}^{(m)} \doteq \left(\tilde{y}_1^{(m)},\, \tilde{u}_1^{(m)+},\, \tilde{u}_1^{(m)-}, \ldots,\, \tilde{y}_N^{(m)},\, \tilde{u}_N^{(m)+},\, \tilde{u}_N^{(m)-} \right),$$

obtained by setting the parameters of the jth traveling profile to zero. More precisely, for every $m \geq 1$ we set

$$\begin{aligned} &\left(\tilde{y}_i^{(m)},\, \tilde{u}_i^{(m)+},\, \tilde{u}_i^{(m)-} \right) = \left(y_i^{(m)},\, u_i^{(m)+},\, u_i^{(m)-} \right) \qquad \text{if } \ i \neq j\,, \\ &\left(\tilde{y}_j^{(m)},\, \tilde{u}_j^{(m)+},\, \tilde{u}_j^{(m)-} \right) = (0,0,0)\,. \end{aligned}$$

We claim that

$$\limsup_{m\to\infty} \mathcal{J}(u;\, \tilde{X}^{(m)}) \leq \lim_{m\to\infty} \mathcal{J}(u;\, X^{(m)})\,. \tag{3.2}$$

If the original sequence had k unbounded components, say, for $j \in \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, N\}$, the above construction yields a new minimizing sequence having $k-1$ unbounded components. By induction, in a finite number of steps we obtain a minimizing sequence where all components are bounded. Hence, by Step 3, a global minimizer exists.

Step 5. It now remains to show that (3.2) holds. Equivalently, for every $\varepsilon > 0$ we will prove that

$$\limsup_{m\to\infty} \mathcal{J}(u;\, \tilde{X}^{(m)}) \leq \lim_{m\to\infty} \mathcal{J}(u;\, X^{(m)}) + \varepsilon\,. \tag{3.3}$$

We shall consider different cases.

Case 1. Assume that, as $m \to \infty$,

$$\|\omega_{j,x}^{(m)}\|_{\mathbf{L}^2} \to 0\,. \tag{3.4}$$

By the assumption $f''(u) \geq \kappa > 0$, the strict convexity of the flux function implies

$$\|\omega_{j,x}^{(m)}\|_{\mathbf{L}^\infty} \to 0\,. \tag{3.5}$$

In this case, observing that $\omega_{i,x}^{(m)} \leq 0$ because all viscous shock profiles are decreasing, we have the estimate

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left| u_x - \sum_{i=1}^{N} \omega_{i,x}^{(m)} \right|^2 dx \\
&\qquad \geq \int_{-\infty}^{\infty} \left| u_x - \sum_{i \neq j} \omega_{i,x}^{(m)} \right|^2 dx - 2 \int_{-\infty}^{\infty} |u_x| \cdot \left| \omega_{j,x}^{(m)} \right| dx \\
&\qquad \geq \int_{-\infty}^{\infty} \left| u_x - \sum_{i \neq j} \omega_{i,x}^{(m)} \right|^2 dx - 2 \cdot \|u_x\|_{\mathbf{L}^1} \cdot \left\| \omega_{j,x}^{(m)} \right\|_{\mathbf{L}^\infty}.
\end{aligned}$$

Here we used the elementary inequality $(a-b-c)^2 \geq (a-b)^2 - 2|ac|$, valid whenever b and c have the same sign. Therefore,

$$\lim_{m\to\infty} \int_{-\infty}^{\infty} \left| u_x - \sum_{i=1}^{N} \omega_{i,x}^{(m)} \right|^2 dx \;\geq\; \limsup_{m\to\infty} \int_{-\infty}^{\infty} \left| u_x - \sum_{i \neq j} \omega_{i,x}^{(m)} \right|^2 dx\,, \tag{3.6}$$

provided that (3.4) holds. Clearly, (3.6) implies (3.2). Notice that the condition (3.4) is certainly satisfied if $u^{(m)+}$ and $u^{(m)-}$ remain uniformly bounded and $|u^{(m)+} - u^{(m)-}| \to 0$.

Case 2. Assume that

$$\liminf_{m\to\infty} \left\| \omega_{j,x}^{(m)} \right\|_{\mathbf{L}^2} \doteq \delta_2 > 0\,. \tag{3.7}$$

This breaks down into three different subcases.

Case 2a. We have the limits $|y_j^{(m)}| \to \infty$ while $u_j^{(m)+} \to u_j^+$, $u_j^{(m)-} \to u_j^-$. To fix the ideas, assume $y_j^{(m)} \to +\infty$. Observe that in this case $\delta_2 = \|\omega_{j,x}\|_{\mathbf{L}^2}$, where ω_j is a viscous shock profile connecting u_j^- with u_j^+. Given $\varepsilon > 0$, choose L so large that

$$\left(\int_L^{\infty} |u_x|^2 \, dx \right)^{1/2} \;<\; \frac{\varepsilon}{2\,\delta_2}\,.$$

We then have the estimate

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left| u_x - \sum_{i=1}^{N} \omega_{i,x}^{(m)} \right|^2 dx \\
&\quad \geq \int_{-\infty}^{\infty} \left| u_x - \sum_{i\neq j} \omega_{i,x}^{(m)} \right|^2 dx - 2\int_{-\infty}^{L} |u_x| \cdot \left|\omega_{j,x}^{(m)}\right| \, dx - 2\int_{L}^{\infty} |u_x| \cdot \left|\omega_{j,x}^{(m)}\right| \, dx \\
&\quad \geq \int_{-\infty}^{\infty} \left| u_x - \sum_{i\neq j} \omega_{i,x}^{(m)} \right|^2 dx - 2\left(\int_{-\infty}^{L} |u_x| \, dx\right) \cdot \sup_{x<L} \left|\omega_{j,x}^{(m)}(x)\right| \\
&\qquad\qquad - 2\left(\int_{L}^{\infty} |u_x|^2 \, dx\right)^{1/2} \cdot \left\|\omega_{j,x}^{(m)}\right\|_{\mathbf{L}^2} .
\end{aligned}$$

Observing that

$$\lim_{m\to\infty} \sup_{x<L} \left|\omega_{j,x}^{(m)}(x)\right| = 0\,, \qquad\qquad \lim_{m\to\infty} \left\|\omega_{j,x}^{(m)}\right\|_{\mathbf{L}^2} = \delta_2$$

from the above estimate we deduce

$$\liminf_{m\to\infty} \int_{-\infty}^{\infty} \left| u_x - \sum_{i=1}^{N} \omega_{i,x}^{(m)} \right|^2 dx \;\geq\; \liminf_{m\to\infty} \int_{-\infty}^{\infty} \left| u_x - \sum_{i\neq j} \omega_{i,x}^{(m)} \right|^2 dx - 2\frac{\varepsilon}{2\delta_2}\delta_2\,.$$

This clearly implies (3.3).

Case 2b. Assume that both sequences $u_j^{(m)+}$ and $u_j^{(m)-}$ diverge to $+\infty$. The case where they both tend to $-\infty$ is entirely similar. We then have

$$\begin{aligned}
&\liminf_{m\to\infty} \int_{-\infty}^{\infty} \left|u - \omega_j^{(m)}\right| \cdot \left|\omega_{j,x}^{(m)}\right|^2 dx \\
&\qquad \geq \lim_{m\to\infty} \left(\inf_{x\in\mathbb{R}} \left|u(x) - \omega_j^{(m)}(x)\right| \right) \cdot \left\|\omega_{j,x}^{(m)}\right\|_{\mathbf{L}^2}^2 \;=\; \infty\,.
\end{aligned}$$

Hence the original sequence was not minimizing. This contradiction shows that this case cannot happen.

Case 2c. Assume that the strength of the jth traveling wave becomes arbitrarily large as $m\to\infty$, so that $u_j^{(m)-} - u_j^{(m)+} \to \infty$. In this case, it is easy to check that $\mathcal{J}(u\,;\, X^{(m)}) \to \infty$. Indeed, let $K \doteq \|u\|_{\mathbf{L}^\infty}$. We then have

$$\liminf_{m\to\infty} \int_{-\infty}^{\infty} \left|u - \omega_j^{(m)}\right| \cdot \left|\omega_{j,x}^{(m)}\right|^2 dx \geq \liminf_{m\to\infty} \int_{\left|\omega_j^{(m)}(x)\right|>K+1} \left|\omega_{j,x}^{(m)}\right|^2 dx\,.$$

Obviously, this integral diverges to infinity. Indeed, let's consider the case when

$$\lim_{m\to\infty} u^{(m)-} = +\infty\,.$$

For the case when $\lim_{m\to\infty} u^{(m)+} = -\infty$ it is entirely similar. We have

$$\begin{aligned}
&\liminf_{m\to\infty} \int_{\left|\omega_j^{(m)}(x)\right|>K+1} \left|\omega_{j,x}^{(m)}\right|^2 dx \;\geq\; \liminf_{m\to\infty} \int_{\omega_j^{(m)}(x)>K+1} \left|\omega_{j,x}^{(m)}\right|^2 dx \\
&\qquad \geq \liminf_{m\to\infty} \min_{\omega_j^{(m)}(x)>K+1} \left|\omega_{j,x}^{(m)}\right| \cdot \int_{\omega_j^{(m)}(x)>K+1} \left|\omega_{j,x}^{(m)}\right| dx \;=\; \infty.
\end{aligned}$$

This proves that the original sequence was not minimizing. We again conclude that this case cannot happen. This completes the proof of Theorem 2.1. □

4. Proof of Theorem 2.2. We shall rewrite the parabolic equation (2.1) using a different set of variables:

$$v = f(u) - u_x\,, \qquad \tau = t\,, \qquad \eta = u\,. \tag{4.1}$$

This change of variable was first introduced in [BB] and then used in later papers [BB1, BB2]. For each fixed time $t > 0$, the solution of (2.1)–(2.2) is smooth. The map

$$x \mapsto \gamma^t(x) \doteq \big(u(t,x)\,,\ v(t,x)\big) \tag{4.2}$$

parameterizes a curve γ^t in the u-v plane. To see how this curve evolves in time, from (2.3) one obtains

$$v_t + f'(u)v_x = v_{xx}\,. \tag{4.3}$$

On regions where $u_x \neq 0$ we can now use (τ, η) as independent variables, instead of (t, x). From (4.1) and (4.3) we obtain

$$u_x = f(u) - v\,, \qquad v_\eta = \frac{v_x}{u_x}\,, \qquad v_{\eta\eta} = \frac{v_{xx}}{u_x^2} - \frac{v_x}{u_x^3}\,u_{xx}\,,$$

$$v_\tau = v_t - \frac{u_t}{u_x}\,v_x = (v_{xx} - f'(u)v_x) - \frac{v_x}{u_x}\,(u_{xx} - f'(u)u_x) = v_{xx} - \frac{u_{xx}}{u_x}\,v_x\,.$$

Therefore

$$v_\tau = (u_x)^2 v_{\eta\eta} = \big(v - f(\eta)\big)^2 v_{\eta\eta}\,. \tag{4.4}$$

In particular, the curve $\gamma = \gamma(\tau, \eta) = (\eta, v(\tau, \eta))$ evolves in the direction of the curvature and its total length is monotone decreasing in time. Another functional which is monotonically decreasing in time is the area functional

$$Q(\gamma) \doteq \frac{1}{2}\int\!\!\int_{\eta<\tilde\eta} \Big|\gamma_\eta(\eta) \wedge \gamma_\eta(\tilde\eta)\Big|\,d\eta d\tilde\eta, \tag{4.5}$$

defined as the double integral of a wedge product. In terms of the original coordinates u, x, we have

$$\begin{aligned} Q(u) = \int\!\!\int_{x<\tilde x} &|u_x(\tilde x)\cdot[f'(u(x))\cdot u_x(x) - u_{xx}(x)] \\ &-u_x(x)\cdot[f'(u(\tilde x))\cdot u_x(\tilde x) - u_{xx}(\tilde x)]|\;dxd\tilde x\,. \end{aligned} \tag{4.6}$$

All these calculations, (4.1)–(4.6), can be found in [BB, BB1]. As proved in [BB1], the decrease of the functional Q controls the area swept by the curve γ in its motion.

By parabolic regularization estimates, at time $t = 1$ we now have

$$Q\big(u(1)\big) \leq C_1 \tag{4.7}$$

for some constant C_1, uniformly valid for all solutions $u \in \mathcal{F}_M$. Therefore

$$\begin{aligned}\int_1^\infty \int \big(v - f(\eta)\big)^2 |v_{\eta\eta}|\, d\eta d\tau \;&\leq\; \int_1^\infty \int \big|v_\tau(\tau,\eta)\big|\, d\eta d\tau \\ &\leq \int_1^\infty -\left\{\frac{d}{dt} Q\big(u(t)\big)\right\} dt \;\leq\; Q\big(u(1)\big) \;\leq\; C_1\,. \end{aligned} \tag{4.8}$$

As a consequence, for any given $\varepsilon > 0$, there exists a set of times $I^u \subset [1, \infty[$ with

$$\text{meas}(I^u) \leq C_1/\varepsilon \tag{4.9}$$

such that

$$\int \big(v(t,\eta) - f(\eta)\big)^2 \cdot \big|v_{\eta\eta}(t,\eta)\big|\, d\eta \leq \varepsilon \qquad \text{for all } t \geq 1, t \notin I^u. \tag{4.10}$$

In addition, the assumption (2.4) of genuine nonlinearity yields the well-known decay estimate $u_x(t,x) \leq (\kappa t)^{-1}$, hence

$$-\infty < u_x(t,x) \leq \varepsilon, \qquad\qquad x \in \mathbb{R}\,, \tag{4.11}$$

for all $t \geq (\kappa\varepsilon)^{-1}$. To achieve a proof of Theorem 2.2, it now suffices to show that at every time t where (4.10)–(4.11) hold with some $\varepsilon > 0$ sufficiently small, the profile of $u(t,\cdot)$ can be suitably approximated by a finite superposition of viscous shock profiles, and (2.12) holds.

As before, set $\delta \doteq M/N$. We can single out finitely many disjoint intervals $I_k = [a_k, b_k]$, $k = 1, \ldots, \nu$, such that

$$\min_{x \in I_k} u_x(t,x) \leq -7\delta^2 \quad \text{for all} \quad k\,,$$

$$u_x(t,x) \leq u_x(t,a_k) = u_x(t,b_k) = -2\delta^3 \quad \text{for all} \quad x \in I_k\,, \tag{4.12}$$

$$u_x(t,x) > -7\delta^2 \quad \text{for all} \quad x \notin I_1 \cup \cdots \cup I_\nu\,. \tag{4.13}$$

The images of these intervals through the mapping $x \mapsto \gamma(x)$ are graphs of functions $v = v^{(k)}(\eta)$, say, with $\eta \in [b_k', a_k'] \doteq \big[u(b_k), u(a_k)\big]$; see Figure 3. For each k we now choose a point $x_k \in [a_k, b_k]$ such that

$$-m_k \;\doteq\; u_x(t,x_k) \;=\; \min_{a_k \leq x \leq b_k} u_x(t,x) \tag{4.14}$$

and call γ^k the segment in the u-v plane with endpoints on the graph of the function f, tangent to the graph of the function $v^{(k)}$ at the point $c_k \doteq u(t,x_k)$, as in Figure 4. Let $u_k^+ < u_k^-$ be the points where γ^k intersects the graph of f, and call ω_k the unique viscous traveling wave profile satisfying

$$\begin{gathered} \omega(-\infty) = u_k^-\,, \quad \omega(\infty) = u_k^+\,, \\ \omega' = f(\omega) - \sigma_k\,\omega - \big[f(u_k^-) - \sigma_k u_k^-\big]\,, \quad \sigma_k = \frac{f(u_k^-) - f(u_k^+)}{u_k^- - u_k^+} = f'\big(u(t,x_k)\big)\,, \\ \omega(x_k) = u(t,x_k)\,, \quad f'(\omega(x_k)) = \sigma_k\,. \end{gathered} \tag{4.15}$$

It is important to notice that, by the previous construction, the image of the one-to-one map

$$x \mapsto \Big(\omega_k(x)\,,\; f\big(\omega_k(x)\big) - \omega_{k,x}(x)\Big)$$

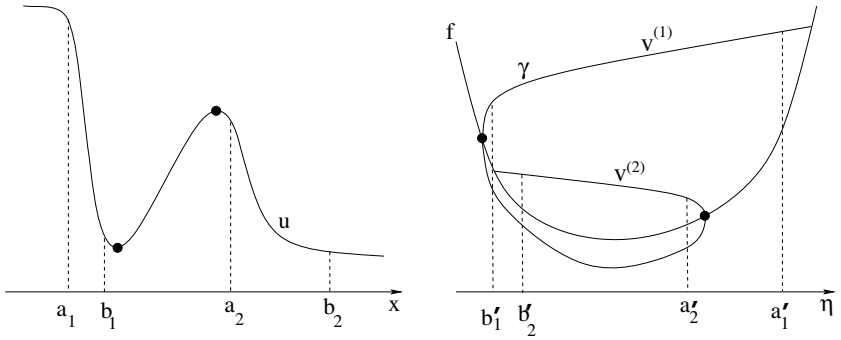


FIG. 3. *Example of a solution of the viscous conservation law and the corresponding curve in the $v-\eta$ plane.*

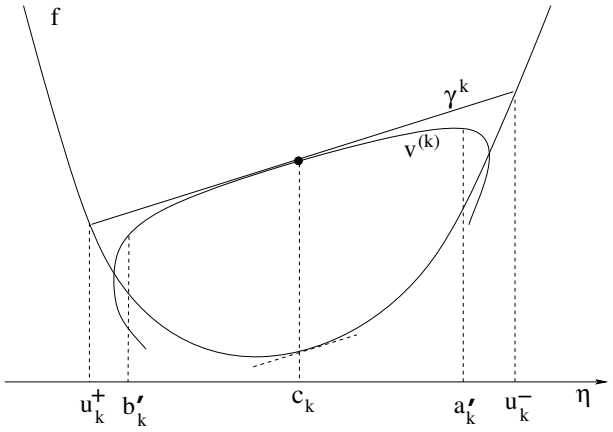


FIG. 4. *Fitting in a viscous shock ω_k, illustrated in the $v-\eta$ plane.*

is precisely the segment γ^k. Moreover, the tangency condition and the maximality condition (4.14) imply that, at $x = x_k$,

$$u_x(t,x_k) = \omega_{k,x}(x_k)\,, \qquad\qquad u_{xx}(t,x_k) = \omega_{k,xx}(x_k) = 0\,.$$

Geometrically, this means that both $u(t,\cdot)$ and $\omega_k(\cdot)$ have an inflection point at $x = x_k$. We now recall that, by (4.10),

$$\sum_k \int_{b'_k}^{a'_k} \Big(v^{(k)}(\eta) - f(\eta)\Big)^2 \cdot \Big|v^{(k)}_{\eta\eta}(t,\eta)\Big|\, d\eta \;\leq\; \varepsilon\,. \tag{4.16}$$

Restricted to the region where $u_x \leq -\delta^3$, the previous inequality implies the key estimate

$$\sum_{k=1}^{\nu} \int_{\{v^{(k)}(\eta)-f(\eta)\geq\delta^3\}} \Big|v^{(k)}_{\eta\eta}(t,\eta)\Big|\, d\eta \;\leq\; \frac{\varepsilon}{\delta^6}\,. \tag{4.17}$$

Since $\varepsilon > 0$ can be chosen arbitrarily small, according to (4.17) every function $v^{(k)}$ is almost affine, hence its graph is very well approximated by the tangent line γ^k. Reverting to the original variables t, x, this in turn implies that $u(t,\cdot)$ is closely approximated by the corresponding traveling profile ω_k on the appropriate interval $x \in [a_k, b_k]$.

LEMMA 4.1. *Assume that the flux function satisfies*

$$0 < \kappa \le f''(u) \le \kappa' \qquad \text{for all } u \in \mathbb{R}\,. \tag{4.18}$$

Then for every $\varepsilon' > 0$ there exists $\varepsilon > 0$ small enough so that (4.17) *implies the following:*

$$\|u - \omega_k\|_{\mathbf{L}^\infty([a_k,b_k])} \le \varepsilon'\,, \quad \|u - \omega_k\|^2_{H^1([a_k,b_k])} \le \varepsilon' \quad \text{for all } k\,. \tag{4.19}$$

Moreover,

$$u(t,a_k) - u(t,b_k) \ge \sqrt{\frac{m_k}{\kappa'}}\,, \tag{4.20}$$

$$\sup_{x \notin [a_k,b_k]} |\omega_{k,x}(x)| \le 3\delta^3\,, \tag{4.21}$$

$$\int_{\mathbb{R}\setminus[a_k,b_k]} |\omega_{k,x}(x)|\,dx \le \frac{6\delta^2}{\sqrt{\kappa}}\,. \tag{4.22}$$

Proof. By choosing $\varepsilon > 0$ sufficiently small, we can assume that the $\mathcal{C}^1$ distance

$$\left\| v^{(k)} - \gamma^k \right\|_{\mathcal{C}^1\left([b'_k,a'_k]\right)} \tag{4.23}$$

is as small as we like. By (4.12), when $x \in [a_k,b_k]$ we have $u_x \le -2\delta^3$. The map $x \mapsto u(x)$ is thus invertible on each interval $[a_k,b_k]$. The two norms in (4.19) can both be estimated in terms of the distance (4.23).

We now prove (4.20). Using (4.23) and recalling (4.12), by taking $\varepsilon > 0$ sufficiently small we can assume that

$$\begin{aligned} \gamma^k(a'_k) - f(a'_k) &\le v^{(k)}(a'_k) - f(a'_k) + \|v^{(k)} - \gamma^k\|_{\mathcal{C}^0} \le 3\delta^3\,, \\ \gamma^k(b'_k) - f(b'_k) &\le v^{(k)}(b'_k) - f(b'_k) + \|v^{(k)} - \gamma^k\|_{\mathcal{C}^0} \le 3\delta^3\,. \end{aligned}$$

The inequality (4.20) now follows from a simple geometrical inequality (see Figure 5(a)). If $f'' < \kappa'$ and γ is a linear function such that

$$\gamma(a') - f(a') \le 3\delta^3, \qquad \gamma(b') - f(b') \le 3\delta^3, \qquad \gamma(c) - f(c) = m_k\,,$$

for some points $b' < c < a'$, then

$$a' - b' \ge \sqrt{\frac{2(m_k - 3\delta^3)}{\kappa'}}\,.$$

Since we are assuming $\delta < 1$, $m_k \ge 7\delta^2 > 2(3\delta^3)$, from the previous inequality we deduce

$$a'_k - b'_k \ge \sqrt{\frac{m_k}{\kappa'}}\,,$$

proving (4.20).

The inequality (4.21) follows from

$$\sup_{x \notin [a_k,b_k]} |\omega_{k,x}(x)| = \max\left\{|\omega_{k,x}(a'_k)|\,,\ |\omega_{k,x}(b'_k)|\right\} \le 2\delta^3 + \left\| v^{(k)} - \gamma^k \right\|_{\mathcal{C}^1([b'_k,a'_k])} \le 3\delta^3\,.$$

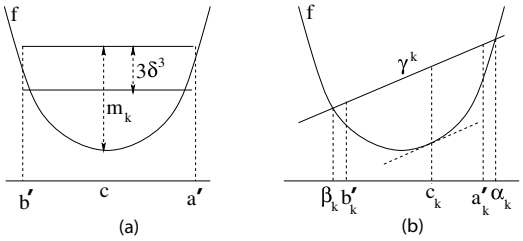


FIG. 5. *Some geometrical properties of convex functions.*

To prove (4.22), call α_k, β_k the points where the line γ^k intersects the graph of f, as in Figure 5(b). Then

$$\int_{\mathbb{R}\setminus[a_k,b_k]} |\omega_{k,x}(x)|\, dx = (\alpha_k - a'_k) + (b'_k - \beta_k)\,.$$

Consider the function $g(u) \doteq \gamma^k(u) - f(u)$. Clearly, we have

$$g(c_k) = \max_u g(u) \geq 7\delta^2,$$

and c_k is the midpoint of the interval $[\beta_k, \alpha_k]$. Assuming $f'' \geq \kappa$, then

$$a'_k - c_k \geq \frac{1}{2}\sqrt{\frac{4(m_k - 3\delta^3)}{\kappa}}.$$

Recalling that

$$g(a'_k) \leq 3\delta^3\,, \qquad g(b'_k) \leq 3\delta^3\,, \qquad g'' = -f'' \leq -\kappa\,,$$

we conclude that

$$-g'(a'_k) = -\int_{c_k}^{a'_k} g''(u)\, du \geq \kappa(a'_k - c_k) \geq \sqrt{\kappa(m_k - 3\delta^3)}.$$

Recalling that $m_k \geq 7\delta^2$ and $\delta < 1$, we obtain

$$\alpha_k - a'_k \leq \frac{3\delta^3}{\sqrt{\kappa(m_k - 3\delta^3)}} \leq \frac{6\delta^3}{\sqrt{8\kappa\delta^2}} \leq \frac{3\delta^2}{\sqrt{\kappa}}\,.$$

The estimate for $b'_k - \beta_k$ is totally similar. Together, these yield (4.22). The proof of the lemma is completed. □

As approximations to $u(t,\cdot)$ we now choose the N traveling profiles ω_k in the above list, corresponding to the N largest values of m_k, say, $m_1 \geq m_2 \geq \cdots \geq m_N$. Notice that (4.20) implies

$$N\sqrt{\frac{m_N}{\kappa'}} \leq M\,, \qquad\qquad m_N \leq \kappa' \left(\frac{M}{N}\right)^2.$$

Hence

$$|u_x(t,x)| \leq \kappa' \left(\frac{M}{N}\right)^2, \qquad\qquad x \notin \bigcup_{k=1}^{N} [a_k, b_k]\,. \tag{4.24}$$

Using the estimates (4.19)–(4.21) we now check that the functional $\mathcal{J}$ at (2.7) is small, as claimed by Theorem 2.2. The first half of the right-hand side in (2.7) can be estimated as

$$\begin{aligned}
&\sum_{k=1}^N \int_{\mathbb{R}} |u(t,x)-\omega_k(x)|\cdot|\omega_{k,x}(x)|^2\,dx\\
&\quad=\sum_{k=1}^N\left(\int_{x\in[a_k,b_k]}+\int_{x\notin[a_k,b_k]}\right)|u(t,x)-\omega_k(x)|\cdot|\omega_{k,x}(x)|^2\,dx\\
&\quad\leq \sum_k \|u-\omega_k\|_{\mathbf{L}^\infty([a_k,b_k])}\cdot\|\omega_{k,x}\|^2_{\mathbf{L}^2(\mathbb{R})}\\
&\qquad+\sum_k\Big(\|u\|_{\mathbf{L}^\infty(\mathbb{R})}+\|\omega_k\|_{\mathbf{L}^\infty(\mathbb{R})}\Big)\sup_{x\notin[a_k,b_k]}|\omega_{k,x}(x)|\cdot\int_{\mathbb{R}\setminus[a_k,b_k]}|\omega_{k,x}(x)|\,dx\\
&\quad\leq \mathcal{O}(1)\cdot N\,\varepsilon'+N\cdot 2M\cdot 3\delta^3\frac{6\delta^2}{\sqrt{\kappa}}\;=\;\mathcal{O}(1)\cdot\frac{1}{N^4}\,.
\end{aligned}$$

For the second half of the right-hand side of (2.7), we can use the inequality $(a+b)^2\leq 2(a^2+b^2)$, valid for all real numbers of a,b, and we get

$$\begin{aligned}
&\int_{\mathbb{R}}\left|u_x(t,x)-\sum_{k=1}^N\omega_{k,x}(x)\right|^2dx\\
&=\sum_k\int_{a_k}^{b_k}\left|u_x(t,x)-\sum_{k=1}^N\omega_{k,x}(x)\right|^2dx+\int_{\mathbb{R}\setminus\cup_k[a_k,b_k]}\left|u_x(t,x)-\sum_{k=1}^N\omega_{k,x}(x)\right|^2dx\\
&\leq\sum_k\int_{a_k}^{b_k}\left(|u_x(t,x)-\omega_{k,x}(x)|+\sum_{j\neq k}|\omega_{j,x}(x)|\right)^2dx\\
&\qquad+\int_{\mathbb{R}\setminus\cup_k[a_k,b_k]}\left(|u_x(t,x)|+\sum_k|\omega_{k,x}(x)|\right)^2dx\\
&\leq 2\sum_k\left\{\int_{a_k}^{b_k}|u_x(t,x)-\omega_{k,x}(x)|^2\,dx+\int_{a_k}^{b_k}\left(\sum_{j\neq k}|\omega_{j,x}(x)|\right)^2dx\right\}\\
&\qquad+2\int_{\mathbb{R}\setminus\cup_k[a_k,b_k]}\left[|u_x(t,x)|^2+\left(\sum_k|\omega_{k,x}(x)|\right)^2\right]dx\\
&\leq I_1+I_2+I_3+I_4\,,
\end{aligned}$$

where

$$I_1=2\sum_k|u-\omega_k|_{H^1([a_k,b_k])}\,,$$

$$I_2 = 2\sum_j \left(\sum_{k\neq j} \int_{a_k}^{b_k} |\omega_{j,x}(x)|^2 \, dx \right) \;\leq\; 2\sum_j \int_{\mathbb{R}\setminus[a_j,b_j]} |\omega_{j,x}(x)|^2 \, dx$$
$$\leq\; 2\sum_j \left\{ \sup_{x\notin[a_j,b_j]} |\omega_{j,x}(x)| \cdot \int_{\mathbb{R}\setminus[a_j,b_j]} |\omega_{j,x}(x)| \, dx \right\},$$
$$I_3 = 2\left\{ \sup_{x\notin\cup[a_k,b_k]} |u_x(t,x)| \right\} \cdot \int_{\mathbb{R}} |u_x(t,x)| \, dx\,,$$
$$I_4 = 2\sum_k \left\{ \sup_{x\notin[a_k,b_k]} |\omega_{k,x}(x)| \right\} \cdot \int_{\mathbb{R}\setminus\cup_k[a_k,b_k]} \sum_k |\omega_{k,x}(x)| \, dx\,.$$

Using the estimates in Lemma 4.1 and (4.24), we get

$$I_1 \leq 2N\varepsilon'\,,$$
$$I_2 \leq 2N\, 3\delta^3\, \frac{6\delta^2}{\sqrt{\kappa}}\,,$$
$$I_3 \leq 2\kappa' \frac{M^2}{N^2}\, M\,,$$
$$I_4 \leq 2\cdot N3\delta^3 \cdot N\, \frac{6\delta^2}{\sqrt{\kappa}}\,.$$

Note that I_1 measures how well the viscous shock profile matches the solution on the interval $[a_k, b_k]$, and this term is arbitrarily small. I_2 measures the H_1 norm of the viscous shock waves outside the interval $[a_k, b_k]$, and it is of $\mathcal{O}(1)/N^4$. And I_3 is the sum of all the shock waves that are not represented. This is the largest term here, and is of $\mathcal{O}(1)/N^2$. Finally, I_4 is similar to I_2, and is of $\mathcal{O}(1)/N^3$. In summary, we have

$$\int_{\mathbb{R}} \left| u_x(t,x) - \sum_{k=1}^{N} \omega_{k,x}(x) \right|^2 dx \;\leq\; \mathcal{O}(1)\cdot\frac{1}{N^2}\,.$$

Putting these two parts together, we get the desired result.

5. Concluding remarks. For solutions to the conservation law (2.1), the transient behavior is nontrivial and can last an arbitrarily long time. This happens because we are considering solutions defined on the whole real line. On the other hand, if the equation is restricted to a bounded interval, say,

$$u_t + f(u)_x = u_{xx}, \qquad x \in]a,b[\,, \tag{5.1}$$

with boundary conditions

$$u(a) = \alpha\,, \qquad u(b) = \beta\,, \tag{5.2}$$

then all solutions would converge at an exponential rate to a unique steady state $w(\cdot)$.

Indeed, from basic theory of parabolic equations [H] it follows that there exists a unique function $w : [a,b] \mapsto \mathbb{R}$ which satisfies the two-point boundary value problem

$$f(w)_x = w_{xx}\,, \qquad w(a) = \alpha\,, \qquad w(b) = \beta\,. \tag{5.3}$$

Linearizing (5.1) around the steady state w, one obtains the existence of some $\delta > 0$ such that, for every initial data $\bar{u} \in \mathbf{L}^2$ the corresponding solution of (5.1)–(5.2) satisfies

$$\|u(t) - w\|_{\mathcal{C}^k([a,b])} \leq C\, e^{-\delta t}.$$

Here one can choose a constant C uniformly valid on bounded subsets of $\mathbf{L}^2$. After an initial time interval, the long-term behavior of the solution is thus trivial.

In the case of a bounded domain, the corresponding equation (4.4) in the (η, v) variables must be supplemented with the boundary conditions

$$v_\eta(\alpha) = v_\eta(\beta) = 0\,.$$

The unique steady state solution of (5.3) corresponds to a constant function:

$$v(\eta) \equiv \kappa \qquad \text{for all } \eta \in [\alpha,\, \beta]\,.$$

Observing that

$$b - a = \int_\alpha^\beta \frac{1}{u_x}\, du = \int_\alpha^\beta \frac{1}{f(\eta) - v(\eta)}\, d\eta\,,$$

one can uniquely determine the constant κ from the relation

$$b - a = \int_\alpha^\beta \frac{1}{f(\eta) - \kappa}\, d\eta\,.$$

Acknowledgments. The authors are grateful to Prof. Alberto Bressan at the same department for proposing the problem and for many useful discussions. We also want to thank one of the referees for carefully reading through the paper and for various remarks.

REFERENCES

[BB] S. Bianchini and A. Bressan, *A case study in vanishing viscosity*, Discrete Contin. Dyn. Syst., 7 (2001), pp. 449–476.

[BB1] S. Bianchini and A. Bressan, *On a Lyapunov functional relating shortening curves and viscous conservation laws*, Nonlinear Anal., 51 (2002), pp. 649–662.

[BB2] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. Math., 161 (2005), pp. 223–342.

[B] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[H] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

[S] D. Serre, *Systems of Conservation Laws* I, Cambridge University Press, Cambridge, UK, 2000.

[Sm] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[T] R. Temam, *Infinite Dimensional Dynamical Systems in Mathematics and Physics*, Springer-Verlag, New York, 1988.

[W] G. Whitham, *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1489–1507

SYNCHRONIZATION OF A STOCHASTIC REACTION-DIFFUSION SYSTEM ON A THIN TWO-LAYER DOMAIN*

TOMÁS CARABALLO†, IGOR D. CHUESHOV‡, AND PETER E. KLOEDEN§

Abstract. A system of semilinear parabolic stochastic partial differential equations with additive space-time noise is considered on the union of thin bounded tubular domains $D_{1,\varepsilon} := \Gamma \times (0, \varepsilon)$ and $D_{2,\varepsilon} := \Gamma \times (-\varepsilon, 0)$ joined at the common base $\Gamma \subset \mathbb{R}^d$, where $d \geq 1$. The equations are coupled by an interface condition on Γ which involves a reaction intensity $k(x', \varepsilon)$, where $x = (x', x_{d+1}) \in \mathbb{R}^{d+1}$ with $x' \in \Gamma$ and $|x_{d+1}| < \varepsilon$. Random influences are included through additive space-time Brownian motion, which depend only on the base spatial variable $x' \in \Gamma$ and not on the spatial variable x_{d+1} in the thin direction. Moreover, the noise is the same in both layers $D_{1,\varepsilon}$ and $D_{2,\varepsilon}$. Limiting properties of the global random attractor are established as the thinness parameter of the domain $\varepsilon \to 0$, i.e., as the initial domain becomes thinner, when the intensity function possesses the property $\lim_{\varepsilon \to 0} \varepsilon^{-1} k(x', \varepsilon) = +\infty$. In particular, the limiting dynamics is described by a single stochastic parabolic equation with the averaged diffusion coefficient and a nonlinearity term, which essentially indicates synchronization of the dynamics on both sides of the common base Γ. Moreover, in the case of nondegenerate noise we obtain stronger synchronization phenomena in comparison with analogous results in the deterministic case previously investigated by Chueshov and Rekalo [*EQUADIFF*-2003, F. Dumortier et al., eds., World Scientific, Hackensack, NJ, 2005, pp. 645–650; *Sb. Math.*, 195 (2004), pp. 103–128].

Key words. thin domains, random attractors, synchronization

AMS subject classifications. Primary, 37L55; Secondary, 60H15, 35M20, 37L30

DOI. 10.1137/050647281

1. Introduction.

Let $D_{1,\varepsilon}$ and $D_{2,\varepsilon}$ be thin bounded domains in $\mathbb{R}^{d+1}$, where $d \geq 1$, of the form

$$D_{1,\varepsilon} = \Gamma \times (0, \varepsilon), \qquad D_{2,\varepsilon} = \Gamma \times (-\varepsilon, 0),$$

where $0 < \varepsilon \leq 1$ and Γ is a bounded C^2-domain in $\mathbb{R}^d$. We write $x \in D_\varepsilon := D_{1,\varepsilon} \cup D_{2,\varepsilon}$ as $x = (x', x_{d+1})$, where $x' \in \Gamma$ and $x_{d+1} \in (0, \varepsilon)$ or $x_{d+1} \in (-\varepsilon, 0)$, and will not distinguish between the sets $\Gamma \times \{0\} \subset \mathbb{R}^{d+1}$ and $\Gamma \subset \mathbb{R}^d$.

We consider the following system of semilinear parabolic equations:

$$\frac{\partial}{\partial t} U^i - \nu_i \Delta U^i + a U^i + f_i(U^i) + h_i(x) = \dot{W}(t, x'), \quad t > 0, \ \ x \in D_{i,\varepsilon}, \ \ i = 1, 2, \tag{1}$$

with the initial data

$$U^i(0, x) = U_0^i(x), \qquad x \in D_{i,\varepsilon}, \ \ i = 1, 2, \tag{2}$$

*Received by the editors December 12, 2005; accepted for publication (in revised form) August 28, 2006; published electronically January 12, 2007. This work was partly supported by Ministerio de Educación y Ciencia (Spain) and FEDER (European Community), project MTM2005-01412.
http://www.siam.org/journals/sima/38-5/64728.html

†Departamento Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Apdo. de Correos 1160, 41080-Sevilla, Spain (caraball@us.es).

‡Department of Mechanics and Mathematics, Kharkov National University, 4 Svobody sq., 61077, Kharkov, Ukraine (chueshov@univer.kharkov.ua).

§Institut für Mathematik, Johann Wolfgang Goethe-Universität, D-60054 Frankfurt am Main, Germany (kloeden@math.uni-frankfurt.de). This author's work was supported by Ministerio de Educación y Ciencia (Spain) under grant SAB2004-0146, within the Programa de Movilidad del Profesorado universitario español y extranjero.

where the $\dot{W}(t,x')$ is a Gaussian white noise depending on the spatial variable $x' \in \Gamma$ (but not on the x_{d+1} spatial variable).

We assume that U^1 and U^2 satisfy the Neumann boundary conditions

$$\big(\nabla U^i, n_i\big) = 0, \qquad x \in \partial D_{i,\varepsilon} \setminus \Gamma, \quad i = 1,2, \tag{3}$$

on the external part of the boundary of the compound domain D_ε, where n is the outer normal to ∂D_ε, and a matching condition on Γ of the form

$$\begin{aligned} \left(-\nu_1 \frac{\partial U^1}{\partial x_{d+1}} + k(x',\varepsilon)(U^1 - U^2)\right)\Big|_\Gamma &= 0, \\ \left(\nu_2 \frac{\partial U^2}{\partial x_{d+1}} + k(x',\varepsilon)(U^2 - U^1)\right)\Big|_\Gamma &= 0. \end{aligned} \tag{4}$$

Here the above constants ν_i and a are positive numbers.

We impose the following assumptions:

- for $i = 1,2$ the function $f_i \in C^1(\mathbb{R})$ possesses the property $f_i'(v) \geq -c$ for all $v \in \mathbb{R}$ and also satisfies the relations

$$v f_i(v) \geq a_0 |v|^{p+1} - c, \quad |f_i'(v)| \leq a_1 |v|^{p-1} + c, \quad v \in \mathbb{R}, \tag{5}$$

  where a_j and c are positive constants and $1 \leq p < 3$;
- $h_i \in H^1(D_{i,1})$, $i = 1,2$;
- the interface reaction intensity $k(x',\varepsilon)$ satisfies

$$k(\cdot,\varepsilon) \in L^\infty(\Gamma), \quad k(x',\varepsilon) > 0 \text{ for } x' \in \Gamma,\ \varepsilon \in (0,1],$$

  and

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} k(x',\varepsilon) = +\infty,\ x' \in \Gamma, \text{ in Lebesgue measure (see Remark 1.1)}; \tag{6}$$

- $W(t)$, $t \in \mathbb{R}$, is a two-sided $L_2(\Gamma)$-valued Wiener process with covariance operator $K = K^* \geq 0$ such that

$$\operatorname{tr}\Big[K\left(-\Delta_N + 1\right)^{2\beta-1}\Big] < \infty \quad \text{for some} \quad \beta > \max\left\{1\,,\ \frac{d}{4}\right\}, \tag{7}$$

  where Δ_N is the Laplace operator in $L_2(\Gamma)$ with the Neumann boundary conditions on $\partial\Gamma$. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the corresponding probability space, and by $\dot{W} \equiv \partial_t W$ the generalized derivative with respect to t.

Remark 1.1. Our main example of the interface reaction intensity is the following function:

$$k(x',\varepsilon) = \varepsilon^\alpha k_0(x') \in L^\infty(\Gamma), \quad k_0(x') > 0 \quad \text{for } x' \in \Gamma,\ \varepsilon \in (0,1],$$

for some $\alpha \in [0,1)$. We also note that the convergence in (Lebesgue) measure to infinity means that

$$\lim_{\varepsilon \to 0} \operatorname{Leb}\left\{x' \in \Gamma\ :\ \varepsilon^{-1} k(x',\varepsilon) \leq N\right\} = 0 \ \text{ for any } \ N > 0.$$

Problem (1)–(4) is a model for a reaction-diffusion system consisting of two components filling thin contacting layers $D_{1,\varepsilon}$ and $D_{2,\varepsilon}$ separated by a penetrable membrane Γ. Reaction of the components is possible on the surface Γ only, and the reaction

intensity $k(x', \varepsilon)$ depends on the thickness of the domains filled by the reactants. The deterministic version of the model was considered by Chueshov and Rekalo [12, 13], while Rekalo [27] investigated the special case of identical equations in both layers with $k(\varepsilon, x)$ independent of ε. The stochastic version considered in the present paper allows for irregularities and random effects on the separating membrane.

Hale and Raugel [21, 22] initiated the analysis of asymptotic dynamics of deterministic semilinear reaction-diffusion equations on thin domains. Some extensions of their results can be also found in [16] and [26]. In all these papers, a reaction-diffusion equation is endowed with homogeneous Neumann boundary conditions. To our knowledge stochastic evolution equations have not previously been investigated on thin domains.

In this paper we investigate the pathwise asymptotic behavior of the above stochastic evolution system by converting it into a system of pathwise random partial differential equations (PDEs) to which deterministic methods can be applied in a pathwise manner.

Our main result deals with properties of random (global) pullback attractors for the random dynamical system generated by (1)–(4) in $L_2(D_\varepsilon)$. In particular, we prove that these pullback attractors are closely related to the corresponding object for the problem

$$\frac{\partial}{\partial t}U - \nu\Delta_{x'}U + aU + f(U) + h(x') = \dot{W}(t, x'), \qquad t > 0, \;\; x' \in \Gamma, \tag{8}$$

on the spatial domain Γ with the Neumann boundary conditions on $\partial\Gamma$. Here we denote

$$\nu = \frac{\nu_1 + \nu_2}{2}, \quad f(U) = \frac{f_1(U) + f_2(U)}{2}, \quad h(x') = \frac{h_1(x', 0) + h_2(x', 0)}{2}. \tag{9}$$

This is essentially a statement about the synchronization of the dynamics of the system in the two thin layers at the level of global pullback attractors. Since, in principle, a global attractor can be a rather complicated set, the synchronization at this level does not imply that *any* pair of trajectories becomes asymptotically synchronized. However, in the case of nondegenerate noise ($Kh = 0$ if and only if $h = 0$ and the image of K is dense in $L_2(\Gamma)$) we can prove, in contrast with the deterministic counterpart, that the global pullback attractor for (8) is a singleton. This means that we also have asymptotic synchronization in our system at the level of trajectories. Thus we observe a stronger synchronizing effect of a nondegenerate stochastic noise in the system under consideration.

The synchronization of stochastic stationary solutions (i.e., single-valued random attractors) of finite dimensional stochastic systems was considered in [5]. See also [1, 23] for similar results in deterministic nonautonomous systems and [7, 28] for autonomous infinite dimensional systems.

The synchronization of coupled systems is a ubiquitous phenomenon in the biological and physical science and is also known to occur in a number of social science contexts. A descriptive account of its diversity of occurrence can be found in the recent book of Strogatz [32], which contains an extensive list of references. In particular, synchronization provides an explanation for the emergence of spontaneous order in the dynamical behavior of coupled systems, which in isolation may exhibit chaotic dynamics. It has been shown to persist in the presence of environmental noise provided that appropriate concepts of random attractors and stochastic stationary solutions are used instead of their deterministic counterparts [5]. As mentioned above, in this

paper we will see that the presence of additive noise can lead to a strengthening of the synchronization, i.e., at the level of trajectories rather than attractors, which does not occur in the absence of noise.

Since most of our analysis is a pathwise analysis applied to pathwise defined random PDEs, i.e., with the stationary Ornstein–Uhlenbeck process appearing as a space-time dependent coefficient, it is reasonable to expect that similar results will also hold for other kinds of noise, for example, with fractional Brownian motion in the original stochastic partial differential equations (SPDEs). The results will be presented in a forthcoming paper.

The paper is organized as follows. We start with preliminary section 2 containing background material from the theory of random systems which we need to state and discuss our main results in section 3. Further sections are devoted to the proof of our main theorem, Theorem 3.1.

2. Random dynamical systems. In order to formulate our results we need some notation and results from the theory of random dynamical systems (with continuous time) and random attractors.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathcal{X}, d_{\mathcal{X}})$ be a complete separable metric (Polish) space. Arnold [2] defined a random dynamical system (RDS) (θ, ϕ) on $\Omega \times \mathcal{X}$ in terms of a metric dynamical system θ on Ω, which represents the noise driving the system, and a cocycle mapping $\phi : \mathbb{R}_+ \times \Omega \times \mathcal{X} \to \mathcal{X}$, which represents the dynamics in the state space $\mathcal{X}$ and satisfies the following properties:

1. $\phi(0, \omega)\phi_0 = \phi_0$ for all $\phi_0 \in \mathcal{X}$ and $\omega \in \Omega$;
2. $\phi(s + t, \omega)\phi_0 = \phi(s, \theta_t\omega)\phi(t, \omega)\phi_0$ for all s, $t \geq 0$, $\phi_0 \in \mathcal{X}$, and $\omega \in \Omega$;
3. $(t, \phi_0) \mapsto \phi(t, \omega)\phi_0$ is continuous for each $\omega \in \Omega$; and
4. $\omega \mapsto \phi(t, \omega)\phi_0$ is $\mathcal{F}$-measurable for all $(t, \phi_0) \in \mathbb{R}_+ \times \mathcal{X}$.

We recall that a metric dynamical system $\theta \equiv (\Omega, \mathcal{F}, \mathbb{P}, \{\theta_t, t \in \mathbb{R}\})$ is a family of measure-preserving transformations $\{\theta_t : \Omega \mapsto \Omega, t \in \mathbb{R}\}$ such that

(i) $\theta_0 = id, \theta_t \circ \theta_s = \theta_{t+s}$ for all $t, s \in \mathbb{R}$;

(ii) the map $(t, \omega) \mapsto \theta_t\omega$ is measurable, and $\theta_t\mathbb{P} = \mathbb{P}$ for all $t \in \mathbb{R}$.

RDSs (with continuous time) are generated by differential equations with random coefficients or stochastic differential equations with a unique and global solution, as well as by infinite dimensional stochastic evolution equations with additive noise. We refer to [2] for more details on the general theory of RDS theory.

To construct an RDS in our case we first need to associate a metric dynamical system θ with the Wiener process W on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $L_2(\Gamma)$. The probability measure $\mathbb{P}$ of this process can be realized on $\mathcal{F} = \mathcal{B}(C_0(\mathbb{R}, L_2(\Gamma)))$, where $C_0(\mathbb{R}, L_2(\Gamma))$ is the Fréchet space of continuous functions on $\mathbb{R}$ with values in $L_2(\Gamma)$ which are zero at time zero. For this realization we introduce the flow $(\theta_t)_{t\in\mathbb{R}}$ given by the *Wiener shift*

$$\theta_t\omega(\cdot) = \omega(\cdot + t) - \omega(t), \quad t \in \mathbb{R}. \tag{10}$$

Interpreting the above Wiener process in the canonical sense $W(\cdot, \omega) = \omega(\cdot)$, it follows that (10) is the well-known helix property of a Wiener process:

$$W(t + s, \omega) - W(s, \omega) = W(t, \theta_s\omega), \quad s, t \in \mathbb{R}, \quad \omega \in \Omega.$$

We now introduce the Ornstein–Uhlenbeck process as a stationary solution of the linear stochastic evolution equation

$$\frac{\partial}{\partial t}U = \nu\Delta_{x'}U - aU + \dot{W}(t, x'), \qquad t > 0, \;\; x' \in \Gamma,$$

on the spatial domain Γ with Neumann boundary conditions on $\partial\Gamma$. Here, as above, we denote $\nu = (\nu_1 + \nu_2)/2$. This process $\eta(t)$ can be written in the form

$$\eta(t,\omega) := \left(\int_{-\infty}^{t} e^{-(t-\tau)A_0} dW(\tau)\right)(\omega), \tag{11}$$

where $A_0 = -\nu\Delta_N + a$ and Δ_N is the Laplace operator in $L_2(\Gamma)$ with the Neumann boundary conditions on $\partial\Gamma$. The integral in (11) exists as an operator stochastic integral (see, e.g., [24] or [19]) We can also involve a perfection procedure to define $\eta(t,\omega) \equiv \bar{\eta}(\theta_t\omega)$ for *all* $\omega \in \Omega$ (for details see [14]). Moreover, under condition (7), $t \mapsto \bar{\eta}(\theta_t\omega)$ is continuous from $\mathbb{R}$ into $D(A_0^{\beta'}) \subset H^{2\beta'}(\Gamma)$ for each $\omega \in \Omega$, where $\beta' \in [0,\beta)$ is arbitrary, and the *temperedness* condition

$$\sup_{t\in\mathbb{R}}\{\| A_0^{\beta'}\bar{\eta}(\theta_t\omega) \| e^{-\gamma|t|}\} < \infty \quad \forall\ \gamma > 0,\ \omega \in \Omega,$$

is satisfied. We also note that under condition (7), since $H^s(\Gamma) \subset C(\overline{\Gamma})$ for $s > d/2$, we have that $t \mapsto \bar{\eta}(\theta_t\omega)$ is a pathwise continuous tempered process with values in $D(A_0) \cap C(\overline{\Gamma})$. In particular

$$\bar{\eta}(\theta_t\omega) \in C\left(\mathbb{R}; C(\overline{\Gamma}) \cap \left\{\psi \in H^2(\Gamma) : \psi \text{ satisfies Neumann b.c. on } \partial\Gamma\right\}\right) \tag{12}$$

for every $\omega \in \Omega$. We will use this observation later.

We recall the following definition of a random set (see [2] or [4]).

DEFINITION 2.1 (random set). *Let $\mathcal{X}$ be a Polish space with a metric $d_{\mathcal{X}}$. A multifunction $\omega \mapsto D(\omega) \neq \emptyset$ is said to be a* random set *if the mapping $\omega \mapsto \operatorname{dist}_{\mathcal{X}}(v, D(\omega))$ is measurable for any $v \in \mathcal{X}$, where $\operatorname{dist}_{\mathcal{X}}(v,B)$ is the distance in $\mathcal{X}$ between the element v and the set $B \subset \mathcal{X}$. For ease of notation we denote the random set $\omega \mapsto D(\omega)$ by $\widehat{D}$ or $\{D(\omega)\}$. If $D(\omega)$ is closed for each $\omega \in \Omega$, then $\widehat{D}$ is called a* random closed set*, while if $D(\omega)$ is a compact set for all $\omega \in \Omega$, then $\widehat{D}$ is called a* random compact set*. A random set $\{D(\omega)\}$ is said to be* tempered *if there exists a $v_0 \in \mathcal{X}$ such that $D(\omega) \subset \{v \in \mathcal{X} : d_{\mathcal{X}}(v, v_0) \le r(\omega)\}$ for all $\omega \in \Omega$, where the random variable $r(\omega) > 0$ is tempered, i.e.,*

$$\sup_{t\in\mathbb{R}}\{r(\theta_t\omega)e^{-\gamma|t|}\} < \infty \quad \forall\ \gamma > 0,\ \omega \in \Omega.$$

We denote by $\mathcal{D}$ the collection of all tempered random sets in $\mathcal{X}$.

Below we also need the concept of a random attractor for RDSs (see, e.g., [2, 17, 18, 29] and the references therein), which extends the corresponding definition of a global attractor in autonomous systems (cf. [3, 9, 33], for example).

DEFINITION 2.2. *Let (θ,ϕ) be an RDS with the phase space $\mathcal{X}$. A random closed set $\{\mathfrak{A}(\omega)\}$ from $\mathcal{D}$ is said to be a* random pullback attractor *for (θ,ϕ) in $\mathcal{D}$ if* (i) *$\widehat{\mathfrak{A}}$ is an invariant set, i.e., $\phi(t,\omega)\mathfrak{A}(\omega) = \mathfrak{A}(\theta_t\omega)$ for $t \ge 0$ and $\omega \in \Omega$, and* (ii) *$\widehat{\mathfrak{A}}$ is pullback attracting in $\mathcal{D}$, i.e.,*

$$\lim_{t\to+\infty} d_{\mathcal{X}}\{\varphi(t,\theta_{-t}\omega)D(\theta_{-t}\omega)\,|\,\mathfrak{A}(\omega)\} = 0, \quad \omega \in \Omega,$$

for all $\widehat{D} \in \mathcal{D}$, where $d_{\mathcal{X}}\{A|B\} = \sup_{a\in A}\operatorname{dist}_{\mathcal{X}}(a,B)$.

Note that a pullback attractor is also a weak *forward* attractor; i.e., we have that

$$\lim_{t\to+\infty}\int_{\Omega} d_{\mathcal{X}}\{\varphi(t,\omega)D(\omega)\,|\,\mathfrak{A}(\theta_t\omega)\}\mathbb{P}(d\omega) = 0 \quad \forall\widehat{D} \in \mathcal{D}.$$

If the random attractor consists of singleton sets, i.e., $\mathfrak{A}(\omega) = \{X^*(\omega)\}$ for some random variable X^* with $X^*(\omega) \in \mathcal{X}$, then $X_t^*(\omega) := X^*(\theta_t\omega))$ is a stationary stochastic process on $\mathcal{X}$.

The following result [18] ensures the existence of a random attractor for an RDS on a Polish space.

THEOREM 2.3. *Let (θ, ϕ) be a continuous or discrete time RDS on $\Omega \times \mathcal{X}$ such that $\phi(t, \omega, \cdot) : \mathcal{X} \to \mathcal{X}$ is a compact operator for each fixed $t > 0$ and $\omega \in \Omega$. If there exists a tempered random set $\widehat{\mathfrak{B}} = \{\mathfrak{B}(\omega), \omega \in \Omega\}$ and a $T_{\widehat{D},\omega} \geq 0$ such that*

$$\phi(t, \theta_{-t}\omega)D(\theta_{-t}\omega) \subset \mathfrak{B}(\omega) \qquad \forall t \geq T_{\widehat{D},\omega},$$

for every tempered random set $\widehat{D}$, then the RDS (θ, ϕ) has a random pullback attractor $\widehat{\mathfrak{A}} = \{\mathfrak{A}(\omega), \omega \in \Omega\}$ with the component subsets defined for each $\omega \in \Omega$ by

$$\mathfrak{A}(\omega) = \bigcap_{s>0} \overline{\bigcup_{t \geq s} \phi(t, \theta_{-t}\omega)\mathfrak{B}(\theta_{-t}\omega)}^{d_{\mathcal{X}}}.$$

The family $\{\mathfrak{B}(\omega)\}$ is called a pullback absorbing random set for the RDS.

3. Main results. Now we are in position to state our main results which we formulate in the theorem below. This says that the limiting dynamics of the system (1)–(4) is given by that of the averaged system (8) on Γ, which one can interpret as the synchronization of dynamics of the original system on the two sides of the membrane Γ. In addition, if the system is the same on both sides of the membrane, then the limiting behavior is independent of the thinness parameter ε when it is sufficiently small.

THEOREM 3.1. *Under the conditions above the following assertions hold.*

1. *Problem* (1)–(4) *generates an RDS $(\theta, \bar{\phi}_\varepsilon)$ in the space*

$$\mathcal{H}_\varepsilon = L_2(D_{1,\varepsilon}) \oplus L_2(D_{2,\varepsilon}) \sim L_2(D_\varepsilon)$$

with the metric dynamical system θ generated by the Wiener process W and the cocycle $\bar{\phi}_\varepsilon$ defined by the formula $\bar{\phi}_\varepsilon(t, \omega)U_0 = U(t, \omega)$, where $U(t, \omega) = (U^1(t, \omega); U^2(t, \omega))$ is a strong (in the sense of stochastic equations [19]*) solution to problem* (1)–(4) *and $U_0 = (U_0^1; U_0^2)$.*

2. *Similarly, problem* (8) *generates an RDS $(\theta, \bar{\phi}_0)$ in the space $L_2(\Gamma)$.*
3. *The cocycles $\bar{\phi}_\varepsilon$ converge to $\bar{\phi}_0$ in the sense that*

$$\lim_{\varepsilon \to 0} \sup_{t \in [0,T]} \frac{1}{\varepsilon} \int_{D_\varepsilon} |\bar{\phi}_\varepsilon(t, \omega)v - \bar{\phi}_0(t, \omega)v|^2 dx = 0 \quad \forall \omega \in \Omega,$$

for any $v(x) \in \mathcal{H}_\varepsilon$ independent of the variable x_{d+1}, and for any $T > 0$.

4. *These RDS $(\theta, \bar{\phi}_\varepsilon)$ and $(\theta, \bar{\phi}_0)$ have random compact pullback attractors $\{\bar{\mathfrak{A}}^\varepsilon(\omega)\}$ and $\{\bar{\mathfrak{A}}^0(\omega)\}$ in their corresponding phase spaces. Moreover, if the correlation operator K of the Wiener process W is nondegenerate in the sense that* (i) *$Kh = 0$ if and only if $h = 0$, and* (ii) *the image of K is dense in $L_2(\Gamma)$, then the attractor $\{\bar{\mathfrak{A}}^0(\omega)\}$ is a singleton, i.e., $\bar{\mathfrak{A}}^0(\omega) = \{\bar{v}_0(\omega)\}$, where $\bar{v}_0(\omega)$ is a tempered random variable with values in $L_2(\Gamma)$.*
5. *The attractors $\{\bar{\mathfrak{A}}^\varepsilon(\omega)\}$ are upper semicontinuous as $\varepsilon \to 0$ in the sense that*

$$\lim_{\varepsilon \to 0} \sup_{v \in \bar{\mathfrak{A}}^\varepsilon(\omega)} \left\{ \inf_{v_0 \in \bar{\mathfrak{A}}^0(\omega)} \frac{1}{\varepsilon} \int_{D_\varepsilon} |v(x', x_{d+1}) - v_0(x')|^2 dx \right\} = 0 \quad \forall \omega \in \Omega. \tag{13}$$

6. *In addition, if*

$$\begin{aligned} &\nu_1 = \nu_2 := \nu, \quad f_1(U) = f_2(U) := f(U), \\ &h_1(x', x_{d+1}) = h(x'), \quad h_2(x', x_{d+1}) = h(x'); \end{aligned} \tag{14}$$

$f(U)$ is globally Lipschitz, i.e., there exists a constant $L > 0$ such that

$$|f(U) - f(V)| \le L|U - V|, \quad U, V \in \mathbb{R}, \tag{15}$$

and also that

$$k(x', \varepsilon) > k_\varepsilon \text{ for } x' \in \Gamma,\ \varepsilon \in (0, 1]; \quad \text{and} \quad \lim_{\varepsilon \to 0} \varepsilon^{-1} k_\varepsilon = +\infty, \tag{16}$$

then there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0]$ the global random pullback attractor $\{\bar{\mathfrak{A}}^\varepsilon(\omega)\}$ for $(\theta, \bar{\phi}_\varepsilon)$ has the form

$$\bar{\mathfrak{A}}^\varepsilon(\omega) \equiv \left\{ v(x', x_{d+1}) \equiv v_0(x') \ : \ v_0 \in \bar{\mathfrak{A}}^0(\omega) \right\},$$

where $\{\bar{\mathfrak{A}}^0(\omega)\}$ is the random pullback attractor for the RDS $(\theta, \bar{\phi}_0)$.

Remark 3.2. In the case when $\bar{\mathfrak{A}}^0(\omega) = \{\bar{v}_0(\omega)\}$ is a singleton, relation (13) turns into the equality

$$\lim_{\varepsilon \to 0} \sup_{v \in \bar{\mathfrak{A}}^\varepsilon(\omega)} \left\{ \frac{1}{\varepsilon} \int_{D_\varepsilon} |v(x', x_{d+1}, \omega) - \bar{v}_0(x', \omega)|^2 dx \right\} = 0 \quad \forall \omega \in \Omega.$$

In particular, this implies that for any $U_0, U_0^* \in \mathcal{H}_\varepsilon$ we have that

$$\lim_{\varepsilon \to 0} \limsup_{t \to +\infty} \left\{ \frac{1}{\varepsilon} \| \bar{\phi}_\varepsilon(t, \theta_{-t}\omega) U_0 - \bar{\phi}_\varepsilon(t, \theta_{-t}\omega) U_0^* \|^2_{L_2(D_\varepsilon)} \right\} = 0 \quad \forall \omega \in \Omega, \tag{17}$$

where we can omit the $\lim_{\varepsilon \to 0}$ under conditions (14)–(16). Thus we obtain the synchronization effect not only at the level of global attractors (see (13)) but also at the level of trajectories in relation (17). We emphasize that this *double* synchronization phenomenon is not true for the deterministic ($K \equiv 0$) counterpart of the problem. In the latter case the global attractor for (8) (without the noise $\dot{W}$) is not a single point when the reaction term $au + f(u)$ has several roots, and thus (17) cannot be true for *all* initial data. In this case we have synchronization at the level of the global attractors only.

Remark 3.3. The statements of Theorem 3.1 deal with the case when the intensity interaction $k(x', \varepsilon)$ between layers is asymptotically strong enough (see condition (6)). However, similarly to [12, 13] we can also consider the case when the limit in (6) is finite by assuming that

$$\lim_{\varepsilon \to 0} \varepsilon^{-1} k(x', \varepsilon) = k(x') \quad \text{strongly in } L_2(\Gamma) \tag{18}$$

for some bounded nonnegative function $k(x') \in L_2(\Gamma)$. In this case the limiting problem for (1)–(4) is a system of two parabolic SPDEs on Γ of the form

$$\frac{\partial}{\partial t} U^i - \nu_i \Delta_{x'} U^i + aU^i + f_i(U^i) + k(x')(-1)^{i+1}(U^1 - U^2) + h_i(x', 0) = \dot{W}(t, x'), \tag{19}$$

where $i = 1, 2$ and $(t; x') \in \mathbb{R}_+ \times \Gamma$, with the Neumann boundary condition on $\partial\Gamma$. Using the same method as for the case (6) in combination with deterministic arguments given in [13] for a particular case of (18), one can prove upper semicontinuity of $\{\bar{\mathfrak{A}}^\varepsilon(\omega)\}$ in the limit $\varepsilon \to 0$ in the case (18). However we will not present the case because (i) our main point of interest is the phenomenon of synchronization, and (ii) under condition (18) synchronization is possible only in some very special cases.

The proof of Theorem 3.1 is given in the remaining sections of the paper. To begin, in section 4 the problem is reformulated in terms of pathwise random PDEs on a scaled domain and appropriate function spaces are introduced. Then we show that (1)–(4) generates an RDS. In section 5 the existence of a random pullback attractor is proved. Then in section 5.1 the limiting dynamics on finite time intervals as $\epsilon \to 0$ is established, and in section 7 the upper continuous dependence of the attractors as $\epsilon \to 0$ is shown. Finally, in section 8 the synchronization of the systems for fixed $\epsilon > 0$ is considered.

4. Generation of an RDS by the two-layer problem.

4.1. Equivalent random PDEs. We introduce the new dependent variables V^i (which are also stochastic processes):

$$V^i(t, x, \omega) := U^i(t, x', x_{d+1}, \omega) - \bar{\eta}(\theta_t\omega, x'), \quad t > 0, \ \ x = (x', x_{d+1}) \in D_{i,\varepsilon}, \ \ i = 1, 2,$$

where $\bar{\eta}(\omega, x')$ is given by (11) after perfection. Let

$$\begin{aligned} h_1(x, \omega) &= -\frac{1}{2}(\nu_1 - \nu_2)\Delta\bar{\eta}(\omega) + h_1(x), \\ h_2(x, \omega) &= \frac{1}{2}(\nu_1 - \nu_2)\Delta\bar{\eta}(\omega) + h_2(x). \end{aligned} \tag{20}$$

Then equations (1)–(4) can be transformed into the pathwise random semilinear parabolic PDEs

$$\partial_t V^i - \nu_i\Delta V^i + aV^i + f_i\left(V^i + \bar{\eta}(\theta_t\omega)\right) + h_i(x, \theta_t\omega) = 0, \quad t > 0, \ x \in D_{i,\varepsilon}, \tag{21}$$

for $i = 1, 2$, with the random initial data

$$V^i(0, x, \omega) = U_0^i(x) - \bar{\eta}(\omega), \qquad x \in D_{i,\varepsilon}, \ \ i = 1, 2. \tag{22}$$

Since the Ornstein–Uhlenbeck process $\bar{\eta}(\theta_t\omega; x')$ does not depend on x_{d+1}, due to (12) we obtain the Neumann boundary conditions

$$\left(\nabla V^i(x), n_i(x)\right) = 0, \qquad x \in \partial D_{i,\varepsilon} \setminus \Gamma, \quad i = 1, 2, \tag{23}$$

on the external part of the boundary of the compound domain D_ε, where n is the outer normal to ∂D_ε. Condition (4) turns into a matching condition on Γ of the form

$$\begin{aligned} \left(-\nu_1\frac{\partial V^1}{\partial x_{d+1}} + k(x', \varepsilon)(V^1 - V^2)\right)\Big|_\Gamma &= 0, \\ \left(\nu_2\frac{\partial V^2}{\partial x_{d+1}} + k(x', \varepsilon)(V^2 - V^1)\right)\Big|_\Gamma &= 0, \end{aligned} \tag{24}$$

which is now pathwise random and homogeneous.

4.2. Scaling and functional spaces. It is convenient to deal with a fixed domain where every equation is defined for $\varepsilon > 0$. Let us introduce the new coordinates $(x, y) \in \mathbb{R}^{d+1}$, as follows:

$$x = x', \quad x \in \Gamma, \quad y = \varepsilon^{-1} x_{d+1}, \quad y \in (-1, 1).$$

In so doing, we transform the domain D_ε into $D = D_1 \cup D_2$, where $D_1 = \Gamma \times (0, 1)$, $D_2 = \Gamma \times (-1, 0)$; the operator $\nabla = (\nabla_{x'}$ and $\partial_{x_{d+1}})$ into $\nabla_\varepsilon = (\nabla_x, \varepsilon^{-1}\partial_y)$; and $\Delta = \Delta_{x'} + \partial^2_{x_{d+1}}$ into $\Delta_\varepsilon = \Delta_x + \varepsilon^{-2}\partial_{yy}$. Problem (21)–(24) takes the form

$$\partial_t v^i - \nu_i \Delta_\varepsilon v^i + a v^i + f_i(v^i + \bar{\eta}) + h_i^\varepsilon(x, y, \theta_t \omega) = 0, \quad t > 0, \ (x, y) \in D_i, \tag{25}$$

for $i = 1, 2$, with the initial data

$$v^i(0, x, y) = V_0^i(x, y), \quad (x, y) \in D_i, \quad i = 1, 2, \tag{26}$$

and the boundary conditions

$$\left.\frac{\partial v^i}{\partial n_i}\right|_{\partial D_i \setminus \Gamma} = 0, \quad i = 1, 2, \tag{27}$$

$$\left.\left(\nu_i \frac{\partial v_i}{\partial y} - \varepsilon k(x, \varepsilon)(v_1 - v_2)\right)\right|_{y=0} = 0, \quad i = 1, 2. \tag{28}$$

Here $h_i^\varepsilon(x, y, \omega) = h_i(x, \varepsilon y, \omega)$ and n_i is the outward normal to the boundary ∂D_i. A solution $V(t, x', x_{d+1})$ to problem (21)–(24) is expressed in terms of a solution $v(t, x, y)$ to problem (25)–(28) by the formula $V(t, x', x_{d+1}) = v(t, x', \varepsilon^{-1} x_{d+1})$.

Let us introduce the space

$$\mathcal{H} = L^2(D_1) \oplus L^2(D_2) \simeq L^2(D)$$

endowed with the norm $\|u\|^2 \equiv \|u_1\|^2_{L^2(D_1)} + \|u_2\|^2_{L^2(D_2)}$, where $u = (u_1; u_2)$, $u_i \equiv u|_{D_i}$, and let us define a family of Sobolev spaces

$$\mathcal{H}^1_\varepsilon = H^1(D_1) \oplus H^1(D_2), \quad \varepsilon \in (0, 1],$$

endowed with the norm

$$\|u\|^2_{1,\varepsilon} \equiv \sum_{i=1}^{2} \left(\|u_i\|^2_{H^1(D_i)} + \varepsilon^{-2} \|\partial_y u_i\|^2_{L^2(D_i)} \right).$$

Every element $v \in H^1(\Gamma) \oplus H^1(\Gamma)$ can be extended naturally to an element $u \in \mathcal{H}^1_\varepsilon$ by the formula $u_i(x, y) \equiv v_i(x)$, $(x, y) \in D_i$, $i = 1, 2$; in what follows, this will be done without further comment.

4.3. Abstract representation. Now we represent problem (25)–(27) in the abstract form. To do this we first consider the bilinear form

$$\begin{aligned} a_\varepsilon(u, v) \quad &= \sum_{i=1}^{2} \nu_i \left[(\nabla_x u_i, \nabla_x v_i)_{L^2(D_i)} + \frac{1}{\varepsilon^2} (\partial_y u_i, \partial_y v_i)_{L^2(D_i)} \right] + a \cdot (u, v)_{\mathcal{H}} \\ &+ \frac{1}{\varepsilon} \int_\Gamma k(x, \varepsilon)(u_1(x, 0) - u_2(x, 0))(v_1(x, 0) - v_2(x, 0))\, dx, \end{aligned}$$

defined on the elements $u = (u_1; u_2)$, $v = (v_1; v_2)$ of the space $\mathcal{H}^1_\varepsilon = H^1(D_1) \oplus H^1(D_2)$. One can show that $a_\varepsilon(u, v)$ is a closed symmetric form in $\mathcal{H}$ possessing the property

$$c_0 \sum_{i=1,2} \|u\|^2_{H^1(D_i)} \le c_1 \|u\|^2_{1,\varepsilon} \le a_\varepsilon(u, u), \quad u \in \mathcal{H}^1_\varepsilon. \tag{29}$$

Here and in what follows we drop the subscript ε in constants which can be chosen independently of $\varepsilon \in (0, 1]$. Therefore, there exists a unique positive self-adjoint operator A_ε such that $\mathcal{D}(A_\varepsilon) \subset \mathcal{H}^1_\varepsilon$ and

$$a_\varepsilon(u, v) = (A_\varepsilon u, v)_{\mathcal{H}}, \quad u \in \mathcal{D}(A_\varepsilon),\ v \in \mathcal{H}^1_\varepsilon.$$

It can be shown that

$$\mathcal{D}(A_\varepsilon) = \big\{u \in H^2(D_1) \oplus H^2(D_2)\ :\ u \text{ satisfies (27) and (28)}\big\}$$

and also that

$$A_\varepsilon u = (-\nu_1 \Delta_\varepsilon u_1 + a u_1,\ -\nu_2 \Delta_\varepsilon u_2 + a u_2), \quad u = (u_1, u_2) \in \mathcal{D}(A_\varepsilon).$$

Moreover, $\mathcal{D}(A^{1/2}_\varepsilon) = \mathcal{H}^1_\varepsilon$, $a_\varepsilon(u, u) = \|A^{1/2}_\varepsilon u\|^2$. For more details concerning the operator A_ε we refer to [13].

Now we can rewrite the pathwise random PDE in problem (25)–(28) in the abstract form

$$\frac{d}{dt} v + A_\varepsilon v = B(v, \theta_t \omega), \quad v|_{t=0} = v_0, \tag{30}$$

in the space $\mathcal{H}$, where

$$B(v, \omega) = \begin{cases} -f_1(v^1 + \bar\eta(\omega)) - h_1(x, \varepsilon y, \omega), & y > 0, \\ -f_2(v^2 + \bar\eta(\omega)) - h_2(x, \varepsilon y, \omega), & y < 0. \end{cases}$$

4.4. Generation of an RDS. By the same method as in [25] (see also [30, Chap. 3]) one can prove that there exists a deterministic constant M such that the nonlinear mapping $A_\varepsilon - B(\cdot, \omega) + M$ is a maximal monotone operator on $\mathcal{D}(A_\varepsilon)$. This observation makes it possible (some details can be found in [8, Chap. 15] for the general nonautonomous case) to prove that for each $\omega \in \Omega$ and $v_0 \in \mathcal{H}$ on any time interval $[0, T]$ there exists a unique weak solution $v(t, \omega)$ to (30) from the class

$$L_{p+1}(0, T; L_{p+1}(D)) \cap L_2(0, T; \mathcal{H}^1_\varepsilon) \cap C(0, T; \mathcal{H}).$$

Since this solution can be constructed as a limit of the corresponding Galerkin approximations, the mapping $(t; \omega) \mapsto v(t, \omega)$ is measurable. Moreover, it is easy to derive from the uniqueness property that the mapping $\phi_\varepsilon(t, \omega) : \mathcal{H} \mapsto \mathcal{H}$ defined by the relation $\phi_\varepsilon(t, \omega) v_0 = v(t, \omega)$, where $v(t, \omega)$ solves (30), satisfies the cocycle property. Thus (30) generates an RDS.

Now using inverse transformation we define the cocycle $\bar\phi_\varepsilon$ for problem (1)–(4) by the formula

$$\bar\phi_\varepsilon(t, \omega) = R^{-1}_\varepsilon(\theta_t \omega) \circ \phi_\varepsilon(t, \omega) \circ R_\varepsilon(\omega),$$

where $R_\varepsilon(\omega) : L_2(D_\varepsilon) \mapsto L_2(D)$ is an affine random mapping of the form

$$[R_\varepsilon(\omega) U](x, y) = U(x, \varepsilon y) - \bar\eta(\omega), \quad U \in L_2(D_\varepsilon).$$

This proves the first statement in Theorem 3.1.

It is clear that $R_\varepsilon(\omega)$ maps tempered random sets in $L_2(D_\varepsilon)$ into tempered sets in $L_2(D)$. Therefore all other statements of Theorem 3.1 can be easily reformulated as statements concerning the RDS $(\theta, \phi_\varepsilon)$ generated by the random evolution equation in (30). In our further considerations we deal with this RDS $(\theta, \phi_\varepsilon)$.

5. Random pullback attractors. In this section we prove the existence of a random pullback attractor for problem (25)–(28) for every *fixed* $\varepsilon \in (0, 1]$ and also for the limiting problem (8).

5.1. The case $\varepsilon > 0$. We first want to emphasize that we do not use any information concerning the behavior of the intensity $k(x', \varepsilon)$ as $\varepsilon \to 0$, and hence our results in this subsection cover both of the cases (6) and (18).

Our main result in this section is the following assertion.

PROPOSITION 5.1. *In the space $\mathcal{H}$ the RDS $(\theta, \phi_\varepsilon)$ generated by problem* (25)–(28) *possesses a compact pullback attractor $\widehat{\mathfrak{A}}^\varepsilon$ which belongs to the space $\mathcal{H}^1_\varepsilon$. Moreover, there exists a tempered random variable $R(\omega)$, which does not depend on ε, such that*

$$\mathfrak{A}^\varepsilon(\omega) \subset \left\{ v \in \mathcal{H}^1_\varepsilon \ : \ a_\varepsilon(v, v) + \|v\|^{p+1}_{L_{p+1}(D)} \le R^2(\omega) \right\}, \quad \omega \in \Omega. \tag{31}$$

We split the proof into several lemmata which are also important for the limit transition on finite time intervals.

LEMMA 5.2 (pullback dissipativity). *The RDS $(\theta, \phi_\varepsilon)$ is pullback dissipative in $\mathcal{D}$; i.e., there exists a tempered random variable $R(\omega) > 0$ such that for any random set $\widehat{D}$ from $\mathcal{D}$ we can find $t_0(\omega, \widehat{D}) > 0$ for which*

$$\|\phi_\varepsilon(t, \theta_{-t}\omega)U\|_{\mathcal{H}} \le R(\omega) \quad \forall\ U \in D(\theta_{-t}\omega),\ t \ge t_0(\omega, \widehat{D}).$$

Thus the random ball $B_0(\omega) = \{U \in \mathcal{H} : \|U\|_{\mathcal{H}} \le R(\omega)\}$ is pullback absorbing. This ball is also forward invariant and absorbing if we take

$$R^2(\omega) = c_1 \int_{-\infty}^0 e^{c_0 \tau} \left(1 + \|\bar\eta(\theta_\tau\omega)\|^{p+1}_{L_{p+1}(\Gamma)} + \|\bar\eta(\theta_\tau\omega)\|^2_{H^1(\Gamma)} \right) d\tau,$$

with appropriate $c_0 > 0$ and $c_1 > 0$ independent of $\varepsilon \in (0, 1]$.

Proof. The calculations below are formal, but can be justified by considering Galerkin approximations.

Multiplying (25) by v^i in $L_2(D_i)$ for $i = 1, 2$, after some calculations we obtain that

$$\frac{1}{2}\frac{d}{dt}\|v\|^2_{\mathcal{H}} + a_\varepsilon(v, v) + \sum_{i=1,2} \left[\int_{D_i} f_i(v^i + \bar\eta) v^i dx + (h^\varepsilon, v^i)_{L_2(D_i)} \right] = 0. \tag{32}$$

From (5) we have that

$$\begin{aligned}
(f_i(v^i + \bar\eta), v^i) &= \int_{D_i} f(v^i) v^i dx + \int_{D_i} \left[\int_0^1 f_i'(v^i + \lambda\bar\eta) d\lambda \right] \bar\eta v^i dx \\
&\ge a_0 \|v^i\|^{p+1}_{L_{p+1}(D_i)} - c_1 \int_{D_i} \left(1 + |v^i|^{p-1} + |\bar\eta|^{p-1} \right) |\bar\eta| |v^i| dx - c_2 \\
&\ge \frac{a_0}{2} \|v^i\|^{p+1}_{L_{p+1}(D_i)} - b_0 \left(1 + \|\bar\eta\|^{p+1}_{L_{p+1}(\Gamma)} \right) \qquad (33)
\end{aligned}$$

and from (20) and (29) we also have that

$$\sum_{i=1,2}(h^\varepsilon, v^i)_{L_2(D_i)} \le C\left(\|\bar\eta\|_{H^1(\Gamma)} + \sum_{i=1,2}\|h\|_{H^1(D_i)}\right)[a_\varepsilon(v,v)]^{1/2}. \tag{34}$$

Now from (32)–(34) we obtain that

$$\frac{d}{dt}\|v\|^2_{\mathcal{H}} + a_\varepsilon(v,v) + a_0\|v\|^{p+1}_{L_{p+1}(D)} \le R_0^2(\theta_t\omega), \tag{35}$$

where

$$R_0^2(\omega) = c\left(1 + \|\bar\eta(\omega)\|^{p+1}_{L_{p+1}(\Gamma)} + \|\bar\eta(\omega)\|^2_{H^1(\Gamma)}\right). \tag{36}$$

Since $a_\varepsilon(v,v) \ge c_0\|v\|^2_{\mathcal{H}} + \frac{1}{2}a_\varepsilon(v,v)$, by differentiating $e^{\nu_* t}\|v\|^2_{\mathcal{H}}$, taking into account (35), and integrating, we have that

$$\|v(t)\|^2_{\mathcal{H}} + \int_0^t e^{-\nu_*(t-\tau)}V^0_\varepsilon(v(\tau))d\tau \le \|v_0\|^2_{\mathcal{H}}e^{-\nu_* t} + \int_0^t e^{-\nu_*(t-\tau)}R_0^2(\theta_\tau\omega)d\tau, \tag{37}$$

for any $0 < \nu_* \le c_0$, where $R_0(\omega)$ is given by (36) and

$$V^0_\varepsilon(v) = \frac{1}{2}a_\varepsilon(v,v) + a_0\|v\|^{p+1}_{L_{p+1}(D)}. \tag{38}$$

This allows us to complete the proof of Lemma 5.2. □

LEMMA 5.3 (compact absorbing set). *For each $\varepsilon \in (0,1]$ there exists a compact, forward invariant tempered absorbing set.*

Proof. Multiplying (25) by $\partial_t v^i$ in $L_2(D_i)$ we find that

$$\begin{aligned}&\partial_t\Psi_\varepsilon(v(t)) + \|\partial_t v(t)\|^2_{\mathcal{H}}\\ &+ \sum_{i=1,2}\int_{D_i}\left[f(v^i+\bar\eta) - f(v^i)\right]\partial_t v^i dxdy + \int_D h^\varepsilon\partial_t v^i dxdy = 0,\end{aligned} \tag{39}$$

where

$$\Psi_\varepsilon(u) = \frac{1}{2}a_\varepsilon(u,u) + \sum_{i=1}^2\int_{D_i}F_i(u^i)\,dxdy, \quad u = (u^1;u^2) \in \mathcal{H}^1_\varepsilon. \tag{40}$$

Here $F_i(u) = \int_0^u f_i(\xi)d\xi$. It is clear from the assumptions concerning f_i that

$$\begin{aligned}&\left|\sum_{i=1,2}\int_{D_i}\left[f(v^i+\bar\eta) - f(v^i)\right]\partial_t v^i dxdy\right|\\ &\le c\sum_{i=1,2}\int_{D_i}\left|f(v^i+\bar\eta) - f(v^i)\right|^2 dxdy + \frac{1}{4}\|\partial_t v(t)\|^2_{\mathcal{H}}\\ &\le c_1 + c_2\int_D|v|^{p+1}\,dxdy + c_3\left[|\bar\eta|^{p+1}_{L_{p+1}(\Gamma)} + |\bar\eta|^{p_*}_{L_{p_*}(\Gamma)}\right] + \frac{1}{4}\|\partial_t v(t)\|^2_{\mathcal{H}},\end{aligned}$$

where $p_* = 2(p+1)/(3-p)$. We also have that

$$\left|\int_D h^\varepsilon\partial_t v^i dxdy\right| \le c_1 + c_2\|\bar\eta\|^2_{H^2(\Gamma)} + \frac{1}{4}\|\partial_t v(t)\|^2_{\mathcal{H}}$$

Therefore from (39) we have that

$$\partial_t \Psi_\varepsilon(v(t)) + \frac{1}{2}\|\partial_t v(t)\|^2_{\mathcal{H}} \leq c_1 + c_2\|v\|^{p+1}_{L_{p+1}(D)} + c_3\left(\|\bar\eta\|^2_{H^2(\Gamma)} + |\bar\eta|^{p+1}_{L_{p+1}(\Gamma)} + |\bar\eta|^{p_*}_{L_{p_*}(\Gamma)}\right). \tag{41}$$

Consequently, choosing positive constants b_0 and b_1 in an appropriate way one can see that

$$V_\varepsilon(u) := b_0\|u\|^2_{\mathcal{H}} + \Psi_\varepsilon(u) + b_1 \tag{42}$$

with Ψ_ε given by (40) satisfies the relations

$$c_0 V^0_\varepsilon(v) \leq V_\varepsilon(v) \leq c_1\left[1 + V^0_\varepsilon(v)\right] \tag{43}$$

with $V^0_\varepsilon(v)$ given by (38). Moreover, due to (35) we can choose b_0 and b_1 such that

$$\frac{d}{dt}V_\varepsilon(v) + \gamma V_\varepsilon(v) + \frac{1}{2}\|\partial_t v_\varepsilon(t)\|^2_{\mathcal{H}} \leq R_1^2(\theta_t\omega), \tag{44}$$

with positive γ, where

$$R_1^2(\omega) = c\left(1 + \|\bar\eta(\omega)\|^{p+1}_{L_{p+1}(\Gamma)} + |\bar\eta(\omega)|^{p_*}_{L_{p_*}(\Gamma)} + \|\bar\eta(\omega)\|^2_{H^2(\Gamma)}\right), \quad p_* = \frac{2p+2}{3-p}. \tag{45}$$

We note that $R_1(\omega)$ is a tempered random variable because $t \mapsto \bar\eta(\theta_t\omega)$ is a tempered process with values in $H^2(\Gamma) \cap C(\overline{\Gamma})$. From (44) we have that

$$V_\varepsilon(v(t)) \leq e^{-\gamma(t-s)}V_\varepsilon(v(s)) + \int_s^t e^{-\gamma(t-\tau)}R_1^2(\theta_\tau\omega)d\tau, \quad t \geq s. \tag{46}$$

By (43) we also have

$$V^0_\varepsilon(v(t)) \leq c_1 e^{-\gamma(t-s)}V^0_\varepsilon(v(s)) + c_2\int_s^t e^{-\gamma(t-\tau)}R_1^2(\theta_\tau\omega)d\tau, \quad t \geq s.$$

Therefore using (37) after integration with respect to s over the interval $[0,t]$ we obtain

$$V^0_\varepsilon(v(t)) \leq \frac{c_1}{t}\|v_0\|^2_{\mathcal{H}}e^{-\gamma_* t} + c_2\left(1 + \frac{1}{t}\right)\int_0^t e^{-\gamma_*(t-\tau)}R_1^2(\theta_\tau\omega)d\tau, \quad t > 0, \tag{47}$$

for some $0 < \gamma_* \leq \gamma$. Relations (46) and (47) makes it possible to conclude that there exists a tempered random variable $R_*(\omega)$ such that the set

$$\mathfrak{B}(\omega) = \left\{v : V_\varepsilon(v) \leq R_*^2(\omega)\right\} \tag{48}$$

is forward invariant and absorbing. It is clear that $\mathfrak{B}(\omega)$ is compact in $\mathcal{H}$ for each $\omega \in \Omega$. Moreover, $R_*^2(\omega)$ does not depend on ε. □

Completion of the proof of Proposition 5.1. The proof follows from Theorem 2.3 and Lemmata 5.2 and 5.3. Relation (31) follows from (47), (48), and properties of the functionals V^0_ε and V_ε given in (38), (42), and (43).

Remark 5.4. It also follows from (44) and (37) that

$$\int_0^t \tau e^{-\gamma_*(t-\tau)}\|\partial_t v_\varepsilon(\tau)\|^2_{\mathcal{H}} d\tau \le c_1\|v_0\|^2_{\mathcal{H}} e^{-\gamma_* t} + c_2\int_0^t (1+\tau)e^{-\gamma_*(t-\tau)} R_1^2(\theta_\tau\omega)d\tau, \tag{49}$$

for all $t \ge 0$, where $\gamma_* > 0$. Below we will also need the next lemma.

LEMMA 5.5. *For any initial data $v, v_* \in \mathcal{H}$ we have the estimate*

$$\|\phi_\varepsilon(t,\omega)v - \phi_\varepsilon(t,\omega)v_*\|_{\mathcal{H}} \le c_1 e^{c_2 t}\|v - v_*\|_{\mathcal{H}}, \quad \omega \in \Omega, \tag{50}$$

where c_1 and c_2 do not depend on $\omega \in \Omega$ and $\varepsilon \in (0,1]$.

Proof. We use the same method as in Lemma 5.2 by considering the difference of two solutions and relying on the property

$$\big(f_i(v^i + \bar\eta) - f_i(v^i_* + \bar\eta)\big)\,(v^i - v^i_*) \ge -c_0|v^i - v^i_*|^2,$$

where c_0 does not depend on ω and ε. ∎

5.2. Limiting system. The same change of unknown variable $U = v + \bar\eta$ transforms equation (8) into the following random PDE on Γ:

$$\begin{cases} \partial_t v - \nu\Delta v + av + f(v + \bar\eta(\theta_t\omega)) + h(x') = 0, \quad t > 0,\ x' \in \Gamma, \\ \\ \frac{\partial v}{\partial n}\big|_{\partial\Gamma} = 0,\ v|_{t=0} = v_0, \end{cases} \tag{51}$$

where ν, $f(v)$, and h are given by (9). The same argument as in section 4 allows us to prove that problem (51) generates an RDS (θ, ϕ_0) in the space $L_2(\Gamma)$ and thus to establish Theorem 3.1(2).

The following assertion states the existence of a pullback attractor for this RDS (θ, ϕ_0).

PROPOSITION 5.6. *In the space $L_2(\Gamma)$, problem* (51) *generates an RDS (θ, ϕ_0) possessing a compact pullback attractor $\{\mathfrak{A}^0(\omega)\}$ which belongs to the space $H^1(\Gamma)$. If the correlation operator K possesses the properties* (i) *$Kh = 0$ if and only if $h = 0$ and* (ii) *the image of K is dense in $L_2(\Gamma)$, then the attractor $\{\mathfrak{A}^0(\omega)\}$ is a singleton; i.e., there exists a tempered random variable $v_0(\omega)$ with values in $H^1(\Gamma)$ such that $\mathfrak{A}^0(\omega) = \{v_0(\omega)\}$ for all $\omega \in \Omega$.*

Proof. To prove the existence of the attractor we argue exactly as in Proposition 5.1 and we do not repeat it again.

As for the second part, we first note that the RDS (θ, ϕ_0) is *monotone*; i.e., the property $v(x) \le v_*(x)$ for almost all $x \in \Gamma$ implies that

$$[\phi_0(t,\omega)v](x) \le [\phi_0(t,\omega)v_*](x) \quad \text{for almost all} \quad x \in \Gamma,$$

for all $t > 0$, and for $\omega \in \Omega$. This monotonicity property can be established by the standard (pathwise) argument (see, e.g., [31]). We also refer to [10] for a general discussion of monotone RDSs. Our next step is to apply a result from [15] which states that, under some conditions, the global pullback attractor of a monotone RDS consists of a single random equilibrium. The main hypothesis in [15] is the weak convergence of distributions of the process $t \mapsto \phi_0(t,\omega)v$ to some limiting probability measure. In our case we can guarantee this property because the noise $\dot W$ is nondegenerate in the phase space of the system (θ, ϕ_0). We refer to [15, subsection 4.5] for details. ∎

Propositions 5.1 and 5.6 imply Theorem 3.1(4).

Remark 5.7. Although it is possible to prove that the RDS $(\theta, \phi_\varepsilon)$ generated by problem (25)–(28) is also monotone, we cannot apply the result from [15] to prove that $\hat{\mathfrak{A}}^\varepsilon$ is a single equilibrium. The point is that the noise $\dot{W}$ is nondegenerate in $L_2(\Gamma)$ (the phase space of the system (θ, ϕ_0)), but it is *degenerate* in $\mathcal{H} = L_2(D)$ (the phase space for $(\theta, \phi_\varepsilon)$), and hence we cannot guarantee the weak convergence of distributions of the process $t \mapsto \phi_\varepsilon(t, \omega)U_0$. Thus the pullback attractor $\hat{\mathfrak{A}}^\varepsilon$ may contain more than one equilibrium. The same conclusion is valid for problem (19). One can prove that (19) generates a monotone RDS with a compact pullback attractor, but to conclude that this attractor is a random equilibrium we need the nondegeneracy of the noise in $L_2(\Gamma) \times L_2(\Gamma)$, which is obviously not true for this case.

Remark 5.8. 1. It is clear from the argument in the proof of Lemma 5.5 that

$$\|\phi_0(t, \omega)v - \phi_0(t, \omega)v_*\|_{L_2(\Gamma)} \le c_1 e^{c_2 t}\|v - v_*\|_{L_2(\Gamma)}, \quad \omega \in \Omega, \tag{52}$$

for some constants c_1 and c_2 independent of ω, where $v, v_* \in L_2(\Gamma)$.

2. Since $L_2(\Gamma)$ can be embedded naturally into $L_2(D) \sim \mathcal{H}$ as the subspace of functions independent of y, we can consider the cocycle ϕ_0 as a mapping from $L_2(\Gamma)$ into $\mathcal{H}$. Therefore we can compare it with ϕ_ε. Below we also consider the image $\widetilde{\mathfrak{A}}^0(\omega)$ of $\mathfrak{A}^0(\omega)$ under this embedding.

6. Limit transition on finite time intervals. Our main result in this section is the following theorem, which implies the third statement in Theorem 3.1.

THEOREM 6.1. *For any time interval we have that*

$$\lim_{\varepsilon \to 0} \sup_{t \in [\delta, T]} \|\phi_\varepsilon(t, \omega)v - \phi_0(t, \omega)v_*\|_{\mathcal{H}} = 0 \quad \forall \delta \in (0, T), \tag{53}$$

where $v_* = \langle v \rangle := \frac{1}{2}\int_{-1}^{1} v(x, y)dy$. *If* v *does not depend on* y, *i.e.,* $v = v_*$, *then*

$$\lim_{\varepsilon \to 0} \sup_{t \in [0, T]} \|\phi_\varepsilon(t, \omega)v - \phi_0(t, \omega)v_*\|_{\mathcal{H}} = 0. \tag{54}$$

Proof. Let $w_\varepsilon(t) = \phi_\varepsilon(t, \omega)v$. It follows from (37), (47), and (49) that

$$\sup_{t \in [0, T]} \sum_{i=1,2} \|w^i_\varepsilon(t)\|^2_{L_2(D_i)} + \int_0^T \|w^i_\varepsilon(t)\|^2_{H^1(D_i)}dt \le C_T(\omega), \tag{55}$$

and, for every $\delta > 0$,

$$\sup_{t \in [\delta, T]} \sum_{i=1,2} \|w^i_\varepsilon(t)\|^2_{H^1(D_i)} + \sum_{i=1,2} \int_\delta^T \|\partial_t w^i_\varepsilon(t)\|^2_{L_2(D_i)}dt \le C_{T,\delta}(\omega), \tag{56}$$

$$\frac{1}{\varepsilon^2}\left[\sup_{t \in [\delta, T]} \sum_{i=1,2} \|\partial_y w^i_\varepsilon(t)\|^2_{L_2(D_i)} + \sum_{i=1,2} \int_0^T \|\partial_y w^i_\varepsilon(t)\|^2_{L_2(D_i)}dt\right] \le C_{T,\delta}(\omega). \tag{57}$$

Moreover, we have that

$$\sup_{t \in [\delta, T]} \int_\Gamma \frac{k(x', \varepsilon)}{\varepsilon}|w^1_\varepsilon(t) - w^2_\varepsilon(t)|^2dx' + \int_0^T dt \int_\Gamma \frac{k(x', \varepsilon)}{\varepsilon}|w^1_\varepsilon(t) - w^2_\varepsilon(t)|^2dx' \le C_{T,\delta}(\omega) \tag{58}$$

for all intervals $[0,T]$ and $\varepsilon \in (0,1]$. Therefore, using relations (55)–(57) and Aubin's compactness theorem we can conclude that there exist a pair of functions

$$u_i \in C(\delta, T; L^2(\Gamma)) \cap L^\infty(\delta, T; H^1(\Gamma)), \quad i = 1,2, \quad \forall \delta > 0$$

and a sequence $\{\varepsilon_n\}$ such that

$$\lim_{n\to\infty} \sum_{i=1,2} \sup_{t\in[\delta,T]} \|w^i_{\varepsilon_n}(t) - u_i(t)\|_{L_2(D_i)} = 0. \tag{59}$$

Moreover, we also have weak convergence in $L_2(0,T;H^1(D))$. One can also see from (58) and (6) that $u_1(t) = u_2(t) \equiv u(t)$ on the set Γ. Considering a variational form of (25)–(28), one can show that $u(t)$ solves problem (51). The corresponding argument is exactly the same as in [13] for the deterministic case and therefore we do not give details here. Thus (53) follows from (59) and from the uniqueness theorem for (51).

To prove (54) we first consider $v \equiv v_*$ from the space $H^1(\Gamma) \cap L_{p+1}(\Gamma)$. In this case relying on (46) with $s = 0$ and using the fact that $V_\varepsilon(v)$ does not depend on ε for this choice of v, we can easily prove estimates (56) and (57) with $\delta = 0$. Thus the same argument as above gives (54) for $v \equiv v_*$ from $H^1(\Gamma) \cap L_{p+1}(\Gamma)$. To obtain (54) for $v_* \in L_2(\Gamma)$ we use an appropriate approximation procedure and relations (50) and (52). □

Remark 6.2. By a standard argument we can prove that (53) and (54) hold uniformly with respect to v in every compact set.

Remark 6.3. Since the arguments given in Lemmata 5.2 and 5.3 do not depend on the behavior of $k(x,\varepsilon)$ as $\varepsilon \to 0$, the estimates in (55)–(58) hold for *both* cases (6) and (18). Thus, in the latter case, we can also conclude from (55)–(57) that w^1_ε and w^2_ε converge to some functions u^1 and u^2 defined on Γ. However, in that case we cannot prove that u^1 and u^2 are the same because under condition (18) estimate (58) does not lead to the conclusion. In the case (18) the same arguments as in [12, 13] give us the convergence of $\bar{\phi}_\varepsilon(t,\omega)$ generated by (1)–(4) to the cocycle generated by (19).

7. Upper semicontinuity of attractors. In this section we prove the following assertion, which is our first result on synchronization.

THEOREM 7.1. *Let $\{\mathfrak{A}^\varepsilon(\omega)\}$ be the global random pullback attractor for the RDS $(\theta, \phi_\varepsilon)$ generated by* (25)–(28). *Then*

$$\lim_{\varepsilon\to 0} \sup \left\{ \operatorname{dist}_{\mathcal{H}} \left(u, \widetilde{\mathfrak{A}}^0(\omega) \right) \,:\, u \in \mathfrak{A}^\varepsilon(\omega) \right\} = 0 \quad \forall \omega \in \Omega, \tag{60}$$

where $\widetilde{\mathfrak{A}}^0(\omega) = \left\{ J(v) \,:\, v \in \mathfrak{A}^0(\omega) \right\} \subset \mathcal{H}$. Here $\{\mathfrak{A}^0(\omega)\}$ is the random pullback attractor for the RDS (θ, ϕ_0) and $J : L_2(\Gamma) \mapsto L_2(D) = \mathcal{H}$ is the natural embedding operator.

Proof. Assume that (60) does not hold for some $\omega \in \Omega$. Then there exist a sequence $\{\varepsilon_n\}$ with $\varepsilon_n \to 0$ and a sequence $u_n \in \mathfrak{A}^{\varepsilon_n}(\omega)$ such that

$$\operatorname{dist}_{\mathcal{H}}(u_n, \widetilde{\mathfrak{A}}^0(\omega)) \geq \delta > 0 \quad \forall\, n = 1,2,\ldots. \tag{61}$$

By the invariance property of the attractor $\mathfrak{A}^{\varepsilon_n}(\omega)$, for every $t > 0$ there exists $v^t_n \in \mathfrak{A}^{\varepsilon_n}(\theta_{-t}\omega)$ such that $u_n = \phi_{\varepsilon_n}(t, \theta_{-t}\omega) v^t_n$. Since $\mathfrak{A}^{\varepsilon_n}(\omega)$ is compact and estimate (31) holds, we can assume that there exist u_* and v^t_* in $H^1(D_1) \oplus H^1(D_2)$ such that

$$\lim_{n\to\infty} \|u_n - u_*\|_{\mathcal{H}} = 0, \quad \lim_{n\to\infty} \|v^t_n - v^t_*\|_{\mathcal{H}} = 0. \tag{62}$$

As in the proof of Theorem 6.1 one can see that

$$u_* = \tilde{u} \oplus \tilde{u}, \quad v_*^t = \tilde{v}^t \oplus \tilde{v}^t,$$

where $\tilde{u}, \tilde{v}^t \in H^1(\Gamma)$. Therefore, if we show that $\tilde{u} \in \mathfrak{A}^0(\omega)$, then we obtain a contradiction to (61).

It follows from Lemma 5.5 and Theorem 6.1 that

$$\tilde{u} = \phi_0(t, \theta_{-t}\omega)\tilde{v}^t.$$

However, it follows from (31) and (62) that $\tilde{v}^t \in B_0(\theta_{-t}\omega)$, where

$$B_0(\omega) = \left\{ v \in H^1(\Gamma) \,:\, \|v\|_{H^1(\Gamma)} \le \tilde{R}(\omega) \right\},$$

where $\tilde{R}(\omega)$ is a tempered random variable. Thus we have that

$$\tilde{u} \in \phi_0(t, \theta_{-t}\omega)B_0(\theta_{-t}\omega) \quad \text{for every} \quad t > 0.$$

Since $\phi_0(t, \theta_{-t}\omega)B_0(\theta_{-t}\omega) \to \mathfrak{A}^0(\omega)$ as $t \to \infty$, this implies that $\tilde{u} \in \mathfrak{A}^0(\omega)$. □

Theorem 3.1(5) follows from Theorem 7.1.

Remark 7.2. In the case (18), similarly to the deterministic case (see [12, 13]), we can prove the upper convergence of the pullback attractors $\hat{\mathfrak{A}}^\varepsilon$ to the corresponding object for the RDS generated by (19). We also refer to [6] and to the references therein for a general study of upper semicontinuity of random and nonautonomous attractors.

8. Synchronization for fixed $\varepsilon > 0$. Now we consider the case when the equations are the same in both domains; i.e., we assume that relations (14), (15), and (16) hold.

Under conditions (14) the cocycle ϕ_ε has a deterministic forward *invariant* subspace $\mathcal{L}$ in $\mathcal{H}$ consisting of functions which are independent of the variable y, i.e.,

$$\mathcal{L} = \{u(x,y) \in L_2(D) \,:\, u(x,y) \equiv u(x,0) \equiv v \in L_2(\Gamma)\}.$$

It is clear that $\phi_\varepsilon(t,\omega)\mathcal{L} \subset \mathcal{L}$ and $\phi_\varepsilon(t,\omega) \equiv \phi_0(t,\omega)$ on $\mathcal{L}$.

THEOREM 8.1. *Under conditions* (14), (15), *and* (16) *there exists* $\varepsilon_0 > 0$ *such that for all* $\varepsilon \in (0, \varepsilon_0]$ *the global random pullback attractor* $\mathfrak{A}^\varepsilon(\omega)$ *for* $(\theta, \phi_\varepsilon)$ *has the form*

$$\mathfrak{A}^\varepsilon(\omega) \equiv \widetilde{\mathfrak{A}}^0(\omega) = \left\{ J(v) \,:\, v \in \mathfrak{A}^0(\omega) \right\} \subset \mathcal{H}, \tag{63}$$

where $J : L_2(\Gamma) \mapsto L_2(D) = \mathcal{H}$ *is the natural embedding operator and* $\mathfrak{A}^0(\omega)$ *is the random pullback attractor for the RDS* (θ, ϕ_0).

Proof. Let P be the orthoprojector in $\mathcal{H}$ onto $\mathcal{L}$. This operator has the form

$$(Pu)(x,y) = \frac{1}{2}\int_{-1}^{1} u(x,\xi)d\xi, \quad u \in \mathcal{H} \sim L_2(D).$$

Let $Q = 1 - P$. Both of the operators P and Q map the domain $\mathcal{D}(A_\varepsilon)$ of the operator A_ε into itself and commute with A_ε. Therefore it follows from (30) that Qv_ε satisfies the equation

$$\frac{d}{dt}Qv_\varepsilon + A_\varepsilon Qv_\varepsilon = QB(v_\varepsilon, \theta_t\omega), \quad Qv|_{t=0} = Qv_0. \tag{64}$$

Multiplying this equation by Qv_ε we obtain

$$\frac{1}{2}\frac{d}{dt}\|Qv_\varepsilon\|^2_{\mathcal{H}} + a_\varepsilon(Qv_\varepsilon, Qv_\varepsilon)_{\mathcal{H}} = (QB(v_\varepsilon, \theta_t\omega), Qv_\varepsilon)_{\mathcal{H}}. \tag{65}$$

From (15) we have that

$$\begin{aligned}(QB(v_\varepsilon, \theta_t\omega), Qv_\varepsilon)_{\mathcal{H}} &= \int_D \left[f(v_\varepsilon(x,y)) - \frac{1}{2}\int_{-1}^1 f(v_\varepsilon(x,\xi))d\xi\right] Qv_\varepsilon(x,y)dxdy \\ &\le \frac{L}{2}\int_D \int_{-1}^1 |v_\varepsilon(x,y) - v_\varepsilon(x,\xi)|\, Qv_\varepsilon(x,y)d\xi dxdy \\ &\le \frac{L}{\sqrt{2}}\left[\int_\Gamma dx \int_{-1}^1 dy \int_{-1}^1 d\xi\, |v_\varepsilon(x,y) - v_\varepsilon(x,\xi)|^2\right]^{1/2} \|Qv_\varepsilon\|_{\mathcal{H}}.\end{aligned}$$

If we add and subtract Pv_ε in the expression under the integral, then we easily arrive at the relation

$$(QB(v_\varepsilon, \theta_t\omega), Qv_\varepsilon)_{\mathcal{H}} \le 2L\|Qv_\varepsilon\|^2_{\mathcal{H}}. \tag{66}$$

Thus from (65) we obtain that

$$\frac{1}{2}\frac{d}{dt}\|Qv_\varepsilon\|^2_{\mathcal{H}} + a_\varepsilon(Qv_\varepsilon, Qv_\varepsilon)_{\mathcal{H}} \le 2L\|Qv_\varepsilon\|^2_{\mathcal{H}}. \tag{67}$$

LEMMA 8.2. *Under conditions* (14) *and* (16) *we have that*

$$\limsup_{\varepsilon\to 0}\left\{\frac{a_\varepsilon(Qv_\varepsilon, Qv_\varepsilon)_{\mathcal{H}}}{\|Qv_\varepsilon\|^2_{\mathcal{H}}} \ : v \in \mathcal{H}^1_\varepsilon\right\} = +\infty. \tag{68}$$

Proof. Basically we use the same calculations of the spectrum of A_ε as in [11]. □

Lemma 8.2 implies that there exists $\varepsilon_0 > 0$ such that

$$\frac{d}{dt}\|Qv_\varepsilon\|^2_{\mathcal{H}} + \gamma_0\|Qv_\varepsilon\|^2_{\mathcal{H}} \le 0$$

for all $0 < \varepsilon \le \varepsilon_0$ and for some $\gamma_0 > 0$. Therefore,

$$\|Qv_\varepsilon(t)\|^2_{\mathcal{H}} \le \|Qv_\varepsilon(0)\|^2_{\mathcal{H}} e^{-\gamma_0 t}, \quad t \ge 0.$$

This implies that the subspace $\mathcal{L}$ attracts all tempered sets (in both the forward and the pullback sense) with exponential (deterministic) speed. Since $\phi_\varepsilon(t,\omega) \equiv \phi_0(t,\omega)$ on $\mathcal{L}$, this implies (63). □

Theorem 8.1 implies Theorem 3.1(6).

REFERENCES

[1] V. S. AFRAIMOVICH AND H. M. RODRIGUES, *Uniform dissipativeness and synchronization of nonautonomous equations*, in International Conference on Differential Equations (Lisboa 1995), World Scientific, River Edge, NJ, 1998, pp. 3–17.

[2] L. ARNOLD, *Random Dynamical Systems*, Springer-Verlag, Berlin, 1998.

[3] A. BABIN AND M. VISHIK, *Attractors of Evolution Equations*, North–Holland, Amsterdam, 1992.

[4] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.

[5] T. Caraballo and P. E. Kloeden, *The persistence of synchronization under environmental noise*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 2257–2267.
[6] T. Caraballo and J. A. Langa, *On the upper semicontinuity of cocycle attractors for non-autonomous and random dynamical systems*, Dyn. Contin. Discrete Impuls. Sys. Ser. A Math. Anal., 10 (2003), pp. 491–513.
[7] A. N. Carvalho, H. M. Rodrigues, and T. Dlotko, *Upper semicontinuity of attractors and synchronization*, J. Math. Anal. Appl., 220 (1998), pp. 13–41.
[8] V. V. Chepyzhov and M. I. Vishik, *Attractors of Equations of Mathematical Physics*, AMS, Providence, RI, 2002.
[9] I. D. Chueshov, *Introduction to the Theory of Infinite-Dimensional Dissipative Systems*, Acta, Kharkov, Ukraine, 2002; also available online from http://www.emis.de/monographs/Chueshov/.
[10] I. D. Chueshov, *Monotone Random Systems: Theory and Applications*, Lecture Notes in Math. 1779, Springer-Verlag, Berlin, 2002.
[11] I. D. Chueshov, G. Raugel, and A. M. Rekalo, *Interface boundary value problem for the Navier-Stokes equations in thin two-layer domains*, J. Differential Equations, 208 (2005), pp. 449–493.
[12] I. D. Chueshov and A. M. Rekalo, *Long-time dynamics of reaction-diffusion equations on thin two-layer domains*, in EQUADIFF-2003, F. Dumortier, H. Broer, J. Mawhin, A. Vanderbauwhede, and S. V. Lunel, eds., World Scientific, Hackensack, NJ, 2005, pp. 645–650.
[13] I. D. Chueshov and A. M. Rekalo, *Global attractor of contact parabolic problem on thin two-layer domain*, Sb. Math., 195 (2004), pp. 103–128.
[14] I. D. Chueshov and M. Scheutzow, *Inertial manifolds and forms for stochastically perturbed retarded semilinear parabolic equations*, J. Dynam. Differential Equations, 13 (2001), pp. 355–380.
[15] I. D. Chueshov and M. Scheutzow, *On the structure of attractors and invariant measures for a class of monotone random systems*, Dyn. Syst., 19 (2004), pp. 127–144.
[16] I. Ciuperca, *Reaction-diffusion equations on thin domains with varying order of thinness*, J. Differential Equations, 126 (1996), pp. 244–291.
[17] H. Crauel and F. Flandoli, *Attractors for random dynamical systems*, Probab. Theory Related Fields, 100 (1994), pp. 365–393.
[18] H. Crauel, A. Debussche, and F. Flandoli, *Random attractors*, J. Dynam. Differential Equations, 9 (1995), pp. 307–341.
[19] G. Da Prato and G. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.
[20] J. K. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monogr. 25, AMS, Providence, 1988.
[21] J. Hale and G. Raugel, *Reaction-diffusion equation on thin domains*, J. Math. Pures Appl. (9), 71 (1992), pp. 33–95.
[22] J. Hale and G. Raugel, *A reaction-diffusion equation on a thin L-shaped domain*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 283–327.
[23] P. E. Kloeden, *Synchronization of nonautonomous dynamical systems*, Electron. J. Differential Equations, 2003 (2003), 39, pp. 1–10.
[24] H. H. Kuo, *Gaussian Measures in Banach Spaces*, Springer-Verlag, New York, 1972.
[25] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[26] M. Prizzi and K. Rybakowski, *Some recent results on thin domain problems*, Topol. Methods Nonlinear Anal., 14 (1999), pp. 239–255.
[27] A. M. Rekalo, *Asymptotic behavior of solutions of nonlinear parabolic equations on two-layer thin domains*, Nonlinear Anal., 52 (2003), pp. 1393–1410.
[28] H. M. Rodrigues, *Abstract methods for synchronization and applications*, Appl. Anal., 62 (1996), pp. 263–296.
[29] B. Schmalfuss, *Backward cocycle and attractors of stochastic differential equations*, in International Seminar on Applied Mathematics-Nonlinear Dynamics: Attractor Approximation and Global Behaviour, V. Reitmann, T. Redrich, and N. J. Kosch, eds., 1992, pp. 185–192.
[30] R. Showalter, *Monotone Operators in Banach Spaces and Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1997.
[31] H. L. Smith, *Monotone Dynamical Systems, An Introduction to the Theory of Competitive and Cooperative Systems*, AMS, Providence, RI, 1996.
[32] S. Strogatz, *Sync: The Emerging Science of Spontaneous Order*, Hyperion, New York, 2003.
[33] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, 2nd ed., Springer-Verlag, New York, 1997.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1508–1527

EFFECTIVE MACROSCOPIC DYNAMICS OF STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS IN PERFORATED DOMAINS*

WEI WANG†, DAOMIN CAO†, AND JINQIAO DUAN‡

Dedicated to Philip Holmes on the occasion of his 60th birthday

Abstract. An effective macroscopic model for a stochastic microscopic system is derived. The original microscopic system is modeled by a stochastic partial differential equation (SPDE) defined on a domain perforated with small holes or heterogeneities. The homogenized effective model is still an SPDE but defined on a unified domain without holes. The solutions of the microscopic model are shown to converge to those of the effective macroscopic model in probability distribution as the size of holes diminishes to zero. Moreover, the long time effectivity of the macroscopic system, in the sense of *convergence in probability distribution*, and the effectivity of the macroscopic system, in the sense of *convergence in energy*, are also proved.

Key words. stochastic PDEs, effective macroscopic model, stochastic homogenization, white noise, probability distribution, perforated domain

AMS subject classifications. 60H15, 86A05, 34D35

DOI. 10.1137/050648766

1. Introduction.

In recent years there has been an explosive growth of activities in multiscale modeling of complex phenomena in many areas, including material science, climate dynamics, chemistry, and biology [15, 31]. Stochastic partial differential equations (SPDEs or stochastic PDEs)—evolutionary equations containing noises—arise naturally as mathematical models of multiscale systems under random influences. In fact, the need to include stochastic effects in mathematical modeling of realistic physical phenomena has become widely recognized in, for example, condensed matter physics, climate and geophysical sciences, and materials sciences. But implementing this idea poses some challenges both in theory and in computation [17, 33].

This paper is devoted to the effective macroscopic dynamics of microscopic systems modeled by parabolic SPDEs in perforated media which exhibit small-scale heterogeneities. One example of such microscopic systems of interest is composite materials with microscopic heterogeneities under the impact of external random fluctuations. The heterogeneity scale is taken to be much smaller than the macroscopic scale, which is equivalent, here, to assuming that the heterogeneities are evenly distributed. From a mathematical point of view, one can assume that microscopic heterogeneities (holes) are periodically placed in the media. This periodicity can be represented by a small positive parameter ϵ (i.e., the period). In fact we work on the space-time cylinder $D_\epsilon \times (0,T)$, with $T > 0$, and D_ϵ is the spatial domain obtained by removing a number N_ϵ of holes, of size ϵ, periodically distributed, from a fixed

*Received by the editors December 31, 2005; accepted for publication (in revised form) August 30, 2006; published electronically January 12, 2007. This work was partly supported by NSF grants DMS-0209326 and DMS-0542450 and by the Outstanding Overseas Chinese Scholars Fund of the Chinese Academy of Sciences.

http://www.siam.org/journals/sima/38-5/64876.html

†Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, 100080, China (wangwei@amss.ac.cn, dmcao@amt.ac.cn).

‡Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616 (duan@iit.edu).

domain D. When taking $\epsilon \to 0$, the holes inside D are smaller and smaller and their numbers go to ∞. This signifies that the heterogeneities become finer and finer.

There has been much work done on the homogenization problem for the deterministic systems defined in such perforated domains or other heterogeneous media; see, for example, [6, 24, 25, 28, 29, 30] for heat transfer in a composite material, [6, 8, 11] for the wave propagation in a composite material, and [21, 23] for the fluid flow in porous media. For an introduction to homogenization, see [2, 9, 18, 27].

Recently there also has been work done on homogenization of partial differential equations (PDEs) in the random context; see [19, 22, 26] for PDEs with random coefficients, and see [5, 35, 36] for PDEs in randomly perforated domains. Also see the survey book [18] about the homogenization results in a random context. A basic assumption in these texts is the ergodic hypotheses on the coefficients for the passing of the limit of $\epsilon \to 0$. Note that the microscopic models in these works are PDEs with random coefficients, so-called random PDEs, instead of stochastic PDEs.

In the present paper, the microscopic model is an SPDE defined in a perforated domain. Homogenization techniques are employed to derive an effective, simplified, macroscopic model. Homogenization is a formal mathematic procedure for deriving macroscopic models from microscopic systems. It has been applied to a variety of problems including composite materials modeling, porous media, and climate modeling; see [9, 10, 18, 27]. Homogenization provides effective macroscopic behavior of the systems with microscopic heterogeneities for which direct numerical simulations are usually too expensive.

We consider a spatially extended system, where stochastic effects are taken into account in the model equation, defined on a deterministic domain but perforated with small scale holes. Specifically, we study a class of SPDEs driven by white noise on a perforated domain in the following form:

$$du_\epsilon(t) = (\mathcal{A}_\epsilon u_\epsilon + F_\epsilon(x,t))dt + G_\epsilon(x,t)dW(t), \quad 0 < t < T, \quad \epsilon > 0, \tag{1.1}$$

which will be described in detail in the next section. For the general theory of SPDEs we refer to [12]. The goal here is to derive the homogenized equation (effective equation), which is still an SPDE, for (1.1) by the homogenization techniques in the sense of *probability*.

Homogenization theory has been developed for deterministic systems, and a compactness discussion for the solutions $\{u_\epsilon\}_\epsilon$ in some function space is a key step in various homogenization approaches [9]. However, due to the appearance of the stochastic term in the above microscopic system considered in this paper, such compactness result does not hold for this stochastic system. Fortunately the compactness in the sense of probability, that is, the tightness of the distributions for $\{u_\epsilon\}$, still holds. So one appropriate approach is to homogenize the stochastic system in the sense of probability. The goal in this paper is to derive an effective macroscopic equation for the above microscopic system by homogenization in the sense of probability. It is shown that the solution u_ϵ of the microscopic or heterogeneous system converges to that of the macroscopic or homogenized system as $\epsilon \downarrow 0$ in probability distribution. This means that the distribution of $\{u_\epsilon\}_\epsilon$ weakly converges, in some appropriate space, to the distribution of a stochastic process which solves the macroscopic effective equation. Moreover, the long time effectivity of the homogenized macroscopic system is demonstrated; that is, the solution $u_\epsilon(t)$ is shown to converge to the stationary solution of the homogenized equation as $t \to \infty$ and $\epsilon \downarrow 0$ in the sense of probability distribution.

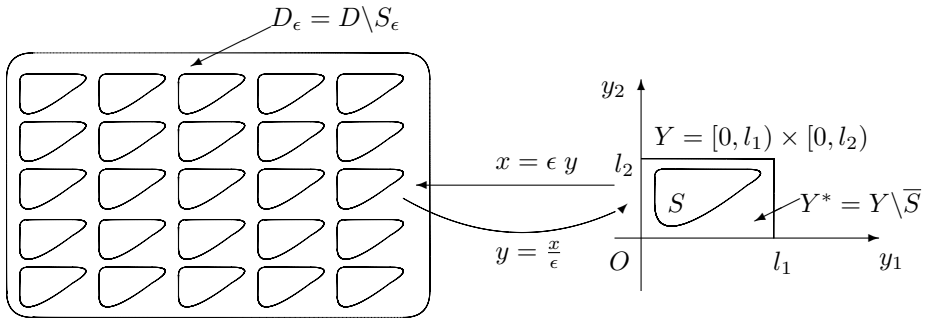


FIG. 1. *Geometric setup in* R^2.

Furthermore, the effectivity of the macroscopic system in the sense of convergence in energy is also shown.

In our approach, one difficulty is that the spatial domain is changing as $\epsilon \to 0$. To overcome this we use the extension operator introduced in [8] and introduce a new probability space depending on a parameter in which the solution is uniformly bounded. One novelty here is that the original microscopic model is a stochastic PDE, instead of a random PDE, as studied by others; see e.g., [19, 26, 7].

This paper is organized as follows. The problem formulation is stated in section 2. Section 3 is devoted to basic properties of the microscopic system. The effective macroscopic equation is derived in section 4. The long time effectivity of the homogenized macroscopic system is considered in section 6. Finally, the effectivity of the macroscopic system in the sense of convergence in energy is shown in section 5. Moreover, in the appendix we present the explicit expression of the homogenization matrix.

2. Problem formulation. Let D be an open bounded set in R^n, $n \geq 2$, with smooth boundary ∂D and $\epsilon > 0$ a small parameter. Let $Y = [0, l_1) \times [0, l_2) \times \cdots \times [0, l_n)$ be a representative (cubic) cell in R^n and S an open subset of Y with smooth boundary ∂S, such that $\overline{S} \subset Y$. Write $l = (l_1, l_2, \ldots, l_n)$. Define $\epsilon S = \{\epsilon y : y \in S\}$. Denote by $S_{\epsilon,k}$ the translated image of ϵS by kl, $k \in Z^n$, $kl = (k_1 l_1, k_2 l_2, \ldots, k_n l_n)$. Also let S_ϵ be the set of all the holes contained in D and $D_\epsilon = D\backslash S_\epsilon$. Then D_ϵ is a periodically perforated domain with holes of the same size as period ϵ. We assume that the holes do not intersect with the boundary ∂D, which implies that $\partial D_\epsilon = \partial D \cup \partial S_\epsilon$. See Figure 1 for the case $n = 2$. This assumption allows us to avoid technicalities, and the results of our paper will remain valid without this assumption; see [1].

In what follows we use the notation

$$Y^* = Y\backslash\overline{S}, \quad \vartheta = \frac{|Y^*|}{|Y|}$$

with $|Y|$ and $|Y^*|$ the Lebesgue measure of Y and Y^*, respectively. Also denote by $\tilde{v}$ the zero extension to the whole D for any function defined on D_ϵ:

$$\tilde{v} = \begin{cases} v & \text{on } D_\epsilon, \\ 0 & \text{on } S_\epsilon. \end{cases}$$

Now for $T > 0$ fixed final time, we consider the following Itô-type nonautonomous SPDE defined on the perforated domain D_ϵ in R^n:

$$du_\epsilon(x,t) = \Big(\operatorname{div}\big(A_\epsilon(x)\nabla u_\epsilon(x,t)\big) + f_\epsilon(x,t)\Big)dt + g_\epsilon(t)dW(t) \quad \text{in } D_\epsilon \times (0,T), \tag{2.1}$$

$$u_\epsilon = 0 \text{ on } \partial D \times (0,T), \tag{2.2}$$

$$\frac{\partial u_\epsilon}{\partial \nu_{A_\epsilon}} = 0 \text{ on } \partial S_\epsilon \times (0,T), \tag{2.3}$$

$$u_\epsilon(0) = u_\epsilon^0 \text{ in } D_\epsilon, \tag{2.4}$$

where the matrix A_ϵ is

$$A_\epsilon = \Big(a_{ij}\Big(\frac{x}{\epsilon}\Big)\Big)_{ij},$$

and

$$\frac{\partial \cdot}{\partial \nu_{A_\epsilon}} = \sum_{ij} a_{ij}\Big(\frac{x}{\epsilon}\Big)\frac{\partial \cdot}{\partial x_j} n_i$$

with n the exterior unit normal vector on the boundary ∂D_ϵ.

We make the following assumptions on the coefficients:

1. $a_{ij} \in L^\infty(R^n)$, $i,j = 1,\dots,n$;
2. $\sum_{i,j=1}^n a_{ij}\xi_i\xi_j \geq \alpha \sum_{i=1}^n \xi_i^2$ for $\xi \in R^n$ and α a positive constant;
3. a_{ij} are Y-periodic.

Furthermore we assume that

$$f_\epsilon \in L^2(D_\epsilon \times [0,T]), \tag{2.5}$$

and for $0 \leq t \leq T$, $g_\epsilon(t)$ is a linear operator from ℓ^2 to $L^2(D_\epsilon)$ defined as

$$g_\epsilon(t)k = \sum_{i=1}^\infty g_\epsilon^i(x,t)k_i, \quad k = (k_1,k_2,\dots) \in \ell^2,$$

where $g_\epsilon^i(x,t) \in L^2(D_\epsilon \times [0,T])$, $i = 1,2,\dots$, are measurable functions with

$$\sum_{i=1}^\infty |g_\epsilon^i(x,t)|^2_{L^2(D_\epsilon)} < C_T, \quad t \in [0,T], \tag{2.6}$$

for some positive constant C_T independent of ϵ. In (2.1), $W(t) = (W_1(t), W_2(t), \dots)$ is a Wiener process in ℓ^2 with covariance operator $Q = Id_{\ell^2}$, and $\{W_i(t) : i = 1,2,\dots\}$ are mutually independent real valued standard Wiener processes on a complete probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with a canonical filtration $(\mathcal{F}_t)_{t\geq 0}$. Then

$$|g_\epsilon(t)|^2_{\mathcal{L}_2^Q} = \sum_{i=1}^\infty |g_\epsilon^i(x,t)|^2_{L^2(D_\epsilon)} < C_T, \quad t \in [0,T]. \tag{2.7}$$

Here $\mathcal{L}_2^Q$ is the space of Hilbert–Schmit operators [12, 16]. Denote by $\mathbf{E}$ the expectation operator with respect to $\mathbf{P}$.

The following compactness result [20] will be used in our approach. Let $\mathcal{X} \subset \mathcal{Y} \subset \mathcal{Z}$ be three reflective Banach spaces and $\mathcal{X} \subset \mathcal{Y}$ with compact and dense embedding. Define the Banach space as

$$G = \left\{ v : v \in L^2(0,T;\mathcal{X}), \frac{dv}{dt} \in L^2(0,T;\mathcal{Z}) \right\}$$

with norm

$$|v|_G^2 = \int_0^T |v(s)|_{\mathcal{X}}^2 ds + \int_0^T \left| \frac{dv}{ds}(s) \right|_{\mathcal{Z}}^2 ds, \quad v \in G.$$

LEMMA 2.1. *If B is bounded in G, then it is precompact in $L^2(0,T;\mathcal{Y})$.*

Let $\mathcal{S}$ be a Banach space and $\mathcal{S}'$ be the strong dual space of $\mathcal{S}$. We recall the definitions and some properties of weak convergence and weak* convergence [34].

DEFINITION 2.2. *A sequence $\{s_n\}$ in $\mathcal{S}$ is said to converge weakly to $s \in \mathcal{S}$ if $\forall s' \in \mathcal{S}'$,*

$$\lim_{n\to\infty} (s', s_n)_{\mathcal{S}',\mathcal{S}} = (s', s)_{\mathcal{S}',\mathcal{S}},$$

which is written as $s_n \rightharpoonup s$ weakly in $\mathcal{S}$. Note that (s', s) denotes the value of the continuous linear functional s' at the point s.

LEMMA 2.3 (Eberlein–Shmulyan). *Assume that $\mathcal{S}$ is reflexive and let $\{s_n\}$ be a bounded sequence in $\mathcal{S}$. Then there exists a subsequence $\{s_{nk}\}$ and $s \in \mathcal{S}$ such that $s_{nk} \rightharpoonup s$ weakly in $\mathcal{S}$ as $k \to \infty$. If all the weakly convergent subsequence of $\{s_n\}$ has the same limit s, then the whole sequence $\{s_n\}$ weakly converges to s.*

DEFINITION 2.4. *A sequence $\{s'_n\}$ in $\mathcal{S}'$ is said to converge weakly* to $s' \in \mathcal{S}'$ if $\forall s \in \mathcal{S}$,*

$$\lim_{n\to\infty} (s'_n, s)_{\mathcal{S}',\mathcal{S}} = (s', s)_{\mathcal{S}',\mathcal{S}},$$

which is written as $s'_n \rightharpoonup s'$ weakly* in $\mathcal{S}'$.

LEMMA 2.5. *Assume that the dual space $\mathcal{S}'$ is reflexive and let $\{s'_n\}$ be a bounded sequence in $\mathcal{S}'$. Then there exists a subsequence $\{s'_{nk}\}$ and $s' \in \mathcal{S}'$ such that $s'_{nk} \rightharpoonup s'$ weakly* in $\mathcal{S}'$ as $k \to \infty$. If all the weakly* convergent subsequence of $\{s'_n\}$ has the same limit s', then the whole sequence $\{s'_n\}$ weakly* converges to s'.*

We also use the following definition of the weak convergence of the Borel probability measures on $\mathcal{S}$; for more details we refer to [14].

DEFINITION 2.6. *Let $\{\mu_\epsilon\}_\epsilon$ be a family of Borel probability measures on the Banach space $\mathcal{S}$. We say μ_ϵ weakly converges to a Borel measure μ on $\mathcal{S}$ if*

$$\int_{\mathcal{S}} h d\mu_\epsilon \to \int_{\mathcal{S}} h d\mu \ \text{ as } \ \epsilon \downarrow 0$$

for any $h \in C_b(\mathcal{S})$, the space of bounded continuous functions on $\mathcal{S}$.

In the following, for a fixed $T > 0$, we always denote by C_T a constant independent of ϵ.

3. Basic properties of the microscopic model. In this section we will present some estimates of the solutions of microscopic model (2.1), useful for the tightness result of the distributions of solution processes in some appropriate space.

Let $H = L^2(D)$ and $H_\epsilon = L^2(D_\epsilon)$. Define the space

$$V_\epsilon = \{u \in H^1(D_\epsilon), u|_{\partial D} = 0\}$$

provided with the norm

$$|v|_{V_\epsilon} = |\nabla_{A_\epsilon} v|_{\oplus_n H_\epsilon} = \left| \left(\sum_{j=1}^n a_{ij}\Big(\frac{x}{\epsilon}\Big) \frac{\partial v}{\partial x_j} \right)_{i=1}^n \right|_{\oplus_n H_\epsilon}.$$

This norm is equivalent to the usual $H^1(D_\epsilon)$-norm, with an embedding constant independent of ϵ, due to the assumptions on a_{ij} in the last section. Here $\oplus_n$ denotes the direct sum of the Hilbert spaces with a usual direct sum norm. Let

$$\mathcal{D}(\mathcal{A}_\epsilon) = \left\{ v \in V_\epsilon : \operatorname{div}(A_\epsilon \nabla v) \in H_\epsilon \text{ and } \left.\frac{\partial v}{\partial \nu_{A_\epsilon}}\right|_{\partial S_\epsilon} = 0 \right\}$$

and define operator $\mathcal{A}_\epsilon v = \operatorname{div}(A_\epsilon \nabla v)$ for $v \in \mathcal{D}(\mathcal{A}_\epsilon)$. Then system (2.1)–(2.4) can be written as the following abstract stochastic evolutionary equation:

$$du_\epsilon = (\mathcal{A}_\epsilon u_\epsilon + f_\epsilon)dt + g_\epsilon dW, \quad u_\epsilon(0) = u_\epsilon^0. \tag{3.1}$$

By the assumptions on a_{ij}, operator $\mathcal{A}_\epsilon$ generates a strongly continuous semigroup $S_\epsilon(t)$ on H_ϵ. The solution of (3.1) can then be written in the mild sense as

$$u_\epsilon(t) = S_\epsilon(t)u_\epsilon^0 + \int_0^t S_\epsilon(t-s)f_\epsilon(s)ds + \int_0^t S_\epsilon(t-s)g_\epsilon(s)dW(s). \tag{3.2}$$

The variational formulation is

$$\begin{aligned} \big(du_\epsilon(t), v\big)_{H_\epsilon^{-1}, V_\epsilon} &= \left(-\int_{D_\epsilon} A_\epsilon(x)\nabla u_\epsilon(x,t)\nabla v(x)dx + \int_{D_\epsilon} f_\epsilon(x,t)v(x)dx \right) dt \\ &\quad + \int_{D_\epsilon} g_\epsilon(x,t)v(x)dW(t), \quad \text{in } \mathcal{D}'(0,T), \quad v \in V_\epsilon, \end{aligned} \tag{3.3}$$

with $u_\epsilon(0,x) = u_\epsilon^0(x)$.

For the well-posedness of system (3.1) we have the following result.

THEOREM 3.1 (global well-posedness of the microscopic model). *Assume that* (2.5) *and* (2.7) *hold. Let* u_ϵ^0 *be an* $\big(\mathcal{F}_0, \mathcal{B}(H_\epsilon)\big)$*-measurable random variable. Then system* (3.1) *has a unique mild solution* $u \in L^2\big(\Omega, C(0,T;H_\epsilon) \cap L^2(0,T;V_\epsilon)\big)$*, which is also a weak solution in the following sense:*

$$\begin{aligned} &(u_\epsilon(t), v)_{H_\epsilon} \\ &= (u_\epsilon^0, v)_{H_\epsilon} + \int_0^t (\mathcal{A}_\epsilon u_\epsilon(s), v)_{H_\epsilon} ds + \int_0^t (f_\epsilon, v)_{H_\epsilon} ds + \int_0^t (g_\epsilon dW, v)_{H_\epsilon} \end{aligned} \tag{3.4}$$

for $t \in [0,T)$ *and* $v \in V_\epsilon$*. Moreover, if* u_ϵ^0 *is independent of* $W(t)$ *with* $\mathbf{E}|u_\epsilon^0|_{H_\epsilon}^2, < \infty$*, then*

$$\mathbf{E}|u_\epsilon(t)|_{H_\epsilon}^2 + \mathbf{E}\int_0^t |u_\epsilon(s)|_{V_\epsilon}^2 ds \le \mathbf{E}|u_\epsilon^0|_{H_\epsilon}^2 + C_T \quad \textit{for } t \in [0,T], \tag{3.5}$$

and

$$\mathbf{E}\int_0^t |\dot{u}_\epsilon(s)|^2_{H_\epsilon^{-1}} ds \le C_T(\mathbf{E}|u_\epsilon^0|^2_{H_\epsilon} + 1) \ \textit{for} \ t \in [0,T]. \tag{3.6}$$

If we further assume that

$$|\nabla_{A_\epsilon} g_\epsilon(t)|^2_{\mathcal{L}_2^Q} = \sum_{i=1}^{\infty} |\nabla_{A_\epsilon} g_\epsilon^i(t)|^2_{\oplus_n H_\epsilon} \le C_T, \ \ \textit{for}\, t \in [0,T] \tag{3.7}$$

and $u_\epsilon^0 \in V_\epsilon$ *with* $\mathbf{E}|u_\epsilon^0|^2_{V_\epsilon} < \infty$, *then*

$$\mathbf{E}|u_\epsilon(t)|^2_{V_\epsilon} + \mathbf{E}\int_0^t |\mathcal{A}_\epsilon u_\epsilon(s)|^2_{H_\epsilon} ds \le \mathbf{E}|u_\epsilon^0|^2_{V_\epsilon} + C_T \ \ \textit{for} \ t \in [0,T]. \tag{3.8}$$

Moreover, system (3.1) *is well-posed on* $[0,\infty)$ *when*

$$f_\epsilon \in L^2(0,\infty; H_\epsilon), \ \ g_\epsilon \in L^2(0,\infty; \mathcal{L}_2^Q). \tag{3.9}$$

Proof. By assumption (2.7), we have

$$|g_\epsilon(t)|^2_{\mathcal{L}_2^Q} = \sum_{i=1}^{\infty} |g_\epsilon^i(t,x)|^2_{H_\epsilon} < \infty.$$

Then the classical result of [12] yields the local existence of u_ϵ. Applying the stochastic Fubini theorem, it is easy to verify that the local mild solution is also a weak solution.

Now we give the following a priori estimate which yields the existence of a weak solution on $[0,T]$, provided (2.5) and (2.7) hold.

Applying the Itô formula to $|u_\epsilon|^2$, we obtain

$$d|u_\epsilon(t)|^2_{H_\epsilon} - 2(\mathcal{A}_\epsilon u_\epsilon, u_\epsilon)_{H_\epsilon} dt = 2(f_\epsilon, u_\epsilon)_{H_\epsilon} dt + 2(g_\epsilon dW, u_\epsilon)_{H_\epsilon} + |g_\epsilon|^2_{\mathcal{L}_2^Q} dt. \tag{3.10}$$

By the assumption on a_{ij}, we see that

$$-(\mathcal{A}_\epsilon u_\epsilon, u_\epsilon)_{H_\epsilon} \ge \lambda |u_\epsilon|^2_{H_\epsilon}$$

for some constant $\lambda > 0$ independent of ϵ. Then integrating (3.10) with respect to t yields

$$\begin{aligned} &|u_\epsilon(t)|^2_{H_\epsilon} + \int_0^t |u_\epsilon|^2_{V_\epsilon} ds \\ &\quad \le |u_\epsilon^0|^2_{H_\epsilon} + \lambda^{-1}|f_\epsilon|^2_{L^2(0,T;H_\epsilon)} + \int_0^t (g_\epsilon dW, u_\epsilon)_{H_\epsilon} ds + \int_0^t |g_\epsilon|^2_{\mathcal{L}_2^Q} ds. \end{aligned}$$

Taking expectation on both sides of the above inequality, we derive (3.5).

In a similar way, application of the Itô formula to $|u_\epsilon|^2_{V_\epsilon} = |\nabla_{A_\epsilon} u_\epsilon|^2_{\oplus_n H_\epsilon}$ results in the relation

$$\begin{aligned} &d|u_\epsilon(t)|^2_{V_\epsilon} + 2(\mathcal{A}_\epsilon u_\epsilon, \mathcal{A}_\epsilon u_\epsilon)_{H_\epsilon} dt \\ &\qquad = -2(f_\epsilon, \mathcal{A}_\epsilon u_\epsilon)_{H_\epsilon} dt - 2(g_\epsilon dW, \mathcal{A}_\epsilon u_\epsilon)_{H_\epsilon} + |\nabla_{A_\epsilon} g_\epsilon|^2_{\mathcal{L}_2^Q} dt. \end{aligned} \tag{3.11}$$

Integrating both sides of (3.11), and by the Cauchy–Schwarz inequality, it is easily to obtain

$$|u_\epsilon(t)|^2_{V_\epsilon} + \int_0^t |\mathcal{A}_\epsilon u_\epsilon|^2_{H_\epsilon} ds \\ \leq |u_\epsilon(0)|^2_{V_\epsilon} + |f_\epsilon|^2_{L^2(0,T;H_\epsilon)} - 2\int_0^t (g_\epsilon dW, \mathcal{A}_\epsilon u_\epsilon)_{H_\epsilon} ds + \int_0^t |\nabla_{A_\epsilon} g_\epsilon|^2_{\mathcal{L}_2^Q} ds,$$

Then taking the expectation, we derive (3.8). By (3.3) and the property of the stochastic integral we easily have (3.6).

Thus, by the above estimates, the solution can be extended to $[0, \infty)$ if (3.9) holds. The proof is complete. □

We recall a probability concept. Let z be a random variable taking values in a Banach space $\mathcal{S}$, namely, $z: \Omega \to z$. Denote by $\mathcal{L}(z)$ the distribution (or law) of z. In fact, $\mathcal{L}(z)$ is a Borel probability measure on $\mathcal{S}$ defined as [12]

$$\mathcal{L}(z)(A) = \mathbf{P}\{\omega : z(\omega) \in A\},$$

for every event (i.e., a Borel set) A in the Borel σ-algebra $\mathcal{B}(\mathcal{S})$, which is the smallest σ-algebra containing all open balls in $\mathcal{S}$.

As stated in section 1, for the SPDE (2.1) we aim at deriving an effective equation in the sense of probability. A solution u_ϵ may be regarded as a random variable taking values in $L^2(0,T;H_\epsilon)$. So for a solution u_ϵ of (2.1)–(2.4) defined on $[0,T]$, we focus on the behavior of the distribution of u_ϵ in $L^2(0,T;H_\epsilon)$ as $\epsilon \to 0$. For this purpose, the tightness [14] of distributions is needed. Note that the function space changes with ϵ, which is a difficulty in obtaining the tightness of distributions. Thus, we will treat $\{\mathcal{L}(u_\epsilon)\}_{\epsilon>0}$ as a family of distributions on $L^2(0,T;H)$ by extending u_ϵ to the whole domain D. Recall that the distribution (or law) of u_ϵ is defined as

$$\mathcal{L}(u_\epsilon)(A) = \mathbf{P}\{\omega : u_\epsilon(\cdot,\cdot,\omega) \in A\}$$

for Borel set A in $L^2(0,T;H_\epsilon)$. First, we define an extension operator P_ϵ in the following lemmas.

In the following we denote by $\mathbf{L}(\mathcal{X},\mathcal{Y})$ the space of a bounded linear operator from Banach space $\mathcal{X}$ to Banach space $\mathcal{Y}$.

LEMMA 3.2. *There exists a bounded linear operator*

$$\hat{Q} \in \mathbf{L}(H^k(Y^*), H^k(Y)), \quad k = 0, 1,$$

such that

$$|\nabla \hat{Q} v|_{\oplus_n L^2(Y)} \leq C |\nabla v|_{\oplus_n L^2(Y^*)}, \quad v \in H^1(Y^*),$$

for some constant $C > 0$.

For the proof of Lemma 3.2 see [8].

In the following lemma we define an extension operator P_ϵ in terms of the above bounded linear operator $\hat{Q}$.

LEMMA 3.3. *There exists an extension operator*

$$P_\epsilon \in \mathbf{L}\big(L^2(0,T;H^k(D_\epsilon)), L^2(0,T;H^k(D))\big), \quad k = 0, 1,$$

such that for any $v \in H^k(D_\epsilon)$,

(1) $P_\epsilon v = v$ *on* $D_\epsilon \times (0,T)$,
(2) $|P_\epsilon v|_{L^2(0,T;H)} \le C_T |v|_{L^2(0,T;H_\epsilon)}$,
(3) $|\nabla_{A_\epsilon}(P_\epsilon v)|_{L^2(0,T;\oplus_n L^2(D))} \le C_T |\nabla_{A_\epsilon} v|_{L^2(0,T;\oplus_n L^2(D_\epsilon))}$,
where C_T *is a constant independent of* ϵ.

Proof. If $\varphi \in H^k(D_\epsilon)$, then

$$\varphi_\epsilon(y) = \frac{1}{\epsilon}\varphi(\epsilon\, y)$$

belongs to $H^k(Y_l^*)$ with Y_l^* the translation of Y^* for some $l \in R^n$. Define

$$\hat{Q}_\epsilon \varphi(x) = \epsilon(\hat{Q}\varphi_\epsilon)\left(\frac{x}{\epsilon}\right). \tag{3.12}$$

Now for $\varphi \in L^2(0,T;H^k(D_\epsilon))$, we define

$$(P_\epsilon\varphi)(x,t) = [\hat{Q}_\epsilon\varphi(t,\cdot)]\Big(\frac{x}{\epsilon}\Big) = \epsilon[\hat{Q}\varphi_\epsilon(t,\cdot)]\Big(\frac{x}{\epsilon}\Big).$$

It is known [8] that the operator $P_\epsilon \in \mathbf{L}\big(L^2(0,T;H^k(D_\epsilon)), L^2(0,T;H^k(D))\big)$, $k = 0,1$, and satisfies conditions (1)–(3) listed in the lemma. This completes the proof. ∎

Remark 3.4. In Lemma 2.1 of [8], the operator P_ϵ defined in $\mathbf{L}(L^\infty(0,T;H^k(D_\epsilon)), L^\infty(0,T;H^k(D)))$, $k = 0,1$, coincides with the operator defined in Lemma 3.3 above.

Remark 3.5. The estimates in Theorem 3.1 for u_ϵ also hold for $P_\epsilon u_\epsilon$. In fact estimates (3.5) and (3.8) are easily derived due to the property of the operator of P_ϵ. Since the operator P_ϵ is defined on $L^2(0,T;H^k(D_\epsilon)$, $k = 0,1$, we define

$$P_\epsilon \dot{u}_\epsilon \equiv \mathcal{A}_\epsilon P_\epsilon u_\epsilon + \tilde{f}_\epsilon + \tilde{g}_\epsilon \dot{W} \ \text{ on } D \times (0,T).$$

By the property of P_ϵ and the estimates of u_ϵ, it is easy to see that

$$P_\epsilon \dot{u}_\epsilon = (\dot{P_\epsilon u_\epsilon}) \ \text{ in } D_\epsilon \times (0,T)$$

and

$$\mathbf{E}|P_\epsilon \dot{u}_\epsilon|_{L^2(0,T;H^{-1})} \le \mathbf{E}|\dot{u}_\epsilon|_{L^2(0,T;H_\epsilon^{-1})}.$$

4. Effective macroscopic model. We now derive the effective macroscopic model for the original model (2.1). Let $u_\epsilon \in L^2(0,T;H_\epsilon)$ be the solution of system (2.1)–(2.4). Then by the estimates in Theorem 3.1, Remark 3.5, and the Chebyshev inequality [12, 14], it is clear that for any $\delta > 0$ there is a bounded set $K_\delta \subset G$ with the spaces $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ in Lemma 2.1 (and in the paragraph immediately before it) replaced by $H_0^1(D)$, H, and $H^{-1}(D)$, respectively, such that

$$\mathbf{P}\{P_\epsilon u_\epsilon \in K_\delta\} > 1 - \delta.$$

Thus, K_δ is compact in $L^2(0,T;H)$ by Lemma 2.1. Then $\{\mathcal{L}(P_\epsilon u_\epsilon)\}_\epsilon$ is tight in $L^2(0,T;H)$. The Prokhorov theorem and the Skorohod embedding theorem [12] assure that for any sequence $\{\epsilon_j\}$ with $\epsilon_j \to 0$ as $j \to \infty$, there exists a subsequence $\{\epsilon_{j(k)}\}$, random variables $\{\hat{u}_{\epsilon_{j(k)}}\} \subset L^2(0,T;H_{\epsilon_{j(k)}})$ and $u \in L^2(0,T;H)$ defined on a new probability space $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mathbf{P}})$, such that

$$\mathcal{L}(P_{\epsilon_{j(k)}}\hat{u}_{\epsilon_{j(k)}}) = \mathcal{L}(P_{\epsilon_{j(k)}} u_{\epsilon_{j(k)}})$$

and

$$P_{\epsilon_{j(k)}}\hat{u}_{\epsilon_{j(k)}} \to u \ \text{ in } \ L^2(0,T;H) \ \text{ as } \ k \to \infty,$$

for almost all $\omega \in \widehat{\Omega}$. Moreover, $P_{\epsilon_{j(k)}}\hat{u}_{\epsilon_{j(k)}}$ solves system (2.1)–(2.4) with W replaced by Wiener process $\widehat{W}_k$ defined on the probability space $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mathbf{P}})$ with the same distribution as W. The limit u is unique; see [4, p. 333]. In the following, we will determine the limiting equation (homogenized effective equation) that u satisfies and see that the limiting equation is independent of ϵ. After this is done we see that $\mathcal{L}(u_\epsilon)$ weakly converges to $\mathcal{L}(u)$ as $\epsilon \downarrow 0$.

We always assume the following conditions

$$\tilde{f}_\epsilon \rightharpoonup f \ \text{ weakly in } \ L^2(0,T;H) \text{ as } \epsilon \to 0, \tag{4.1}$$

and

$$\tilde{g}^i_\epsilon \rightharpoonup g^i \ \text{ weakly in } \ L^2(0,T;H) \text{ as } \epsilon \to 0. \tag{4.2}$$

Define a new probability space $(\Omega_\delta, \mathcal{F}_\delta, \mathbf{P}_\delta)$ as

$$\Omega_\delta = \{\omega \in \Omega : u_\epsilon(\omega) \in K_\delta\},$$

$$\mathcal{F}_\delta = \{F \cap \Omega_\delta : F \in \mathcal{F}\},$$

and

$$\mathbf{P}_\delta(F) = \frac{\mathbf{P}(F \cap \Omega_\delta)}{\mathbf{P}(\Omega_\delta)} \ \text{ for } \ F \in \mathcal{F}_\delta.$$

Denote by $\mathbf{E}_\delta$ the expectation operator with respect to $\mathbf{P}_\delta$.

Now we restrict the system on the probability space $(\Omega_\delta, \mathcal{F}_\delta, \mathbf{P}_\delta)$. In the following discussion we aim at obtaining $L^2(\Omega_\delta)$ convergence for any $\delta > 0$, which means the convergence in probability [3, 14].

From the estimates (3.5), (3.6), Remark 3.5, and the compact embedding of $G \hookrightarrow L^2(0,T;H)$, there exists a subsequence of u_ϵ in K_δ, still denoted by u_ϵ, such that for a fixed $\omega \in \Omega_\delta$,

$$P_\epsilon u_\epsilon \rightharpoonup u \ \text{ weakly}^* \text{ in } \ L^\infty(0,T;H), \tag{4.3}$$

$$P_\epsilon u_\epsilon \rightharpoonup u \ \text{ weakly in } \ L^2(0,T;H^1), \tag{4.4}$$

$$P_\epsilon u_\epsilon \to u \ \text{ strongly in } \ L^2(0,T;H), \tag{4.5}$$

$$P_\epsilon \dot{u}_\epsilon \rightharpoonup \dot{u} \ \text{ weakly in } \ L^2(0,T;H^{-1}). \tag{4.6}$$

Define

$$\xi_\epsilon = \left(\sum_{j=1}^n a_{ij}\Big(\frac{x}{\epsilon}\Big)\frac{\partial u_\epsilon}{\partial x_j}\right) = A_\epsilon \nabla u_\epsilon$$

which satisfies

$$-\text{div } \xi_\epsilon = f_\epsilon + g_\epsilon \dot{W} - \dot{u}_\epsilon \ \ in \ \ D_\epsilon \times (0,T), \tag{4.7}$$

$$\xi_\epsilon \cdot n = 0 \ \text{ on } \ \partial S_\epsilon \times (0,T). \tag{4.8}$$

By the hypothesis of a_{ij} and the fact that $(\tilde{u}_\epsilon)_\epsilon$ is bounded in $L^2(0,T;H_0^1)$, we have

$$\tilde{\xi}_\epsilon \rightharpoonup \xi \ \text{ weakly in } \ L^2(0,T;\oplus_n H). \tag{4.9}$$

We make use of Tartar's method of oscillating test functions to determine the limiting equation [9].

Note that

$$\begin{aligned}\int_0^T\int_D \tilde{\xi}_\epsilon\cdot\nabla v\varphi dxdt &= \int_0^T\int_D \tilde{f}_\epsilon v\varphi dxdt + \sum_{i=1}^\infty \int_0^T\int_D \tilde{g}_\epsilon^i v dx\varphi dW_i(t)\\ &\quad + \int_0^T\int_D P_\epsilon u_\epsilon \chi D_\epsilon \dot{\varphi} v dxdt\end{aligned} \tag{4.10}$$

$\forall\ v\in H_0^1(D)$ and $\varphi\in\mathcal{D}(0,T)$. We pass to the limit in (4.10) as $\epsilon\to 0$. Due to the facts

$$P_\epsilon u_\epsilon \to u \ \text{ strongly in } \ L^2(0,T;H), \tag{4.11}$$

$$\chi_{D_\epsilon} \rightharpoonup \vartheta \ \text{ weakly}^* \ \text{ in } \ L^\infty(D), \tag{4.12}$$

and the estimate

$$\mathbf{E}\Big|\sum_{i=1}^\infty \int_0^T\int_D \tilde{g}_\epsilon^i v dx\varphi dW_i(t)\Big|^2 \le \sum_{i=1}^\infty |\tilde{g}_\epsilon^i|^2_{L^2(0,T;H)} |v\varphi|^2_{L^2(0,T;H)},$$

by assumption (4.2) we see that

$$\sum_{i=1}^\infty \int_0^T\int_D \tilde{g}_\epsilon^i v dx\varphi dW_i(t) \to \sum_{i=1}^\infty \int_0^T\int_D g^i v dx\varphi dW_i(t), \ \text{ in } \ L^2(\Omega).$$

Thus, letting $\epsilon\to 0$ in (4.10), and since $L^2(\Omega_\delta)$ is a subspace of $L^2(\Omega)$, one finds that in $L^2(\Omega_\delta)$

$$\begin{aligned}\int_0^T\int_D \xi\cdot\nabla v\varphi dxdt &= \int_0^T\int_D f v\varphi dxdt + \sum_{i=1}^\infty \int_0^T\int_D g^i v dx\varphi dW_i(t)\\ &\quad + \int_0^T\int_D \vartheta u\dot{\varphi} v dxdt.\end{aligned} \tag{4.13}$$

Hence

$$-\text{div}\ \xi(x,t) = f(x,t) + g(x,t)\dot{W} - \vartheta\dot{u} \ \text{ in } \ D\times(0,T). \tag{4.14}$$

In the following we identify the limit ξ. We follow the approach of the deterministic case for the elliptic problem with homogeneous Neumann boundary condition [9].

For any $\lambda\in R^n$, let w_λ be the solution of

$$-\sum_{j=1}^n \frac{\partial}{\partial y_j}\left(\sum_{i=1}^n a_{ij}(y)\frac{\partial w_\lambda}{\partial y_i}\right) = 0 \ \text{ in } \ Y^*, \tag{4.15}$$

$$w_\lambda - \lambda\cdot y \quad \text{is } \ Y\text{-periodic}, \tag{4.16}$$

$$\frac{\partial w_\lambda}{\partial \nu_A} = 0 \ \text{ on } \ \partial S, \tag{4.17}$$

and define

$$w_\lambda^\epsilon = \epsilon(\hat{Q}w_\lambda)\Big(\frac{x}{\epsilon}\Big),$$

where $\hat{Q}$ is as in Lemma 3.2. Then we have [9]

$$w_\lambda^\epsilon \rightharpoonup \lambda \cdot x \ \text{ weakly in } \ H^1(D), \tag{4.18}$$

$$\nabla w_\lambda^\epsilon \rightharpoonup \lambda \ \text{ weakly in } \ \oplus_n L^2(D). \tag{4.19}$$

Now we define

$$(\eta_j^\lambda(y))_j = \left(\sum_{i=1}^n a_{ji}(y)\frac{\partial w_\lambda(y)}{\partial y_i}\right)_j, \ \ y \in Y^*,$$

and $(\eta_\epsilon^\lambda)(x) = (\eta_j^\lambda(x/\epsilon))_j = A_\epsilon^t \nabla w_\lambda^\epsilon$. Then

$$-\text{div}\, \tilde{\eta}_\epsilon^\lambda = 0 \ \text{ in } \ D, \tag{4.20}$$

and due to (4.18) and (4.19)

$$\tilde{\eta}_\epsilon^\lambda \rightharpoonup \mathcal{M}_Y(\eta^\lambda) \ \text{ weakly in } \ L^2(D). \tag{4.21}$$

It is easy to see that $\mathcal{M}_Y(\eta^\lambda) = B^t\lambda$ with $B^t = (\beta_{ji})$ a constant matrix, which is determined in the appendix.

Using test function $\varphi v w_\lambda^\epsilon$ with $\varphi \in \mathcal{D}(0,T)$, $v \in \mathcal{D}(D)$ in (4.10) and multiplying both sides of (4.20) with $\varphi v P_\epsilon u_\epsilon$, we thus obtain

$$\begin{aligned}
&\int_0^T\int_D \tilde{\xi}_\epsilon \cdot \nabla v\varphi w_\lambda^\epsilon dxdt + \int_0^T\int_{D_\epsilon} \xi_\epsilon \cdot \nabla w_\lambda^\epsilon v\varphi dxdt \\
&- \int_0^T\int_D \tilde{\eta}_\epsilon^\lambda \cdot \nabla v\varphi P_\epsilon u_\epsilon dxdt - \int_0^T\int_D \tilde{\eta}_\epsilon^\lambda \cdot \nabla(P_\epsilon u_\epsilon) v\varphi dxdt \\
&= \int_0^T\int_D \tilde{f}_\epsilon \varphi v w_\lambda^\epsilon dxdt + \sum_{i=1}^\infty \int_0^T\int_D \tilde{g}_\epsilon^i v w_\lambda^\epsilon dx\varphi dW_i(t) + \int_0^T\int_D P_\epsilon u_\epsilon \chi_{D_\epsilon}\dot{\varphi} v w_\lambda^\epsilon dxdt.
\end{aligned}$$

Then by the definition of ξ_ϵ, η_ϵ^λ and assumptions (4.1), (4.2), using the convergences (4.9), (4.11), (4.12), (4.18), (4.19) and (4.21), we have in $L^2(\Omega_\delta)$

$$\begin{aligned}
&\int_0^T\int_D \xi \cdot \nabla v\varphi\lambda \cdot x dxdt - \int_0^T\int_D B^t\lambda \cdot \nabla v\varphi u dxdt \\
&\quad = \int_0^T\int_D f\varphi v\lambda \cdot x dxdt + \sum_{i=1}^\infty \int_0^T\int_D g^i v\lambda \cdot x dx\varphi dW_i(t) + \int_0^T\int_D \vartheta u v\dot{\varphi}\lambda \cdot x dxdt.
\end{aligned}$$

That is,

$$\begin{aligned}
&\int_0^T\int_D \xi \cdot \nabla(v\lambda \cdot x)\varphi dxdt - \int_0^T\int_D \xi \cdot \lambda v\varphi dxdt - \int_0^T\int_D B^t\lambda \cdot \nabla v\varphi u dxdt \\
&\quad = \int_0^T\int_D f\varphi v\lambda \cdot x dxdt + \sum_{i=1}^\infty \int_0^T\int_D g^i v\lambda \cdot x dx\varphi dW_i(t) + \int_0^T\int_D \vartheta u v\dot{\varphi}\lambda \cdot x dxdt.
\end{aligned}$$

Then by using (4.13) with the test function replaced by $v\lambda \cdot x\varphi$, one has

$$\int_0^T \int_D \xi \cdot \lambda v\varphi dxdt = \int_0^T \int_D B^t\lambda \cdot \nabla u\varphi v dxdt,$$

which yields

$$\xi \cdot \lambda = B^t\lambda \cdot \nabla u = B\nabla u \cdot \lambda.$$

Then

$$\xi = B\nabla u$$

since λ is arbitrary. Then u satisfies the equation

$$\vartheta du = \big(\mathrm{div}(B\nabla u) + f\big)dt + gdW(t). \tag{4.22}$$

Assume that

$$\tilde{u}_\epsilon^0 \rightharpoonup u^0, \text{ weakly in } H \text{ as } \epsilon \to 0. \tag{4.23}$$

We now determine the initial value by suitable test functions. In fact, taking $v \in \mathcal{D}(D)$ and $\varphi \in \mathcal{D}([0,T])$ with $\varphi(T) = 0$, we have

$$\begin{aligned}\int_0^T \int_D \tilde{\xi}_\epsilon \cdot \nabla v\varphi dxdt = &\int_0^T \int_D \tilde{f}_\epsilon v\varphi dxdt + \sum_{i=0}^{\infty} \int_0^T \int_D \tilde{g}\epsilon^i v dx\varphi dW_i(t)\\ &- \int_0^T \int_D \tilde{u}_\epsilon v\dot{\varphi} dxdt + \int_D \tilde{u}_\epsilon^0 \varphi(0) v dx.\end{aligned}$$

Now let $\epsilon \to 0$. Noticing that

$$\begin{aligned}\int_0^T \int_D \tilde{u}_\epsilon v\dot{\varphi} dxdt &= \int_0^T \int_D \chi_{D\epsilon} P_\epsilon \tilde{u}_\epsilon v\dot{\varphi} dxdt \to \int_0^T \int_D \vartheta u v\dot{\varphi} dxdt\\ &= -\int_0^T \int_D \vartheta \dot{u} v\varphi dxdt + \int_D \vartheta u(0)\varphi(0) v dx\end{aligned}$$

by (4.14), we have

$$u(0) = \frac{u^0}{\vartheta}.$$

Here one should notice that the above result is in the sense of $L^2(\Omega_\delta)$. Then the above analysis yields

$$\lim_{\epsilon \to 0} \mathbf{E}_\delta |P_\epsilon u_\epsilon - u|^2_{L^2(0,T;H)} = 0 \tag{4.24}$$

and

$$\lim_{\epsilon \to 0} \mathbf{E}_\delta \int_0^T \int_D (\mathcal{A}_\epsilon P_\epsilon u_\epsilon - B\nabla u) v\varphi dxdt = 0 \tag{4.25}$$

for any $v \in \mathcal{D}(D)$ and $\varphi \in \mathcal{D}([0,T])$.

Now we are in the position to present the homogenized effective equation in the following theorem.

THEOREM 4.1 (effective macroscopic model). *For any $T > 0$, assume that* (4.1), (4.2), *and* (4.23) *hold. Let u_ϵ be the solution of* (2.1)–(2.4). *Then the distribution $\mathcal{L}(P_\epsilon u_\epsilon)$ converges weakly to μ in the space of probability measures on $L^2(0,T;H)$ as $\epsilon \downarrow 0$, with μ being the distribution of u, which is the solution of the following homogenized effective SPDE*

$$\vartheta du = \big(\operatorname{div}(B\nabla u) + f\big)dt + gdW(t) \quad \text{in } D \times (0,T), \tag{4.26}$$

$$u = 0 \text{ on } \partial D \times (0,T), \tag{4.27}$$

$$u(x,0) = \frac{u^0}{\vartheta} \quad \text{in } D, \tag{4.28}$$

where the constant coefficient $\vartheta = \frac{|Y^*|}{|Y|}$ is defined in the beginning of section 2, *and the effective matrix $B = (\beta_{ij})$ is determined by* (7.4) *in the appendix at the end of this paper. Moreover, the coefficients f, g and initial datum u^0 are defined in* (4.1), (4.2), *and* (4.23), *respectively.*

Remark 4.2. This theorem implies that the macroscopic model (4.26) is an effective approximation for the microscopic model (2.1), on any finite time interval $0 < t < T$, in the sense of probability distribution. In other words, if we intend to numerically simulate the microscopic model up to finite time, we could use the macroscopic model as an approximation when ϵ is sufficiently small.

Remark 4.3. Due to the appearance of the stochastic integral term (see (4.10)), this theorem on weak convergence of probability measures does not follow from the deterministic homogenization results and the mild formulation (3.2).

Remark 4.4. The SPDE (4.26) is defined on the homogenized domain D. By the analysis in [12], for any fixed $T > 0$, the macroscopic system (4.26)–(4.28) is well-posed, as long as $f \in L^2(0,T;H)$ and $g \in L^2(0,T;\mathcal{L}_2^Q)$.

Proof of Theorem 4.1. Noticing the arbitrariness of δ, we see that this is a direct result of the analysis of the first part in this section by the Skorohod theorem and the $L^2(\Omega_\delta)$ convergence of $P_\epsilon u_\epsilon$ on $(\Omega_\delta, \mathcal{F}_\delta, \mathbf{P}_\delta)$. □

We finish this section with the following remark.

Remark 4.5. Note that there are several papers on effective dynamics for PDEs with random coefficients (so-called random PDEs, not stochastic PDEs); see [19, 26, 32] and references therein. In [19, 26], a random PDE is obtained as the homogenized effective equation for a random system with fast or small scales on its time or spatial variable. The distribution of the solution of the heterogeneous system converges weakly to that of the homogenized equation. However, in [32], the effective equation is obtained as an averaged deterministic equation for a random system with fast scales in time. The fluctuation of the solution of the random equation around the solution of the averaged equation converges to a generalized Ornstein–Uhlenbeck process in distribution. In the present paper, the original microscopic model is an SPDE (i.e., PDE with white noise) and the effective macroscopic equation is still an SPDE.

5. Long time effectivity of the macroscopic model. In this section we consider the long time effectivity of the homogenized system (4.26) in the autonomous case, i.e., when f_ϵ and g_ϵ (and thus f and g) are independent of time t. It is proved in section 4 that for fixed $T > 0$ the macroscopic behavior of the microscopic system (2.1)–(2.4) can be approximated by the macroscopic model (4.26) in the sense of probability distribution. In fact we can show the long time approximation. More

specifically, we now prove that in the sense of distribution, all solutions of (2.1)–(2.4) converge to the unique stationary solution of (4.26) as $T \to \infty$ and $\epsilon \to 0$, under the assumption that $f_\epsilon \in H_\epsilon$ and $g_\epsilon^i \in V_\epsilon$ are *independent* of time t and

$$\sum_{i=1}^{\infty} |\nabla_{A_\epsilon} g_\epsilon^i(x)|^2_{\oplus_n H_\epsilon} < C^*. \tag{5.1}$$

Here C^* is a positive constant independent of ϵ.

By the above assumptions, as well as the properties of a_{ij} and β_{ij}, a standard argument (see [13, section 6]) yields that the system (3.1) and (4.26) have unique stationary solutions $u_\epsilon^*(x,t)$ and $u^*(x,t)$, defined for $t > 0$. We denote by μ_ϵ^* and μ^* the distributions of $P_\epsilon u_\epsilon^*$ and u^* in the space H, respectively. Then if $\mathbf{E}|u_\epsilon^0|^2 < \infty$ and $\mathbf{E}|u^0|^2 < \infty$,

$$\Big| \int_H h d\mu_\epsilon(t) - \int_H h d\mu_\epsilon^* \Big| \le C(u_\epsilon^0) e^{-\gamma t}, \ \ t > 0, \tag{5.2}$$

$$\Big| \int_H h d\mu(t) - \int_H h d\mu^* \Big| \le C(u^0) e^{-\gamma t}, \ \ t > 0, \tag{5.3}$$

for some constant $\gamma > 0$ and any $h : H \to R^1$ with $\sup|h| \le 1$ and $\mathrm{Lip}(h) \le 1$. Here $\mu_\epsilon(t) = \mathcal{L}(P_\epsilon u_\epsilon(t, u_\epsilon^0))$, $\mu(t) = \mathcal{L}(u(t, \frac{u^0}{\vartheta}))$, and $C(u_\epsilon^0)$ and $C(u^0)$ are positive constants depending only on the initial values u_ϵ^0 and u^0, respectively. The above convergence also yields that $\mu_\epsilon(t)$ and $\mu(t)$ weakly converge to μ_ϵ^* and μ^*, respectively, as $t \to \infty$.

We will give some additional a priori estimates which are uniform with respect to ϵ to ensure the tightness of the stationary distributions. For Banach space U and $p > 1$, we define $W^{1,p}(0,T;U)$ as the space of functions $h \in L^p(0,T;U)$ such that

$$|h|^p_{W^{1,p}(0,T;U)} = |h|^p_{L^p(0,T;U)} + \Big|\frac{dh}{dt}\Big|^p_{L^p(0,T;U)} < \infty.$$

For any $\alpha \in (0,1)$, define $W^{\alpha,p}(0,T;U)$ as the space of function $h \in L^p(0,T;U)$ such that

$$|h|^p_{W^{\alpha,p}(0,T;U)} = |h|^p_{L^p(0,T;U)} + \int_0^T \int_0^T \frac{|h(t) - h(s)|^p_U}{|t-s|^{1+\alpha p}} ds dt < \infty.$$

For $\rho \in (0,1)$, we denote by $C^\rho(0,T;U)$ the space of functions $h : [0,T] \to \mathcal{X}$ that are Hölder continuous with exponent ρ.

In the remaining part of this section, we always assume that f_ϵ and g_ϵ^i are independent of time t with (5.1) holding. For $T > 0$ denote by $\mathbf{u}_{\epsilon,T}^*$ (respectively, $\mathbf{u}_T^*$) the distribution of stationary process $P_\epsilon u_\epsilon^*(\cdot)$ (respectively, $u^*(\cdot)$) in the space $L^2(0,T;H^1)$. Then we have the following result.

LEMMA 5.1. *For any $T > 0$ the family $\mathbf{u}_{\epsilon,T}^*$ is tight in the space $L^2(0,T;H^{2-\iota})$ with $\iota > 0$.*

Proof. Since u_ϵ^* is stationary, by (3.8), we see that

$$\mathbf{E}|u_\epsilon^*|^2_{L^2(0,T;H_\epsilon^2)} < C_T. \tag{5.4}$$

Now represent u_ϵ^* in the form

$$u_\epsilon^*(t) = u_\epsilon^*(0) + \int_0^t \mathcal{A}_\epsilon u_\epsilon^*(s) ds + \int_0^t f_\epsilon(x) ds + \int_0^t g_\epsilon(x) dW(s).$$

Also by the stationarity of u_ϵ^* and (3.8) we obtain

$$\mathbf{E}\Big|\int_0^t \mathcal{A}_\epsilon P_\epsilon u_\epsilon^*(s)ds + \int_0^t \tilde{f}_\epsilon(x)ds\Big|^2_{W^{1,2}(0,T;H)} \le C_T. \tag{5.5}$$

Let $M_\epsilon(s,t) = \int_s^t \tilde{g}_\epsilon(x)dW(s)$. By Lemma 7.2 of [12] and the Hölder inequality, we derive that

$$\begin{aligned}\mathbf{E}|M_\epsilon(s,t)|^4_{V_\epsilon} \le c\left(\int_s^t |\nabla_{A_\epsilon}\tilde{g}_\epsilon(x)|^2_{\mathcal{L}_2^Q}d\tau\right)^2 &\le K(t-s)\int_s^t |\nabla_{A_\epsilon}\tilde{g}_\epsilon(x)|^4_{\mathcal{L}_2^Q}d\tau \\ &\le KC^{*2}|t-s|^2\end{aligned}$$

for $t \in [s,T]$, where K is a positive constant independent of ϵ, s, and t. Then

$$\mathbf{E}\int_0^T |M_\epsilon(0,t)|^4_{V_\epsilon}dt \le C_T \tag{5.6}$$

and

$$\mathbf{E}\int_0^T\int_0^T \frac{|M_\epsilon(0,t)-M_\epsilon(0,s)|^4_{V_\epsilon}}{|t-s|^{1+4\alpha}}dsdt \le C_T. \tag{5.7}$$

Combining (5.4)–(5.7) and the compact embedding of

$$L^2(0,T;H^2)\cap W^{1,2}(0,T;H) \subset L^2(0,T;H^{2-\iota})$$

and

$$L^2(0,T;H^2)\cap W^{\alpha,4}(0,T;H^1) \subset L^2(0,T;H^{2-\iota}),$$

we obtain the tightness of $\mathbf{u}^*_{\epsilon,T}$. This completes the proof. □

The above lemma directly yields the following result.

COROLLARY 5.2. *The family $\{\mu_\epsilon^*\}$ is tight in the space H^1.*

By Lemma 5.1, for any fixed $T > 0$, the Skorohod embedding theorem asserts that for any sequence $\{\epsilon_n\}_n$ with $\epsilon_n \to 0$ as $n \to \infty$, there is a subsequence $\{\epsilon_{n(k)}\}_k$, a new probability space $(\overline{\Omega}, \overline{\mathcal{F}}, \overline{\mathbf{P}})$, and random variables $\overline{u}^*_{\epsilon_{n(k)}} \in L^2(0,T;V_\epsilon)$, $\overline{u}^* \in L^2(0,T;H^1)$, such that

$$\mathcal{L}(P_\epsilon \overline{u}^*_{\epsilon_{n(k)}}) = \mathbf{u}^*_{\epsilon_{n(k)},T}, \quad \mathcal{L}(\overline{u}^*) = \mathbf{u}^*_T,$$

and

$$\overline{u}^*_{\epsilon_{n(k)}} \to \overline{u}^* \text{ in } L^2(0,T;H^1) \text{ as } k \to \infty.$$

Moreover, $\overline{u}^*_{\epsilon_{n(k)}}$ (respectively, $\overline{u}^*$) is the unique stationary solution of (3.1) (respectively, (4.26)) with W replaced by $\overline{W}_k$ (respectively, $\overline{W}$). $\overline{W}_k$ and $\overline{W}$ are some Wiener processes defined on $(\overline{\Omega}, \overline{\mathcal{F}}, \overline{\mathbf{P}})$ with the same distribution as W. Then by the analysis of section 4 and the uniqueness of the invariant measure,

$$\mathbf{u}^*_{\epsilon,T} \rightharpoonup \mathbf{u}^*_T \text{ as } \epsilon \to 0$$

for any $T > 0$.

To show the long time effectivity, let $u_\epsilon(t)$, $t \geq 0$, be a weak solution of system (2.1)–(2.4), and define $u_\epsilon^t(\cdot) = u_\epsilon(t+\cdot)$ which is in the space $L^2_{loc}(R_+; V_\epsilon)$ by Theorem 3.1. Then by (5.2)

$$\mathcal{L}(P_\epsilon u_\epsilon^t(\cdot)) \rightharpoonup \mathcal{L}(P_\epsilon u_\epsilon^*(\cdot)), \quad t \to \infty,$$

in the space of probability measures on $L^2_{loc}(R_+; H^1)$. From the above analysis we draw the following result which implies the long time effectivity of the homogenized effective equation (4.26).

THEOREM 5.3 (long time effectivity of the macroscopic model). *Assume that $f_\epsilon \in H_\epsilon$ and $g_\epsilon^i \in V_\epsilon$ are independent of time t with* (5.1) *being satisfied, and further assume that* (4.1) *and* (4.2) *hold in H. Denote by $u_\epsilon(t)$, $t \geq 0$, the solution of* (2.1)–(2.4) *and by u^* the unique stationary solution of* (4.26). *Then*

$$\lim_{\epsilon \downarrow 0} \lim_{t \to \infty} \mathcal{L}(P_\epsilon u_\epsilon^t(\cdot)) = \mathcal{L}(u^*(\cdot)), \tag{5.8}$$

where the limits are understood in the sense of weak convergence of Borel probability measures in the space $L^2_{loc}(R_+; H^1)$. That is, the solution of (2.1)–(2.4) *converges to the stationary solution of* (4.26) *in probability distribution as $t \to \infty$ and $\epsilon \to 0$.*

Remark 5.4. This theorem implies that the macroscopic model (4.26) is an effective approximation for the microscopic model (2.1), on a very long time scale. In other words, if we intend to numerically simulate the long time behavior of the microscopic model, we could just simulate the macroscopic model as an approximation when ϵ is sufficiently small.

6. Effectivity in energy convergence. In sections 4 and 5, we have considered finite time and long time effectivity of the macroscopic model (4.26), in the sense of convergence in probability distribution. In this section we focus on the finite time effectivity of the macroscopic model (4.26), but in the sense of convergence in energy. Namely, we show that the solution of the microscopic model (2.1) or (3.1) converges to the solution of the macroscopic model (4.26), in an energy norm.

Let u_ϵ be a weak solution of (3.1) and u be a weak solution of (4.26). We introduce the following energy functionals:

$$\mathcal{E}^\epsilon(u_\epsilon)(t) = \frac{1}{2}\mathbf{E}|\tilde{u}_\epsilon|_H^2 + \mathbf{E}\int_0^t \int_D \chi_{D_\epsilon} A_\epsilon \nabla\big(P_\epsilon u_\epsilon(x,\tau)\big)\nabla\big(P_\epsilon u_\epsilon(x,\tau)\big)dxd\tau \tag{6.1}$$

and

$$\mathcal{E}^0(u)(t) = \frac{1}{2}\mathbf{E}|u|_H^2 + \mathbf{E}\int_0^t \int_D B\nabla u(x,\tau)\nabla u(x,\tau)dxd\tau. \tag{6.2}$$

By the Itô formula, it is clear that

$$\mathcal{E}^\epsilon(u_\epsilon)(t) = \frac{1}{2}\mathbf{E}|\tilde{u}_\epsilon^0|_H^2 + \mathbf{E}\int_0^t \int_D \tilde{f}_\epsilon(x,\tau)\tilde{u}_\epsilon(x,\tau)dxd\tau + \frac{1}{2}\mathbf{E}\int_0^t |\tilde{g}_\epsilon(x,\tau)|^2_{\mathcal{L}_2^Q} d\tau$$

and

$$\mathcal{E}^0(u)(t) = \frac{1}{2}\mathbf{E}|u^0|_H^2 + \mathbf{E}\int_0^t \int_D f(x,\tau)u(x,\tau)dxd\tau + \frac{1}{2}\mathbf{E}\int_0^t |g(x,\tau)|^2_{\mathcal{L}_2^Q} d\tau.$$

Then we have the following result on effectivity of the macroscopic model in the sense of convergence in energy.

THEOREM 6.1 (effectivity in energy convergence). *Assume that* (4.1) *and* (4.2) *hold. If*

$$\tilde{u}_\epsilon^0 \to u^0 \ \text{ strongly in } \ H \ \text{ as } \epsilon \to 0,$$

then

$$\mathcal{E}^\epsilon(u_\epsilon) \to \mathcal{E}^0(u) \quad \text{in } \ C([0,T]) \ \text{ as } \epsilon \to 0.$$

Proof. By the analysis of section 4, for any $\delta > 0$, $u_\epsilon \to u$ strongly in $L^2(0,T;H)$ on Ω_δ, and then by the arbitrariness of δ, it is easy to see that

$$\mathbf{E}\int_0^t \int_D \tilde{f}_\epsilon(x,\tau)\tilde{u}_\epsilon(x,\tau)dxd\tau \to \mathbf{E}\int_0^t \int_D f(x,\tau)u(x,\tau)dxd\tau \ \text{ for } t \in [0,T].$$

Then by $\tilde{g}_\epsilon \rightharpoonup g$ weakly in $L^2(0,t;\mathcal{L}_2^Q)$, we have

$$\mathcal{E}^\epsilon(u_\epsilon)(t) \to \mathcal{E}^0(u)(t) \ \text{ for any } \ t \in [0,T]. \tag{6.3}$$

We now need only show that $\{\mathcal{E}^\epsilon(u_\epsilon)(t)\}_\epsilon$ is equicontinuous, as then the Ascoli–Arzela theorem [14] will imply the result in the theorem.

In fact, given any $t \in [0,T]$, and $h > 0$ small enough, we have

$$\begin{aligned}
&|\mathcal{E}^\epsilon(u_\epsilon)(t+h) - \mathcal{E}^\epsilon(u_\epsilon)(t)| \\
&\quad \le \Big|\mathbf{E}\int_t^{t+h}\int_D \tilde{f}_\epsilon(x,\tau)\tilde{u}_\epsilon(x,\tau)dxd\tau\Big| + \mathbf{E}\int_t^{t+h} |\tilde{g}_\epsilon(x,\tau)|^2_{\mathcal{L}_2^Q} d\tau \\
&\quad \le \mathbf{E}\left\{|\tilde{f}_\epsilon|_{L^2(0,T;H)}\int_t^{t+h} |\tilde{u}_\epsilon(x,\tau)|_H^2 dxd\tau\right\} + \mathbf{E}\int_t^{t+h} |\tilde{g}_\epsilon(x,\tau)|^2_{\mathcal{L}_2^Q} d\tau.
\end{aligned}$$

Noting that $\tilde{u}_\epsilon \in L^2(0,T;H)$ a.s. and (2.7), we have

$$|\mathcal{E}^\epsilon(u_\epsilon)(t+h) - \mathcal{E}^\epsilon(u_\epsilon)(t)| \to 0 \ \text{ as } h \to 0,$$

uniformly on ϵ, which means the equicontinuity of the family $\{\mathcal{E}^\epsilon(u_\epsilon)\}_\epsilon$. This completes the proof. ∎

7. Appendix: The homogenized matrix. In this appendix, we give the explicit expression of the homogenized matrix B; for more details see [9]. Let χ^i, $i = 1,\dots,n$, be the solutions of

$$-\sum_{l,k=1}^n \frac{\partial}{\partial y_l}\left(a_{kl}\frac{\partial(\chi^i - y_i)}{\partial y_k}\right) = 0 \ \text{ in } \ Y^*, \tag{7.1}$$

$$\sum_{l,k=1}^n a_{kl}\frac{\partial(\chi^i - y_i)}{\partial y_k} n_l = 0 \ \text{ on } \partial S, \tag{7.2}$$

$$\chi^i \ \text{ is } \ Y\text{-periodic}. \tag{7.3}$$

It is easy to calculate that $\chi^i = -w_{e_i} + e_i$ with $\{e_i\}_{i=1}^n$ the canonical basis of R^n. Then

$$\beta_{ij} = \frac{1}{|Y|}\int_Y \sum_{k=1}^n a_{kj}\frac{\partial w_{e_i}}{\partial y_k}dy = \frac{1}{|Y|}\int_Y a_{ij}dy - \frac{1}{|Y|}\int_Y \sum_{k=1}^n a_{kj}\frac{\partial \chi_i}{\partial y_k}dy. \tag{7.4}$$

Moreover, the operator $B = (\beta_{ij})$ satisfies the uniform ellipticity condition: there is a constant $b > 0$ such that

$$\sum_{i,j=1}^{n} \beta_{ij}\xi_i\xi_j \geq b\sum_{i=1}^{n} \xi_i^2 \quad \text{for } \xi = (\xi_1, \ldots, \xi_n) \in R^n.$$

Acknowledgments. The authors thank the referees for very helpful suggestions and comments.

REFERENCES

[1] G. Allaire, M. Murat, and A. Nandakumar, *Appendix of "Homogenization of the Neumann problem with nonisolated holes,"* Asymptotic Anal., 7 (1993), pp. 81–95.
[2] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structure*, North–Holland, Amsterdam, New York, 1978.
[3] P. Billingsley, *Weak Convergence of Probability Measures*, John Wiley, New York, 1968.
[4] P. Billingsley, *Probability and Measure*, 3rd ed., John Wiley, New York, 1995.
[5] M. Briane and L. Mazliak, *Homogenization of two randomly weakly connected materials*, Portugal. Math., 55 (1998), pp. 187–207.
[6] S. Brahim-Otsmane, G. A. Francfort, and F. Murat, *Correctors for the homogenization of the wave and heat equations*, J. Math. Pures Appl., 71 (1998), pp. 197–231.
[7] L. A. Caffarelli, P. Souganidis, and L. Wang, *Homogenization of fully nonlinear, uniformly elliptic and parabolic partial differential equations in stationary ergodic media*, Comm. Pure Appl. Math., 68 (2005), pp. 319–361.
[8] D. Cioranescu and P. Donato, *Exact internal controllability in perforated domains*, J. Math. Pures Appl., 68 (1989), pp. 185–213.
[9] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford University Press, New York, 1999.
[10] A. Cherkaev and R. V. Kohn, *Topics in the Mathematical Modelling of Composite Materials*, Birkhäuser Boston, Boston, MA, 1997.
[11] D. Cioranescu, P. Donato, F. Murat, and E. Zuazua, *Homogenization and correctors results for the wave equation in domains with small holes*, Ann. Scuola Norm. Sup. Pisa, 18 (1991), pp. 251–293.
[12] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.
[13] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite Dimensional Systems*, Cambridge University Press, Cambridge, UK, 1996.
[14] R. M. Dudley, *Real Analysis and Probability*, Cambridge University Press, Cambridge, UK, 2002.
[15] W. E, X. Li, and E. Vanden-Eijnden, *Some recent progress in multiscale modeling*, in Multiscale Modelling and Simulation, Lecture Notes in Comput. Sci. Eng. 39, Springer, Berlin, 2004, pp. 3–21.
[16] Z. Huang and J. Yan, *Introduction to Infinite Dimensional Stochastic Analysis*, Science Press/Kluwer Academic Publishers, Beijing/New York, 1997.
[17] P. Imkeller and A. Monahan, eds., *Stochastics and Dynamics*, special issue on stochastic climate dynamics, Vol. 2, No. 3, World Scientific, River Edge, NJ, 2002.
[18] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[19] M. L. Kleptsyna and A. L. Piatnitski, *Homogenization of a random non-stationary convection-diffusion problem*, Russian Math. Surveys, 57 (2002), pp. 729–751.
[20] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes Non Linéaires*, Dunod, Paris, 1969.
[21] P. L. Lions and N. Masmoudi, *Homogenization of the Euler system in a 2D porous medium*, J. Math. Pures Appl., 84 (2005), pp. 1–20.
[22] G. D. Maso and L. Modica, *Nonlinear stochastic homogenization and ergodic theory*, J. Reine Angew. Math., 368 (1986), pp. 28–42.
[23] A. Mikelić and L. Paloi, *Homogenization of the inviscid incompressible fluid flow through a 2D porous medium*, in Proc. Amer. Math. Soc., 127 (1999), pp. 2019–2028.
[24] A. K. Nandakumaran and M. Rajesh, *Homogenization of a parabolic equation in a perforated domain with Neumann boundary condition*, Proc. Indian Acad. Sci. (Math. Sci.), 112 (2002), pp. 195–207.

[25] A. K. Nandakumaran and M. Rajesh, *Homogenization of a parabolic equation in a perforated domain with Dirichlet boundary condition*, Proc. Indian Acad. Sci. (Math. Sci.), 112 (2002), pp. 425–439.
[26] E. Pardoux and A. L. Piatnitski, *Homogenization of a nonlinear random parabolic partial differential equation*, Stochastic Process Appl., 104 (2003), pp. 1–27.
[27] E. Sanchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Physics 127, Springer-Verlag, Berlin, 1980.
[28] J. Souza and A. Kist, *Homogenization and corrector results for a nonlinear reaction-diffusion equation in domains with small holes*, in Proceedings of the 7th Workshop on Partial Differential Equations II, Mat. Contemp. 23, Soc. Brasil. Mat. Estrada Dona Castorina, Rio de Janeiro, Brazil, 2002, pp. 161–183.
[29] C. Timofte, *Homogenization results for parabolic problems with dynamical boundary conditions*, Romanian Rep. Phys., 56 (2004), pp. 131–140.
[30] M. B. Taghite, K. Taous, and G. Maurice, *Heat equations in a perforated composite plate: Influence of a coating*, Internat. J. Engrg. Sci., 40 (2002), pp. 1611–1645.
[31] R. Temam and A. Miranville, *Mathematical Modeling in Continuum Mechanics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2005.
[32] H. Watanabe, *Averaging and fluctuations for parabolic equations with rapidly oscillating random coefficients*, Prob. Theory Related Fields, 77 (1988), pp. 359–378.
[33] E. Waymire and J. Duan, eds., *Probability and Partial Differential Equations in Modern Applied Mathematics*, IMA Vol. Math. Appl. 140, Springer-Verlag, New York, 2005.
[34] K. Yosida, *Functional Analysis*, 5th ed., Springer-Verlag, Berlin, 1978.
[35] V. V. Zhikov, *On homogenization in random perforated domains of general type*, Mat. Zametki, 53 (1993), pp. 41–58.
[36] V. V. Zhikov, *On homogenization of nonlinear variational problems in perforated domains*, Russian J. Math. Phys., 2 (1994), pp. 393–408.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1528–1541

LOCAL SMOOTHING AND LOCAL SOLVABILITY FOR THIRD ORDER DISPERSIVE EQUATIONS*

HERBERT KOCH† AND JEAN-CLAUDE SAUT‡

Abstract. Two-dimensional deep water waves can be described by various third order dispersive equations, modifying and generalizing the KdV equations as well as nonlinear Schrödinger equations. We establish local well-posedness for initial data $u_0 \in H^{3/2}(\mathbb{R}^2)$ for many of the proposed models using local smoothing estimates with respect to the forcing term and the initial data.

Key words. third order dispersive equations, local smoothing, nonlinear waves

AMS subject classifications. 35Q55, 35Q35, 35Q60, 35B65

DOI. 10.1137/050630659

1. Introduction. The weakly nonlinear dynamics of a quasi-monochromatic wave train propagating at the surface of water can be described by nonlinear Schrödinger-type equations. This has been established in one dimension and two dimensions by Zakharov [19] for infinitely deep water and by Hasimoto and Ono [10] in the one-dimensional finite depth case. They obtained the cubic nonlinear Schrödinger equation. For nearly one-dimensional wave trains in finite depth water, Davey and Stewartson [4] and Djordjevic and Redekopp [7] derived the so-called Davey–Stewartson systems for, respectively, purely gravity and gravity-capillary waves. Those systems involve a wave-induced mean flow resulting in a nonlocal nonlinear Schrödinger equation. They reduce to the cubic nonlinear Schrödinger equation in the infinitely deep water limit. Let a be a typical wave amplitude and k be the modulus of the mean wave number. All the models above are found by a perturbation analysis up to $O(\varepsilon^3)$ when $\varepsilon = k\,a \ll 1$ is the wave steepness.

Taking perturbation analysis one step further to $O(\varepsilon^4)$, Dysthe [8] has derived a system which improves significantly upon the results on the stability of finite amplitude waves in infinite depth. One of the dominant new effects is the wave-induced mean flow with potential Φ. Solving the equation for Φ in terms of the complex amplitude of the wave packet allows one to put the Dysthe system (in dimensionless variables) in the following form (see [9]):

$$\begin{aligned} 2i\left(\frac{\partial A}{\partial t}+\frac{1}{2}\frac{\partial A}{\partial x}\right)+\frac{1}{2}\frac{\partial^2 A}{\partial y^2}-\frac{1}{4}\frac{\partial^2 A}{\partial x^2}-A|A|^2=\frac{i}{8}\left(\frac{\partial^3}{\partial x^3}-6\frac{\partial^3}{\partial x\partial y^2}\right)A \\ +\frac{i}{2}A\left(A\frac{\partial\overline{A}}{\partial x}-\overline{A}\frac{\partial A}{\partial x}\right)-\frac{5i}{2}|A|^2\frac{\partial A}{\partial x}+AR_1\frac{\partial}{\partial x}|A|^2, \end{aligned} \tag{1.1}$$

where R_1 is the Riesz transform in $\mathbb{R}^2$, that is,

$$F(R_1\psi)=i\frac{\xi_1}{|\xi|}\hat{\psi}.$$

Here F and $\hat{\psi}$ denote the Fourier transform.

*Received by the editors May 4, 2005; accepted for publication (in revised form) September 5, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sima/38-5/63065.html
†Mathematisches Institut, Universität Bonn, Beringstr. 1, 35115 Bonn, Germany (koch@math.uni-bonn.de).
‡UMR de Mathematiques, Bât. 425, Université de Paris-Sud, 91405 Orsay, France (Jean-Claude.saut@math.u-psud.fr).

The usual nonlinear Schrödinger equation (NLS) is obtained by neglecting the right-hand side of (1.1), which is of order ε^4 in the dimensional variables. Note that going one step further in the perturbation introduces higher dispersive terms as well as new nonlinear ones.

A similar derivation of the fourth order (in ε) evolution equations for the amplitude of a train of nonlinear gravity-capillary waves on the surface of an ideal fluid of infinite depth was performed by Hogan [11]. The equation reads

$$\begin{aligned} 2i\left(\frac{\partial A}{\partial t}+c_g\frac{\partial A}{\partial x}\right)+p\frac{\partial^2 A}{\partial x^2}+q\frac{\partial^2 A}{\partial y^2}-\gamma|A|^2A=-is\frac{\partial^3 A}{\partial x\partial y^2}-ir\frac{\partial^3 A}{\partial x^3} \\ -iuA^2\frac{\partial\overline{A}}{\partial x}+iv|A|^2\frac{\partial A}{\partial x}+AR_1\frac{\partial}{\partial x}|A|^2, \end{aligned} \tag{1.2}$$

where c_g is the group velocity and γ, p, q, s, r, u, and v are real parameters depending on the surface tension parameter. Note that q and s are strictly positive, while p can achieve both signs (in particular, the sign is negative for purely gravity waves as in the Dysthe system and positive for purely capillary waves). We refer to the survey of Dias and Kharif [6] for a more complete description of the models of water waves.

Similar problems occur in nonlinear optics (see, for instance, [20]), in particular in the modeling of the dynamics of femtosecond laser pulses in a nonlinear media with temporal dispersion. The evolution of the complex envelope $E(x,y,z,t)$ of the field is described by the third order NLS

$$i\frac{\partial}{\partial z}E+(1-i\varepsilon_1\partial_t)\Delta_\perp E-\frac{\partial^2 E}{\partial t^2}-i\varepsilon_2\frac{\partial^3 E}{\partial t^3}+(1+i\varepsilon_1\partial_t)g(|E|^2)E=0, \tag{1.3}$$

where $\Delta_\perp=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}$ is the transverse Laplacian and $\varepsilon_2\in\mathbb{R}$, $\varepsilon_1>0$, and typically, $g(z)=z$. As is usual in nonlinear optics, the evolution variable (which plays, mathematically, the role of time, whereas t will be treated as spatial variable later) is z. The transverse Laplacian accounts for diffraction, while the second and third time derivatives describe group velocity and third order dispersion.

Very little is known concerning the Cauchy problem for (1.1)–(1.3). The only results available are local existence for analytic Cauchy data for (1.1) and (1.2) by de Bouard [5], and, very recently, well-posedness in $H^3(\mathbb{R}^2)$ by Chihara [3], who applies techniques developed for derivative nonlinear Schrödinger equations to these third order equations.

Sharp estimates on the fundamental solution of the linearization of (1.2) at 0 have been recently derived in [2]. They lead naturally to Strichartz estimates. Because of the presence of derivatives in the nonlinear terms, Strichartz estimates do not suffice by themselves to obtain the local well-posedness of (1.2) by a contraction argument.

The aim of this paper is to give a proof of local well-posedness of (1.2) and (1.3) in two space dimensions for data in the Sobolev space $H^{3/2}$ by a contraction argument involving spaces related to the local smoothing properties of the group associated to the linear equation

$$\begin{aligned} iu_t+P(D)u &= f && \text{in } \mathbb{R}^n\times\mathbb{R}, \\ u(x,0) &= u_0(x) && \text{for } x\in\mathbb{R}^n, \end{aligned} \tag{1.4}$$

where $P(D)$ is a differential operator of order $\kappa\geq 2$ with real symbol $p(\xi)$. Let p_κ be the part of degree κ.

Assumption. We assume that there exist $c_1, c_2 > 0$ such that

$$|\nabla p(\xi)| \geq c_1 |\xi|^{\kappa-1} - c_2. \tag{1.5}$$

Straightforward calculations show that (1.5) is satisfied for problem (1.1). It is satisfied for (1.2) provided s and r are nonzero and for (1.3) if ε_1 and ε_2 are nonzero.

Let

$$\tilde{Q}_x = \left\{ y | \max_j |x_j - y_j| \leq \frac{1}{2} \right\}$$

and, for $T > 0$,

$$Q_{x,T} = \tilde{Q}_x \times [0, T].$$

The local smoothing estimate is the following.

PROPOSITION 1.1. *Suppose that p satisfies* (1.5). *Let $T > 0$. Then we have*

$$\sup_x \|(1-\Delta)^{\frac{\kappa-1}{4}} u\|_{L^2(Q_{x,T})} + \sup_{0\leq t\leq T} \|u(t)\|_{L^2} \leq c \left(\|u_0\|_{L^2} + \sum_{k\in\mathbb{Z}^n} \|(1-\Delta)^{-\frac{\kappa-1}{4}} f\|_{L^2(Q_{k,T})} \right).$$

Note that the cubes $Q_{k,T}$ cover $\mathbb{R}^n \times (0,T)$. We apply it with $\kappa = 3$ and $n = 2$ to construct a solution to

$$\begin{array}{rcll} iu_t + P(D)u &=& f(u, \partial_\nu u) & \text{in } \mathbb{R}^2 \times \mathbb{R}, \\ u(x,0) &=& u_0(x) & \text{for } x \in \mathbb{R}^2, \end{array} \tag{1.6}$$

where P is of order 3, ν is a unit vector in $\mathbb{R}^2$, and, with the Riesz transform R_ν defined by the Fourier multiplier $\frac{\xi\cdot\nu}{|\xi|}$ and with real or complex constants a_j,

$$f(u, \partial_\nu u) = a_0 |u|^2 u + a_1 u^2 \partial_\nu \bar{u} + a_2 |u|^2 \partial_\nu u + a_3 u \partial_\nu R_\nu |u|^2. \tag{1.7}$$

Note that in (1.3) z becomes the time variable and t a spatial variable. In most models the L^2 norm is preserved. This is the case when

$$a_0, a_3 \in \mathbb{R}, \qquad a_1, a_2 \in i\mathbb{R}. \tag{1.8}$$

We fix a smooth compactly supported function ϕ, identically 1 on $\tilde{Q}_0$.

DEFINITION 1.2. *We define the function spaces $X_{t_0} \subset C([0,t_0]; H^{3/2}(\mathbb{R}^2))$ and Y_{t_0} through the norms*

$$\begin{aligned} \|u\|_{X_{t_0}} = {} & \sup_{0\leq t\leq t_0} \|u(t)\|_{H^{3/2}(\mathbb{R}^2)} + \sup_{x\in\mathbb{R}^2} \|(1-\Delta)_x^{5/4} u\|_{L^2(Q_{x,t_0})} \\ & + t_0^{-1/2} \left(\sum_{k\in\mathbb{Z}^2} \sup_{0\leq t\leq t_0} \left(\|\phi(.-k)u(t)\|^2_{H^{1/2}(\mathbb{R}^2)} - \|\phi(.-k)u(0)\|^2_{H^{1/2}(\mathbb{R}^2)} \right) \right)^{1/2} \end{aligned} \tag{1.9}$$

and

$$\|f\|_{Y_{t_0}} = \inf_{f_1+f_2=f} \left\{ \sum_{k\in\mathbb{Z}^2} \|(1-\Delta)^{-1/4} f_1\|_{L^2(Q_{k,t_0})} + \int_0^{t_0} \|f_2(\tau)\|_{H^{3/2}(\mathbb{R}^2)} d\tau \right\}.$$

After these preparations we can formulate the main result.

THEOREM 1.3. *Suppose that p satisfies* (1.5) *and that the nonlinearity is of the form* (1.7). *Given $c > 1$ there exists $t_0 \sim c^{-2}$ such that for $u_0 \in H^{3/2}(\mathbb{R}^2)$ with $\|u_0\|_{H^{3/2}} \leq c$ there exists a unique solution $u \in X_{t_0}$ to* (1.6) *and* (1.7) *with initial value u_0. Moreover, the map $H^{\frac{3}{2}}(\mathbb{R}^2) \to X_{t_0}, u_0 \to u$ is analytic. In the particular case of* (1.2) *and* (1.3) *we have, furthermore, $\|u(t)\|_{L^2(\mathbb{R}^2)} = \|u_0\|_{L^2(\mathbb{R}^2)}$.*

Proposition 1.1 is proved in section 2 by a reduction to a one-dimensional estimate. In section 3 we prove the main result, which implies in particular that the Cauchy problem for (1.1), for (1.2) if s and r are nonzero, and for (1.3) if ε_1 and ε_2 are nonzero and $g(z) = z$, is locally well-posed for data in $H^{3/2}(\mathbb{R}^2)$. It is likely that one could consider data in Sobolev spaces of lower order by using the Strichartz estimates of [1] and maximal function estimates. Note that for solutions of (1.2) and (1.3) the L^2 norm is formally conserved, which could possibly be used to get global existence from local existence in $L^2(\mathbb{R}^2)$ since $L^2(\mathbb{R}^2)$ is a critical space for scaling. The dispersive estimates obtained in [2] might be useful to establish global existence for small data.

2. Local smoothing estimates. In this part we consider dispersive equations in all space dimensions and of all orders. Local smoothing estimates with respect to the initial data for general dispersive equations are well known; see, for example, [13]. There are fewer results concerning the smoothing property with respect to the forcing term f. As far as we know, such results have been established in the case when $\kappa = 2$ (Schrödinger-type equations) [18, 16], in one space dimension for all κ [15, 14, 17], and more recently, by Hoshiro [12] who established related local smoothing estimates using arguments which are very different from those in this section.

Let a, b, $s \in \mathbb{R}$ and $b \neq 0$. Then by the residue theorem we have

$$\frac{1}{2\pi i} \int_{\mathbb{R}} \frac{e^{is\sigma}}{\sigma - a - ib}\, d\sigma = \begin{cases} 0 & \text{if } bs < 0, \\ e^{ias - bs} & \text{if } bs > 0. \end{cases} \tag{2.1}$$

This simple identity has important and simple consequences for similar integrals for polynomials.

LEMMA 2.1. *Let $p(\sigma)$ be a real polynomial of degree κ. Then for all real $\varepsilon \neq 0$,*

$$\left| \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{e^{is\sigma} p'(\sigma)}{p(\sigma) - i\varepsilon}\, d\sigma \right| \leq \kappa.$$

Proof. We may assume without loss of generality that the degree κ is at least 1. Let $b_i \in \mathbb{C}$, $i = 1, \ldots, m \leq \kappa$, be the zeros of the polynomial $p - i\varepsilon$ of multiplicity N_i. Then

$$\frac{p'}{p - i\varepsilon} = \sum_{i=1}^{m} \frac{N_i}{s - b_i}.$$

No zero of $p - i\varepsilon$ is real, and the assertion follows from the previous calculation (2.1).

There is even a precise formula for $\varepsilon \to 0$. Let $H(s) = 1$ for $s < 0$ and 0 otherwise. Then

$$\lim_{\varepsilon \to 0+} \int_{\mathbb{R}} \frac{e^{is\sigma} p'(\sigma)}{p(\sigma) - i\varepsilon}\, d\sigma = \sum_{j=1}^{k_1} N_j e^{ia_j s} H(\pm s) + \sum_{\{j : \operatorname{Im} a_j s > 0\}} N_j e^{ia_j s},$$

where the first sum is over all zeros on the real line and the second one over all zeros where the imaginary part is positive. The sign in $H(\pm s)$ is determined by the sign of the imaginary part of the root of $p - i0$. □

Let p be a real polynomial of degree κ in $\mathbb{R}^n$. We denote the group generated by $P(D)$ by $S(t)$, which has the kernel function

$$k(x,t) = \begin{cases} 0 & \text{if } t \le 0, \\ \frac{1}{(2\pi)^n} \int e^{ix\xi + itp(\xi)}\, d\xi & \text{otherwise,} \end{cases}$$

where the integral is a standard oscillatory integral, and

$$Kf(x,t) = \int_{-\infty}^{t} \int_{\mathbb{R}^n} k(x-y, t-s) f(y,s) dy\, ds.$$

Then

$$i(\tau - p(\xi))\hat{Kf}(\xi,\tau) = \hat{f}(\xi,\tau)$$

and

$$k(x,t) = \lim_{\varepsilon > 0, \varepsilon \to 0} F^{-1}\Big(\frac{1}{\tau + p(\xi) - i\varepsilon}\Big).$$

In what follows we fix a unit vector $v \in \mathbb{R}^n$ and define for $r \in \mathbb{R}$ the n-dimensional hyperplane

$$\mathcal{H}^v_r = \{(x,t) | \langle v, x\rangle = r\}. \tag{2.2}$$

Let u and f be measurable functions. We define function spaces by their norm

$$\begin{aligned} \|u\|_{L^\infty(\mathbb{R}, L^2(\mathcal{H}^v))} &= \operatorname*{ess\,sup}_r \|u|_{\mathcal{H}^v_r}\|_{L^2(\mathcal{H}^v_r)}, \\ \|f\|_{L^1(\mathbb{R}, L^2(\mathcal{H}^v))} &= \int_{\mathbb{R}} \|u|_{\mathcal{H}^v_{r'}}\|_{L^2(\mathcal{H}^v_{r'})} dr'. \end{aligned}$$

Let p_v be the derivative of the symbol p in direction v.

LEMMA 2.2. *The following inequality holds for all τ and all ξ perpendicular to v and ε small:*

$$\Big\| F^v\Big[\frac{p_v}{\tau + p(\xi) - i\varepsilon}\Big]\Big\|_{L^\infty(\mathbb{R})} \le \sqrt{2\pi}\kappa,$$

where F^v is the partial inverse Fourier transform in the direction of v.

Proof. We fix $(\xi,\tau) \in \mathcal{H}^v$ and set

$$q(\sigma) = \tau + p(\xi + \sigma v)$$

which is a polynomial of order at most κ. The assertion follows now from Lemma 2.1. □

THEOREM 2.3. *Let $A = p_v(D)$ be the differential operator defined by the Fourier multiplier p_v. Suppose that $f \in L^1(\mathbb{R}; L^2(\mathcal{H}^v_r))$. Then*

$$\sup_r \|(AKf)|_{\mathcal{H}^v_r}\|_{L^2(\mathcal{H}^v_r)} \le \sqrt{2\pi}\kappa \int_{\mathbb{R}} \|f|_{\mathcal{H}^v_{r'}}\|_{L^2(\mathcal{H}^v_{r'})}\, dr'.$$

Proof. The Fourier transform of K is

$$\lim_{\varepsilon \to 0} \frac{1}{\tau + p(\xi) - i\varepsilon}.$$

The Fourier transform in the direction v satisfies the desired bounds. Hence, if $f \in L^2(\mathcal{H}^v)$ and $r', r \in \mathbb{R}$, then

$$\|K(\delta_{r'}(.)f)|_{\mathcal{H}^v_r}\|_{L^2(\mathcal{H}^v_r)} \le \sqrt{2\pi}\kappa\|f|_{\mathcal{H}^v_{r'}}\|_{L^2(\mathcal{H}^v_{r'})},$$

where δ denotes the Dirac measure, because this map is defined by a bounded Fourier multiplier. Integration with respect to r' yields the desired result. □

We consider the initial boundary value problem

$$i\partial_t u - P(D)u = f, \qquad u(.,0) = u_0,$$

where we assume that

$$|p_v(D)|^{-1/2} f \in L^1(\mathbb{R}; L^2(\mathcal{H}^v))$$

and $u(0) = u_0 \in L^2$. If $f = 0$, then there is a unique solution in $C(\mathbb{R}; L^2(\mathbb{R}^n))$. Then formally the solution for positive t is given by

$$u(t) = k(.,t) * u_0 + \int_0^t k(.,t-s) * f(.,s)ds.$$

In what follows we study this solution.

PROPOSITION 2.4. *We have*

$$\sup_t \|u(t)\|_{L^2} + (2\pi)^{-1/4}\kappa^{-1/2}\||p_v|^{1/2}(D)u\|_{L^\infty(\mathbb{R},L^2(\mathcal{H}^v))} \le \|u_0\|_{L^2} + (2\pi)^{1/4}\sqrt{\kappa}\||p_v|^{-1/2}(D)f)\|_{L^1(\mathbb{R},L^2(\mathcal{H}^v))}.$$

It is part of the statement that the left-hand side is finite if the right-hand side is finite.

Proof. There are four estimates to prove. We suppose that f is defined on $\mathbb{R}^n \times \mathbb{R}$. The estimate

$$(2\pi)^{-1/4}\kappa^{-1/2}\||p_v|^{1/2}(D)u\|_{L^\infty(\mathbb{R},L^2(\mathcal{H}^v))} \le (2\pi)^{1/4}\sqrt{\kappa}\||p_v|^{-1/2}(D)f\|_{L^1(\mathbb{R},L^2(\mathcal{H}^v))}$$

for $u(t) = \int_{-\infty}^t S(t-s)f(s)\,ds$ is the content of Theorem 2.3.

The estimate

$$\sup_t \|u(t)\|_{L^2} \le (2\pi)^{1/4}\sqrt{\kappa}\||p_v|^{-1/2}(D)f\|_{L^1(\mathbb{R},L^2(\mathcal{H}^v))}$$

follows from

$$\|u(0)\|_{L^2} \le (2\pi)^{1/4}\sqrt{\kappa}\||p_v|^{-1/2}(D)f\|_{L^1(\mathbb{R},L^2(\mathcal{H}^v))} \tag{2.3}$$

for $u = Kf$. More precisely we define

$$T_1 : L^1(\mathbb{R}; L^2(\mathcal{H}^v \cap \{t < 0\})) \to L^2(\mathbb{R}^n),$$
$$T_1 f = (Kf)(0),$$

and

$$T_2 : L^2(\mathbb{R}^n) \to L^\infty(\mathbb{R}; L^2(\mathcal{H}^v \cap \{t < 0\})),$$
$$T_2(u_0)(t) = S(-t)u_0.$$

Then

$$T_2 = T_1^*$$

since

$$\begin{aligned}\int_{\mathbb{R}^n}\int_{-\infty}^{0} T_2 v\overline{f}\,dt\,dx &= \int_0^\infty \int_{\mathbb{R}^n} S(-t)v\overline{f(-t)}dxdt\\ &= \int_{\mathbb{R}^n} v\int_0^\infty \overline{S(t)f(-t)}dt\,dx\\ &= \int_{\mathbb{R}^n} v\overline{T_1(f)}dx.\end{aligned}$$

We observe that the kernel of K is supported in the half space $t \le 0$, and hence we may assume that f is supported in the same half space. On the other hand, if f is supported in the half space $\{t \le 0\}$, and $t > 0$, then

$$Kf(t) = T_2T_1f(-t).$$

Hence, the bounds for T_1 and T_2 follow by an application of Theorem 2.3.

Now we define

$$v(t) = u(t) - \int_{-\infty}^{t} S(t-s)f(s)\,ds,$$

and we obtain

$$i\partial_t v - P(D)v = 0,$$

$$v(0) = u_0 - \int_{-\infty}^{0} S(-s)f(s)\,ds.$$

The theorem is proved if we show $\|v(t)\|_{L^2} = \|v(0)\|_{L^2}$ and

$$\|v\|_{L^\infty(\mathbb{R},L^2(\mathcal{H}^v)\cap\{t>0\})} \le (2\pi)^{1/4}\sqrt{\kappa}\|v(0)\|_{L^2}.$$

The first estimate follows from the energy equality and the second from the estimate for T_2. □

In what follows we will need an elementary estimate. We study

$$i\partial_t v - P(D)v = f,$$

where p is a real polynomial of order κ in n variables. We recall that

$$\tilde{Q}_x := \left\{ y \in \mathbb{R}^2 : \max\{|x_i - y_i|\} \le \frac{1}{2} \right\} \quad \text{and } Q_{x,t_0} = \tilde{Q}_x \times (0,t_0).$$

LEMMA 2.5. *Let $\phi \in C_0^\kappa(\mathbb{R}^n)$ be supported in $B_R(x)$. Then*

$$\left|\|\phi v(t)\|_{L^2}^2 - \|\phi v(0)\|_{L^2}^2\right| \le 4t\int_0^t \|\phi f(s)\|_{L^2}^2 ds + c\int_0^t \|v(s)\|^2_{H^{\frac{\kappa-1}{2}}(B_R(x))}\,ds.$$

Proof. We write

$$\begin{aligned}\|\phi v(t)\|_{L^2}^2 - \|\phi v(0)\|_{L^2}^2 = &\int_0^t \langle \phi v(s), -i\phi f(s) - i\phi P(D)v(s)\rangle\\ &+ \langle -i\phi f(s) - i\phi P(D)v(s), \phi v(s)\rangle ds\end{aligned}$$

and estimate

$$\int_0^t |\langle \phi v(s), \phi f(s)\rangle| ds \le 2t \int_0^t \|\phi f(s)\|_{L^2}^2 + (8t)^{-1} \int_0^t \|\phi v(s)\|_{L^2}^2,$$

where the last term is controlled by $\sup_{0\le s\le t} 8^{-1}\|\phi v(s)\|_{L^2}^2$. Moreover, since $P(D)$ is self-adjoint,

$$\begin{aligned}\langle \phi v(s), \phi P(D) v(s)\rangle - \langle \phi P(D) v(s), \phi v(s)\rangle \\ = \langle \phi v(s), [\phi, P(D)] v(s)\rangle - \langle [\phi, P(D)] v(s), \phi v(s)\rangle.\end{aligned}$$

Clearly $[\phi, P(D)]$ is an operator of order $\kappa - 1$, and we can distribute the derivatives on both terms. □

Now we are in a position to formulate the local smoothing estimate.

PROPOSITION 2.6. *Suppose that p satisfies* (1.5). *Let $T > 0$. Then we have*

$$\sup_x \|(1-\Delta)^{\frac{\kappa-1}{4}} u\|_{L^2(Q_{x,T})} + \sup_{0\le t\le T} \|u(t)\|_{L^2} \lesssim \|u_0\|_{L^2} + \sum_{k\in\mathbb{Z}^n} \|(1-\Delta)^{-\frac{\kappa-1}{4}} f\|_{L^2(Q_{k,T})}.$$

Proof. We fix a finite number of unit vectors v_i, $1 \le i \le M$, such that

$$\sum_{j=1}^M |p_{v_j}| \ge \delta |\xi|^{\kappa-1}$$

for ξ sufficiently large. We choose $\eta \in C^\infty(\mathbb{R})$ to be identically 0 in $[-\delta/n, \delta/n]$, 1 outside $[-2\delta/n, 2\delta/n]$, and define

$$p_j(\xi) = p_{v_j}(\xi)\eta(p_{v_j}(\xi)(1+|\xi|^2)^{1-\kappa}).$$

By construction,

$$\delta|\xi|^{\kappa-1} \le n \sum |p_j(\xi)|$$

for $|\xi| \gg 1$. This inequality may fail, however, in a bounded set. If $p_j(\xi)$ is small, then we do not obtain a gain in the local smoothing estimate. We use the function η to localize the local smoothing estimates to values of ξ, where we gain in the local smoothing estimate.

By the previous arguments, we may assume $u(0) = 0$ because we obtain the dependence on the initial data by a T-T^* argument, as above, from the smoothing with respect to forcing. Let x be given. We fix a smooth function ψ, identically 1 on $\tilde{Q}_x$, supported in a hypercube of twice the size of $\widetilde{Q}_x$. Then it suffices to bound

$$\|(1-\Delta)^{\frac{\kappa-1}{4}}(\psi u)\|_{L^2((0,T)\times\mathbb{R}^n)}.$$

Clearly for fixed t, with $\bar{Q}_x$, respectively, $\bar{Q}_{x,T}$, the hypercube of twice the spatial scale of $\tilde{Q}_x$, and F the spatial Fourier transform,

$$\begin{aligned}\|(1-\Delta)^{\frac{\kappa-1}{4}}(\psi u)\|_{L^2(\mathbb{R}^n)}^2 &= \langle |\xi|^{\kappa-1} F(\psi u), F(\psi u)\rangle \\ &\lesssim \sum_j \langle |p_j(D_x)|(\psi u), \psi u\rangle + \|(1-\Delta)^{-\frac{\kappa-1}{4}} \psi u\|_{L^2} \\ &\lesssim \sum_j \|\psi |p_j|^{1/2} u\|_{L^2(\mathbb{R}^n)}^2 + \|(1-\Delta)^{\frac{\kappa-3}{4}} u\|_{L^2(\bar{Q}_x)}^2.\end{aligned} \tag{2.4}$$

Integration over t together with the bound of Theorem 2.3 gives

$$\begin{aligned}\|(1-\Delta)^{\frac{\kappa-1}{4}}\psi u\|_{L^2(\bar{Q}_{x,T})} \lesssim & \|(1-\Delta)^{\frac{\kappa-3}{4}}u\|_{L^2(\bar{Q}_{x,T})} \\ & + \sum_{k\in\mathbb{Z}^2}\sum_j \|(|p_j|^{1/2}/|p_{v_j}|)(D_x)|f\|_{L^2(Q_{k,T})} \\ \lesssim & \|(1-\Delta)^{\frac{\kappa-3}{4}}u\|_{L^2(\bar{Q}_{x,T})} + \sum_j\sum_{k\in\mathbb{Z}^2}\|(1-\Delta)^{-\frac{\kappa-1}{4}}f\|_{L^2(Q_{k,T})}.\end{aligned}$$

Now, with H^s the standard Sobolev space, we denote the space of function with values in $H^s(Q_x)$ by $L^2(H^s(Q_{x,T}))$. Since the kernel of $1-\Delta$ decays exponentially, and by compact embeddings, we obtain for some $\delta>0$ and all $\varepsilon>0$

$$\begin{aligned}\|(1-\Delta)^{\frac{\kappa-3}{4}}u\|_{L^2(Q_{x,T})} \le & c_1\sum_{k\in\mathbb{Z}^2} e^{-\delta|x-k|}\|u\|_{L^2(H^{\frac{\kappa-3}{2}}(Q_{k,T}))} \\ \le & c_2\sup_{k\in\mathbb{Z}^2}\|u\|_{L^2(H^{\frac{\kappa-3}{2}}(Q_{x+k,T}))} \\ \le & \sup_k \varepsilon\|u\|_{L^2(H^{\frac{\kappa-1}{2}}(Q_{k,T}))} + c_\varepsilon\|u\|_{L^2(H^{-\frac{\kappa-1}{2}}(Q_{k,T}))} \\ \le & \sup_k \varepsilon\|(1-\Delta)^{\frac{\kappa-1}{4}}u\|_{L^2(Q_{k,T})} + c_\varepsilon\|(1-\Delta)^{-\frac{\kappa-1}{4}}u\|_{L^2(Q_{k,T})}.\end{aligned}$$

We take the sup with respect to x and choose ε small to arrive at

$$\begin{aligned}\sup_x\|(1-\Delta)^{\frac{\kappa-1}{4}}u\|_{L^2(Q_{x,T})} \lesssim & \sup_x\|(1-\Delta)^{-\frac{\kappa-1}{4}}u\|_{L^2(Q_{x,T})} \\ & + \sum_{k\in\mathbb{Z}^n}\|(1-\Delta)^{-\frac{\kappa-1}{4}}f\|_{L^2(Q_{k,T})}.\end{aligned}$$

Using Lemma 2.5 for $(1-\Delta)^{-\frac{\kappa-1}{4}}u$, since $u(0)=0$, we may bound

$$\begin{aligned}\sup_x\|(1-\Delta)^{-\frac{\kappa-1}{4}}u\|_{L^2(Q_{x,T})} & \lesssim T^{\frac12}\sup_x\sup_t\|(1-\Delta)^{-\frac{\kappa-1}{4}}u(t)\|_{L^2(Q_x)} \\ & \lesssim T^{1/2}\sup_x\|(1-\Delta)^{-\frac{\kappa-1}{4}}f\|_{L^2(Q_{x,T})} + T^{1/2}\sup_x\|u\|_{L^2(Q_{x,T})} \\ & \lesssim T^{1/2}\sup_x\|(1-\Delta)^{-\frac{\kappa-1}{4}}f\|_{L^2(Q_{x,T})} + T^{1/2}\sup_x\|(1-\Delta)^{\frac{\kappa-1}{4}}u\|_{L^2(Q_{x,T})}.\end{aligned}$$

We choose T small to absorb the last term and obtain

$$\sup_x\|(1-\Delta)^{\frac{\kappa-1}{4}}u\|_{L^2(Q_{x,T})} \lesssim \|u_0\|_{L^2(\mathbb{R}^n)} + \sum_{k\in\mathbb{Z}^2}\|(1-\Delta)^{-\frac{\kappa-1}{4}}f\|_{L^2(Q_{k,T})}.$$

Next we shall add $\sup_t\|u(t)\|_{L^2}$ on the left-hand side. That this can be done follows from the standard energy estimate if $f=0$; hence we can reduce our attention to the case $u_0=0$. Let $\tilde{Y}$ and $\tilde{X}$ be the spaces defined by $\sum_k\|(1-\Delta)^{-\frac{\kappa-1}{4}}f\|_{L^2(Q_{k,T})}$ and $\sup_k\|(1-\Delta)^{\frac{\kappa-1}{4}}u\|_{L^2(Q_{k,T})}$, respectively. Let $K: L^2\ni u_0\to u\in\tilde{X}$ be the map from u_0 to the solution u to the homogeneous equation with initial data u_0. Its adjoint is the map $K^*: Y\ni f\to u(T)\in L^2$ which maps the right-hand side to the solution with zero initial data evaluated at time T. Thus K^* is bounded since K is bounded if T is small. The arguments work uniformly for small T and we obtain the claimed bound for small T. This last restriction is removed by an iterative application of the estimate on small time intervals. □

From now on we restrict ourselves to $\kappa=3$ and $n=2$.

LEMMA 2.7. *Suppose that* $|\nabla p(\xi)| \sim |\xi|^2$ *for large* ξ,

$$iu_t - P(D)u = f \text{ in } \mathbb{R}^2 \times (0,T), \tag{2.5}$$

and $u(0) = u_0 \in H^{3/2}(\mathbb{R}^2)$. *Then*

$$\|u\|_{X_{t_0}} \lesssim \|u_0\|_{H^{3/2}(\mathbb{R}^2)} + \|f\|_{Y_{t_0}}.$$

Proof. Proposition 2.6 implies

$$\sup_{0\le t\le t_0} \|u\|_{H^1} + \sup_{x\in\mathbb{R}^2} \|D_x^2 u\|_{L^2(Q_{x,t_0})} \lesssim \|D_x u_0\|_{L^2} + \inf_{f_1+f_2=f} \left(\sum_{k\in\mathbb{Z}^2} \|f_1\|_{L^2(Q_{k,t_0})} + \int_0^{t_0} \|f_2(t)\|_{L^2} \right).$$

We apply $(1-\Delta)^{1/4}$ to (2.5) and obtain

$$\sup_{0\le t\le t_0} \|u\|_{H^{3/2}(\mathbb{R}^2)} + \sup_{x\in\mathbb{R}^2} \|(1-\Delta)^{5/4} u\|_{L^2(Q_{x,t_0})} \lesssim \|u_0\|_{H^{3/2}(\mathbb{R}^2)} + \|f\|_{Y_{t_0}}.$$

We combine this inequality with Lemma 2.5 and choose $\phi \in C_0^\infty(\mathbb{R}^2)$, which is identically 1 in $[-1,1]^2$ and supported in $[-2,2]^2$. Then by Lemma 2.5

$$\sum_{k\in\mathbb{Z}^2} \left| \|\phi(x-k)u(t)\|_{L^2}^2 - \|\phi(x-k)u(0)\|_{L^2}^2 \right| \lesssim t \int_0^t \|f(s)\|_{L^2}^2 ds + \int_0^t \|u\|_{H^1(\mathbb{R}^n)}^2 ds.$$

We apply this estimate to $(1-\Delta)^{1/4}u$. The commutators $[(1-\Delta)^{1/4}, \phi]$ are of order $-1/2$. Their kernels decay quickly and the corresponding terms can be easily controlled. Hence,

$$t_0^{-1} \sum_{k\in\mathbb{Z}^2} \left| \|\phi(x-k)u(t)\|_{H^{1/2}(\mathbb{R}^2)}^2 - \|\phi(x-k)u(0)\|_{H^{1/2}(\mathbb{R}^2)}^2 \right| \lesssim \int_0^t \|f(s)\|_{H^{1/2}(\mathbb{R}^2)}^2 ds + \sup_{0\le t\le t_0} \|u(t)\|_{H^{3/2}(\mathbb{R}^2)}^2.$$

This completes the proof. □

3. The main result. Here we study the problem

$$iu_t - P(D)u = f(u), \qquad u(x,0) = u_0(x),$$

where the symbol $p(\xi)$ satisfies (1.5) and where f is as in (1.7). We can write the problem in the form

$$f(u) = F[u,u,u]$$

with

$$F[u,v,w] = \frac{1}{6}\left(f(u+v+w) + f(u) + f(v) + f(w) - f(u+v) - f(u+w) - f(v+w)\right),$$

where F is symmetric and linear in each argument. A typical term is

$$F[u,v,w] = u\partial_\nu R_\nu(vw) + v\partial_\nu R_\nu(wu) + w\partial_\nu R_\nu(uv). \tag{3.1}$$

After these preparations we turn to the proof of Theorem 1.3.

Proof. We will choose $w_0 \in H^4(\mathbb{R}^2)$ below and define w as the solution to

$$iw_t - P(D)w = 0, \qquad u(0,x) = w_0(x).$$

Instead of u we search the solution to

$$iv_t - P(D)v = f(v+w), \qquad v(0) = u_0 - w_0$$

by a fixed point argument. As a by-product of the proof we will obtain Lipschitz—and even analytic—dependence on the initial data.

Clearly by Sobolev's embedding

$$\|D_x^2 w\|_{L^\infty(\mathbb{R}\times\mathbb{R}^2)} \lesssim \|w_0\|_{H^4(\mathbb{R}^2)}.$$

Let $c_0 = \|u_0\|_{H^{3/2}}$. We choose ε small and $w_0 \in H^4(\mathbb{R}^2)$ with $\|u_0 - w_0\|_{H^{1/2}} \le \varepsilon$, $\|u_0 - w_0\|_{H^{3/2}} \lesssim 2c_0$, and $\|w_0\|_{H^4(\mathbb{R}^2)} \lesssim c_0^{5/2}\varepsilon^{-3/2}$. The result will follow from Lemma 2.7 and from the following.

LEMMA 3.1. *Suppose that $f = f(u, \nabla u)$ is of the form* (1.7). *Then, with F related to f as in* (3.1), *if $t_0 \le 1$,*

$$\begin{aligned}\|F[u,v,w]\|_{Y_{t_0}} \lesssim &\Big(\|u(0)\|_{H^{1/2}} + t_0^{\frac12}\|u\|_{X_{t_0}}\Big)\Big(\|v(0)\|_{H^{1/2}} + t_0^{\frac12}\|v\|_{X_{t_0}}\Big)\|w\|_{X_{t_0}}\\ &+ \Big(\|v(0)\|_{H^{1/2}} + t_0^{\frac12}\|v\|_{X_{t_0}}\Big)\Big(\|w(0)\|_{H^{1/2}} + t_0^{\frac12}\|w\|_{X_{t_0}}\Big)\|u\|_{X_{t_0}}\\ &+ \Big(\|w(0)\|_{H^{1/2}} + t_0^{\frac12}\|w\|_{X_{t_0}}\Big)\Big(\|u(0)\|_{H^{1/2}} + t_0^{\frac12}\|u\|_{X_{t_0}}\Big)\|v\|_{X_{t_0}}.\end{aligned}$$

Proof. It suffices to show that

$$\|f(u)\|_{Y_{t_0}} \le c\Big(\|u(0)\|_{H^{1/2}}^2 + t_0\|u\|_{X_{t_0}}^2\Big)\|u\|_{X_{t_0}}. \tag{3.2}$$

This can be seen either by polarization or by checking the proof below.

It suffices to estimate the L^2 norm of $(1-\Delta)^{1/4}f$ in cylinders as in (3.2). More precisely, since by a Sobolev embedding

$$\|h(s)\|_{H^{1/2}(\tilde{Q}_x)} \le \|h(s)\|_{W^{1,4/3}(\tilde{Q}_x)},$$

we have to show that

$$\sum_{k\in\mathbb{Z}^2} \||g| + |\nabla g|\|_{L^{4/3}(Q_k)} \le c\Big(\|u(0)\|_{H^{1/2}} + t_0^{1/2}\|u\|_{X_{t_0}}^2\Big)^2\|u\|_{X_{t_0}}, \tag{3.3}$$

where g is one of the following terms:

$$u^3,\ u^2\partial_\nu u,\ u\partial_\nu R_\nu|u|^2. \tag{3.4}$$

The treatment of $u\partial_\nu R_\nu|u|^2$ is typical and contains all the difficulties. For clarity we first consider the term $\nabla(u^2Du)$. By Hölder's inequality,

$$\|\nabla(u^2(s)Du(s))\|_{L^{4/3}(\tilde{Q}_x)} \lesssim \|u(s)\|_{L^8(\tilde{Q}_x)}^2\|D^2u(s)\|_{L^2(\tilde{Q}_x)} + \|u(s)\|_{L^8(\tilde{Q}_x)}\|\nabla u\|_{L^{16/5}(\tilde{Q}_x)}^2,$$

and by interpolation,

$$\|\nabla u\|_{L^{16/5}(\tilde{Q}_x)}^2 \lesssim \|u\|_{L^8(\tilde{Q}_x)}(\|u\|_{L^2(\tilde{Q}_x)} + \|D^2u\|_{L^2(\tilde{Q}_x)});$$

hence

$$\int_0^{t_0} \|\nabla(u^2(s)Du(s))\|^2_{L^{4/3}(\tilde{Q}_x)}ds \lesssim \int_0^{t_0} \|u(s)\|^4_{L^8(\tilde{Q}_x)}(\|u\|^2_{L^2(\tilde{Q}_x)} + \|D^2u(s)\|^2_{L^2(\tilde{Q}_x)})ds$$
$$\leq c_1 \int_0^{t_0} \|u(s)\|^4_{H^{1/2}(\tilde{Q}_x)}\|u(s)\|^2_{H^{5/2}(\tilde{Q}_x)}ds$$
$$\leq c_2 \sup_{0\leq t\leq t_0} \|(1-\Delta)^{1/4}u(t)\|^2_{L^2(\tilde{Q}_x)}\|(1-\Delta)^{5/4}u\|_{L^2(Q_{x,t_0})},$$

and thus, by the definition of the spaces X_{t_0},

$$\sum_{k\in Z^2}\left\{\int_0^t \|u^2(s)Du(s)\|^2_{H^{1/2}(\tilde{Q}_k)}ds\right\} \leq c\Big(\|(1-\Delta)^{1/4}u(0)\|_{L^2} + t_0^{1/2}\|u\|_{X_{t_0}}\Big)^2\|u\|_{X_{t_0}}.$$

We now turn to the most difficult term

$$\sum_k \|\nabla(u\partial_\nu R_\nu|u|^2)\|_{L^2([0,t_0],L^{4/3}(Q_k))}.$$

There are two terms to control:

$$\sum_k \|(\nabla u)\partial_\nu R_\nu|u|^2\|_{L^2([0,t_0],L^{4/3}(Q_k))} \tag{3.5}$$

and

$$\sum_k \|u\nabla\partial_\nu R_\nu|u|^2\|_{L^2([0,t_0],L^{4/3}(Q_k))}. \tag{3.6}$$

By Hölder's inequality, for fixed t,

$$\|(\nabla u)\partial_\nu R_\nu|u|^2\|_{L^{4/3}(Q_k)} \lesssim \|\nabla u\|_{L^{16/5}(Q_k)}\|\partial_\nu R_\nu|u|^2\|_{L^{16/7}}.$$

The kernel $K(x,y)$ of $\partial_\nu R_\nu$ decays fast for $|x-y|\to\infty$,

$$|K(x,y)| \leq c|x-y|^{-3};$$

hence we estimate

$$\|\partial_\nu R_\nu|u|^2\|_{L^{\frac{16}{7}}(Q_k)} \lesssim \sum_{\bar{k}}(1+|k-\bar{k}|)^{-3}\sup_{\bar{k}}\|u\|^2_{L^2(Q_k)} + \sum_{|k-\bar{k}|\leq 2}\|\partial_\nu|u|^2\|_{L^{\frac{16}{7}}(Q_{\bar{k}})}.$$

Now we proceed as above and bound $\sum_k \|\nabla(u\partial_v R_v|u|^2)\|_{L^2([0,t_0],L^{4/3}(Q_k))}$ following the arguments above. □

We complete the proof of Theorem 1.3 using Lemmas 2.7 and 3.1. First we obtain

$$\|w\|_{X_{t_0}} \leq c(n)\|w_0\|_{H^{3/2}} \leq c(n)c_0.$$

Let $J: X_{t_0}\to X_{t_0}, J:\tilde{v}\to v$, be defined as a solution to

$$v_t + P(D)v = f(w+\tilde{v}), \qquad v(0) = u_0 - w_0.$$

Then by Lemmas 2.7 and 3.1,

(3.7)
$$\|v\|_{X_{t_0}} \lesssim c_0 + \Big(\|w_0\|_{H^{1/2}} + \|u_0 - w_0\|_{H^{1/2}} + t_0^{1/2}(\|w_0\|_{H^{3/2}} + \|\tilde v\|_{X_{t_0}})\Big)^2(\|w_0\|_{H^{3/2}} + \|\tilde v\|_{X_{t_0}})$$
$$\lesssim c_0 + t_0(c_0 + \|\tilde v\|_{X_{t_0}})^3.$$

We shall see that this map is a contraction—at least if we decrease t_0 if necessary. The lower bound on the life span follows from (3.7).

Let

$$v_t^j + P(D)v^j = f(w + \tilde v^j), \qquad v^j(0) = u_0 - w_0,$$

for $j = 1, 2$. We expand the trilinear term

$$f(w + \tilde v^2) - f(w + \tilde v^1) = 3F[w, w, \tilde v^2 - \tilde v^1] + 3F[w, \tilde v^2 + \tilde v^1, \tilde v^2 - \tilde v^1]$$
$$+ F[\tilde v^2 + \tilde v^1, \tilde v^2 + \tilde v^1, \tilde v^2 - \tilde v^1].$$

It is straightforward to estimate

$$\|F[w(t), w(t), \tilde v^2 - \tilde v^1]\|_{L^1(H^{3/2})} \lesssim t_0^{1/2}\|w_0\|_{H^4}^2\|\tilde v^2 - \tilde v^1\|_{X_{t_0}},$$

and hence,

$$\|v^2 - v^1\|_{X_{t_0}} \le \gamma\|\tilde v^2 - \tilde v^1\|_{X_{t_0}},$$

where γ can be chosen as

$$\gamma \sim t_0^{1/2}\|w_0\|_{H^4}^2 + (\|w_0\|_{H^{1/2}} + t_0^{1/2}\|w\|_{X_{t_0}} + \mu)\mu \lesssim t_0^{1/2}c_0^5\varepsilon^{-3} + (c_0 + \mu)\mu$$

with

$$\mu = \varepsilon + t_0^{1/2}(\|\tilde v^2\|_{X_{t_0}} + \|\tilde v^1\|_{X_{t_0}}).$$

Suppose that $\|\tilde v^j\|_{X_{t_0}} \le R$ with $R > c_0$. Then

$$\|v^2 - v^1\|_{X_{t_0}} \le \frac{1}{2}\|\tilde v^2 - \tilde v^1\|_{X_{t_0}}$$

provided $\varepsilon R \ll 1$, $t_0^{1/2}R^2 \ll 1$, $t_0^{1/2}(c_0 + 1) \ll 1$ and $t_0 \ll c_0^{-10}\varepsilon^6$. Given R we can satisfy all these inequalities. Let $R_0 = \|J(0)\|_{X_{t_0}}$, $R \ge 2R_0$, and ε, t_0 be as above. Then

$$\|J(\tilde v)\|_{X_{t_0}} \le R_0 + \frac{1}{2}R \le R$$

if $\|\tilde v\|_{X_{t_0}} \le R$ and $\|\tilde v(0)\|_{H^{1/2}(\mathbb{R}^2)} \le \varepsilon$. Now J maps this ball into itself and is a contraction. The same argument gives uniqueness of solutions in that class. Finally, we may reinterpret the considerations above as an application of the implicit function theorem with analytic nonlinearities, which implies analytic dependence on the initial data. □

By similar arguments one could establish the local well-posedness of (1.3) for $n = 3$ in $\mathbb{R}^3$ in the cubic case for initial data in $H^2(\mathbb{R}^2)$, provided ε_1 and ε_2 are nonzero.

REFERENCES

[1] M. Ben-Artzi, H. Koch, and J.-C. Saut, *Dispersion estimates for fourth order Schrödinger equations*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 87–92.

[2] M. Ben-Artzi, H. Koch, and J.-C. Saut, *Dispersion estimates for third order equations in two dimensions*, Comm. Partial Differential Equations, 28 (2003), pp. 1943–1974.

[3] H. Chihara, *Third Order Semilinear Dispersive Equations Related to Deep Water Waves*, AriXiv e-print, http:www.arxiv.org/abs/math.AP/0404005 (2004).

[4] A. Davey and K. Stewartson, *On three-dimensional packets of surface waves*, Proc. Roy. Soc. London Ser. A, 338 (1974), pp. 101–110.

[5] A. de Bouard, *Analytic solutions to nonelliptic nonlinear Schrödinger equations*, J. Differential Equations, 104 (1993), pp. 196–213.

[6] F. Dias and C. Kharif, *Nonlinear gravity and capillary-gravity waves*, Annu. Rev. Fluid Mech., 31 (1999), pp. 301–346.

[7] V. D. Djordjevic and L. G. Redekopp, *On two-dimensional packets of capillary-gravity waves*, J. Fluid Mech., 79 (1977), pp. 703–714.

[8] K. B. Dysthe, *Note on a modification to the nonlinear Schrödinger equation for applications to deep water waves*, Proc. Roy. Soc. London Ser. A, 369 (1979), pp. 105–114.

[9] J.-M. Ghidaglia and J.-C. Saut, *Nonelliptic Schroedinger equations*, J. Nonlinear Sci., 3 (1993), pp. 169–195.

[10] H. Hasimoto and M. Ono, *Nonlinear modulation of gravity waves*, J. Phys. Soc., 33 (1972), pp. 805–811.

[11] S. J. Hogan, *The fourth-order evolution equation for deep-water gravity-capillary waves*, Proc. Roy. Soc. London Ser. A, 402 (1985), pp. 359–372.

[12] T. Hoshiro, *Decay and regularity for dispersive equations with constant coefficients*, J. Anal. Math., 91 (2003), pp. 211–230.

[13] C. E. Kenig, G. Ponce, and L. Vega, *Oscillatory integrals and regularity of dispersive equations*, Indiana Univ. Math. J., 40 (1991), pp. 33–69.

[14] C. E. Kenig, G. Ponce, and L. Vega, *Small solutions to nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré. Anal. Non Linéaire, 10 (1993), pp. 255–288.

[15] C. E. Kenig, G. Ponce, and L. Vega, *On the generalized Benjamin-Ono equation*, Trans. Amer. Math. Soc., 342 (1994), pp. 155–172.

[16] C. E. Kenig, G. Ponce, and L. Vega, *On the Zakharov and Zakharov-Schulman systems*, J. Funct. Anal., 127 (1995), pp. 204–234.

[17] C. Laurey, *The Cauchy problem for a third order nonlinear Schrödinger equation*, Nonlinear Anal., 29 (1997), pp. 121–158.

[18] F. Linares and G. Ponce, *On the Davey-Stewartson systems*, Ann. Inst. H. Poincaré. Anal. Non Linéaire, 10 (1993), pp. 523–548.

[19] V. E. Zakharov, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, J. Appl. Mech. Tech. Phys., 9 (1968), pp. 86–94.

[20] A. A. Zozulya, S. A. Diddams, A. G. Van Engen, and T. S. Clement, *Propagation dynamics of intense femtosecond pulses: Multiple splittings, coalescense and continuum generation*, Phys. Rev. Lett., 82 (1999), pp. 1430–1439.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1542–1564

RESONANCE AND INTERIOR LAYERS IN AN INHOMOGENEOUS PHASE TRANSITION MODEL*

MANUEL DEL PINO†, MICHAŁ KOWALCZYK‡, AND JUNCHENG WEI§

Abstract. We consider the problem $\varepsilon^2\Delta u + (u - a(x))(1 - u^2) = 0$ in Ω, $\frac{\partial u}{\partial \nu} = 0$ on $\partial\Omega$, where Ω is a smooth and bounded domain in $\mathbb{R}^2$, $-1 < a(x) < 1$. Assume that $\Gamma = \{x \in \Omega, a(x) = 0\}$ is a closed, smooth curve contained in Ω in such a way that $\Omega = \Omega_+ \cup \Gamma \cup \Omega_-$ and $\frac{\partial a}{\partial n} > 0$ on Γ, where n is the outer normal to Ω_+. Fife and Greenlee [*Russian Math. Surveys*, 29 (1974), pp. 103–131] proved the existence of an interior transition layer solution u_ε which approaches -1 in Ω_- and $+1$ in Ω_+, for *all* ε sufficiently small. A question open for many years has been whether an interior transition layer solution approaching 1 in Ω_- and -1 in Ω_+ exists. In this paper, we answer this question affirmatively when $n = 2$, provided that ε is small and away from certain critical numbers. A main difficulty is a resonance phenomenon induced by a large number of small critical eigenvalues of the linearized operator.

Key words. interior transition layer, Fife–Greenlee problem, infinite-dimensional reduction, spectral gap

AMS subject classifications. 35J25, 35J20, 35B33, 35B40

DOI. 10.1137/060649574

1. Introduction and statement of main result.

Let Ω be a bounded, smooth domain in $\mathbb{R}^2$. In the gradient theory of phase transitions it is common to seek critical points in $H^1(\Omega)$ of energy of the form

$$J_\varepsilon(u) = \frac{\varepsilon}{2}\int_\Omega |\nabla u|^2 + \varepsilon^{-1}\int_\Omega W(x,u),$$

where $W(x,\cdot)$ is a double-well potential with exactly two strict local minimizers at $u = +1$ and $u = -1$, which also correspond to trivial local minimizers of J_ε in $H^1(\Omega)$. For simplicity of exposition we shall restrict ourselves to a potential of the form

$$W(x,u) = \int_{-1}^{u} (s^2 - 1)(s - a(x))ds, \tag{1.1}$$

for a smooth function $a(x)$ with

$$-1 < a(x) < 1 \text{ for all } x \in \Omega.$$

Critical points of J_ε correspond to solutions of the problem

$$\begin{cases} \varepsilon^2\Delta u + (u - a(x))(1 - u^2) = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.2}$$

*Received by the editors January 10, 2006; accepted for publication (in revised form) September 13, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sima/38-5/64957.html

†Departamento de Ingeniería Matemática and CMM, Universidad de Chile, Casilla 170 Correo 3, Santiago, Chile (delpino@dim.uchile.cl). This author is partially supported by grants Fondecyt 108340 and FONDAP, Chile.

‡Department of Mathematical Sciences, Kent State University, Kent, OH 44242, and Departamento de Ingeniería Matemática, Universidad de Chile, Casilla 170 Correo 3, Santiago, Chile (kowalcyk@math.kent.edu). This author has been supported by the Fondecyt grant 1050311 and Núcleo Milenio grant P04-069-F.

§Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). This author is partially supported by an Earmarked Grant from RGC of Hong Kong.

where $\varepsilon > 0$ is a small parameter and ν denotes unit outer normal to $\partial\Omega$. The function $u(x)$ represents a continuous realization of the phase present in a material confined to the region Ω at the point x which, except for a narrow region, is expected to take values close to $+1$ or -1. Of interest are, of course, nontrivial steady state configurations in which the two phases coexist.

The case $a \equiv 0$ corresponds to the standard Allen–Cahn equation [6]

$$\begin{cases} \varepsilon^2 \Delta u + u(1-u^2) = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.3}$$

for which extensive literature on transition layer solutions is available; see, for instance, [4, 17, 18, 24] and the references therein. We observe that in this case, $+1$ and -1 are both global minimizers of the potential (1.1). We are interested in an inhomogeneous situation in which $+1$ is the absolute minimizer of $W(x,\cdot)$ in one region of the domain, while -1 is such a minimizer in its complement. More precisely, we shall assume that the set

$$\Gamma = \{x \in \Omega \ / \ a(x) = 0\}$$

is a smooth, simple, closed curve in Ω which separates the domain into two disjoint components,

$$\Omega = \Omega_- \cup \Omega_+ \cup \Gamma \tag{1.4}$$

such that

$$a(x) < 0 \ \text{ in } \Omega_+, \quad a(x) > 0 \ \text{ in } \Omega_-, \quad \frac{\partial a}{\partial n} > 0 \ \text{ on } \Gamma\,. \tag{1.5}$$

Observe in particular that for the potential (1.1) we have

$$W(x,-1) < W(x,+1) \quad \text{in } \Omega_-, \quad W(x,+1) < W(x,-1) \quad \text{in } \Omega_+\,.$$

Thus, if one considers a global minimizer u_ε for J_ε, which exists by standard arguments, then u_ε should minimize $W(x,u)$; namely, u_ε should intuitively have the following asymptotic behavior as $\varepsilon \to 0$:

$$u_\varepsilon \to -1 \text{ in } \Omega_-, \quad u_\varepsilon \to +1 \text{ in } \Omega_+. \tag{1.6}$$

A solution u_ε to problem (1.2) with these characteristics was constructed, and precisely described, by Fife and Greenlee [15] in 1974 via matching asymptotic and bifurcation arguments.

Supersubsolutions were later used by Angenent, Mallet-Paret, and Peletier in the one-dimensional case [7] for construction and classification of stable solutions. Radial solutions were found variationally by Alikakos and Simpson [5]. The construction of the Fife–Greenlee solution allowing Γ to be any closed subset of Ω in any dimension was given by the first author in [10]. Further constructions have been found recently by Dancer and Yan [9] and Do Nascimento [13]. In particular, it was found in [9] that this solution is precisely a minimizer of J_ε. Related results can be found in [1, 3].

On the other hand, a solution exhibiting a transition layer in the *opposite direction*, namely,

$$u_\varepsilon \to +1 \text{ in } \Omega_-, \qquad u_\varepsilon \to -1 \text{ on } \Omega_+, \tag{1.7}$$

has been believed to exist for many years. Hale and Sakamoto [19] established the existence of this solution in the one-dimensional case, while this was done in the radial case in a ball in [11]; see also [8]. The opposite direction layer (1.7) in this scalar problem is meaningful in finding transition layer solutions in pattern-formation–reaction-diffusion systems such as the Gierer–Meinhardt system with saturation; see [11, 14, 25, 28, 27] and the references therein. While the singular perturbation methods used in these one-dimensional or radial equations and systems do not see a substantial difference between the stable and unstable layers, except for the sign of the principal $O(\varepsilon)$ eigenvalue of the linearization, one faces a dramatically different situation in higher-dimensional, nonsymmetric situations. This is clearly seen when linearizing around a spherically symmetric solution like (1.7), as bifurcations of nonradial solutions along certain infinite discrete sets of values for $\varepsilon \to 0$ take place, as established in [27]. In particular, the radial solution has a large ε-dependent Morse index. This poses an important difficulty for a general construction. A phenomenon of this type was previously observed in the one-dimensional case by Alikakos, Bates, and Fusco [2] in a construction of solutions with any prescribed Morse index.

In this paper we are able to prove that the opposite-layer solution (1.7) exists as long as ε remains properly away from a set of critical values. More precisely, there is an explicit number $\lambda_* > 0$ such that given $c > 0$, if ε is sufficiently small and satisfies the gap condition

$$|k^2\varepsilon - \lambda_*| \geq c\sqrt{\varepsilon} \quad \text{for all } k \in \mathbb{N}, \tag{1.8}$$

then a solution u_ε with the required concentration property indeed exists. In other words, this will be the case whenever ε is small and away from the critical numbers $\frac{\lambda_*}{k^2}$, in the sense that for fixed and arbitrarily small $c < \lambda_*$,

$$\varepsilon \notin \left[\frac{\lambda_*}{k^2} - \frac{c}{k^3}\,,\, \frac{\lambda_*}{k^2} + \frac{c}{k^3} \right] \quad \text{for all } k \in \mathbb{N}.$$

Here λ_* is defined by

$$\lambda_* = \frac{1}{3\pi^2 \int_{\mathbb{R}} H_x^2 dx} \left(\int_\Gamma \sqrt{\frac{\partial a}{\partial \nu}} \right)^2, \tag{1.9}$$

where $H(y)$ is the unique heteroclinic solution of

$$H'' + H - H^3 = 0,\ H(0) = 0,\ H(\pm\infty) = \pm 1. \tag{1.10}$$

We can now state our main result.

THEOREM 1. *Given $c > 0$, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$ satisfying the gap condition* (1.8), *problem* (1.2) *has a solution u_ϵ satisfying*

$$u_\varepsilon(x) \to +1 \quad in \quad \Omega_-, \quad u_\varepsilon(x) \to -1 \quad in \quad \Omega_+$$

as $\varepsilon \to 0$.

Much more accurate information on the solution will be provided by its construction; in particular, its shape near Γ is governed by the heteroclinic solution H, in the sense that

$$u_\varepsilon(x) \sim H\left(\frac{t - \varepsilon f(\theta)}{\varepsilon} \right),$$

where f is a bounded function of θ, a choice of arclength coordinate of Γ, and t is the (signed) normal coordinate along the outer normal to Ω_+ on Γ.

The main difficulty in the construction of the interior layer solution in the opposite direction is the appearance of a large number of small *critical eigenvalues*, or *resonance*. This kind of phenomenon has been dealt with in various problems, for example, in the study of periodic orbits for strongly attractive potentials [21, 29] and in boundary concentrations for singularly perturbed Neumann problems [22, 23]. It also arises in our previous work [12] on the construction of a concentrating solution on weighted geodesics for nonlinear Schrodinger equations. The scheme employed here follows the general lines set in [12].

More precisely, the solution to the full problem is roughly decomposed into the form

$$u_\varepsilon(x) = H\left(s - f(\varepsilon z)\right) + \phi_1\left(s - f(\varepsilon z)\right) + \tilde{\phi}(s, z), \tag{1.11}$$

where $x = (t, \theta) = (\varepsilon s, \varepsilon z)$, $t = \varepsilon s$ is the signed distance to Γ, $\theta = \varepsilon z$ is the arclength coordinate of Γ whose length is l, f is an l-periodic function left as a parameter, and ϕ_1 is the correction term to be defined, while $\tilde{\phi}(s, z)$ is $L^2(ds)$-orthogonal for each z to $H_s(s - f(\varepsilon z))$. Solving first in $\tilde{\phi}$ a natural projected problem, where the linear operator is uniformly invertible, the resolution of the full problem becomes reduced to a nonlinear, nonlocal second order system of differential equations in f which turns out to be directly solvable thanks to the assumptions made. This approach is familiar when the parameter f lies in a finite-dimensional space (as in the papers [5, 9, 13, 19]), corresponding this time to adjusting infinitely many parameters. To stress the difference in the radial case, we note that the parameter f is just a single number. The analysis we make takes special advantage through Fourier analysis of the fact that the objects to be adjusted are one-variable functions, while we still believe that the current approach may be modified to the higher-dimensional case. We also believe that the gap condition may be improved to size ε^q, any $q > \frac{1}{2}$.

Additionally we point out the following:

1. The results of Theorem 1 remain true when Ω is an unbounded domain, for instance, $\Omega = \mathbb{R}^2$. Indeed, our proofs, and in particular the matching argument below, can easily be adapted to handle this case.

2. The method and results presented in Theorem 1 can be generalized to more general bistable equations of the form

$$\varepsilon^2 \Delta u - h(x, u, \varepsilon) = 0 \text{ in } \Omega, \quad \varepsilon \frac{\partial u}{\partial \nu} - \sigma(x, \epsilon) u = f(x, \epsilon) \text{ on } \partial\Omega,$$

as treated originally by Fife and Greenlee [15].

3. Our general approach seems also to work when $N = 3$. It will be an interesting problem to consider $N \geq 4$. Note that there is no restriction of dimension in the construction of Fife–Greenlee solutions (1.6); see [10].

The organization of this paper is as follows. In section 2, we set up the local coordinates near Γ and transform (1.2) into a new equation in the stretched variable (s, z). We then introduce the first correction term ϕ_1 and estimate the errors. In section 3, we use a gluing procedure to reduce the nonlinear problem to one on the infinite cylinder and another one away from the interface. Then we solve the inner problem modulo the projections in section 4 and the full problem modulo projections in section 5. In section 6 and section 7, we derive a nonlinear ODE for f, which will be solved in section 8 using the gap condition.

2. The setup near the curve. Let $\Gamma = \{x \in \Omega, a(x) = 0\}$ be a simple, closed, smooth curve in $\Omega \subset \mathbb{R}^2$ and let $\ell = |\Gamma|$ denote its total length. We consider the natural parameterization $\gamma(\theta)$ of Γ with positive orientation, where θ denotes the arclength parameter measured from a fixed point of Γ. Let $\nu(\theta)$ denote outer unit normal to Γ. Points y which are δ_0-close Γ for sufficiently small δ_0 can be represented in the form

$$y = \gamma(\theta) + t\,\nu(\theta), \quad |t| < \delta_0,\ \theta \in [0, \ell)\,, \tag{2.1}$$

where map $y \mapsto (t, \theta)$ is a local diffeomorphism. By slight abuse of notation we denote $a(t, \theta)$ to actually mean $a(y)$ for y in (2.1). Let $k(\theta)$ denote the curvature of Γ.

Stretching variables, absorbing ε from Laplace's operator, and replacing $u(y)$ with $u(\varepsilon y)$, (1.2) becomes

$$\Delta u + (u - a(\varepsilon y))(1 - u^2) = 0 \text{ in } \Omega_\varepsilon, \qquad \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega_\varepsilon, \tag{2.2}$$

where $\Omega_\varepsilon = \frac{\Omega}{\varepsilon}$.

Let $(s, z) = \varepsilon^{-1}(t, \theta)$ be the natural stretched coordinates associated with the curve $\Gamma_\varepsilon = \varepsilon^{-1}\Gamma$, now defined for

$$z \in [0, \varepsilon^{-1}\ell), \quad s \in (-\varepsilon^{-1}\delta_0, \varepsilon^{-1}\delta_0). \tag{2.3}$$

Equation (2.2) for u expressed in these coordinates becomes

$$u_{zz} + u_{ss} + B_1(u) + B_2(u) + u - u^3 = 0, \tag{2.4}$$

in the region (2.3), where

$$B_1(u) = -u_{zz}\left[1 - \frac{1}{(1 + \varepsilon k(\varepsilon z)s)^2}\right] + \frac{\varepsilon k(\varepsilon z)u_s}{1 + \varepsilon k(\varepsilon z)s} - \frac{\varepsilon^2 s\, k'(\varepsilon z)u_z}{(1 + \varepsilon k(\varepsilon z)s)^3},$$

$$B_2(u) = -a(\varepsilon s, \varepsilon z)(1 - u^2).$$

For further reference, it is convenient to expand B_1 in the form

$$B_1(u) \,=\, (\varepsilon k(\varepsilon z) - \varepsilon^2 s k^2(\varepsilon z))u_s \,+\, B_0(u), \tag{2.5}$$

where

$$B_0(u) \,=\, \varepsilon^2 s a_1(\varepsilon s, \varepsilon z)u_z + \varepsilon s a_2(\varepsilon s, \varepsilon z)u_{zz} + \varepsilon^3 s^2 a_3(\varepsilon s, \varepsilon z)u_s, \tag{2.6}$$

for certain smooth functions $a_j(t, \theta), j = 1, 2, 3$. Observe that all terms in the operator B_1 have ε as a common factor.

We consider now a further change of variables in (2.4). Let $f(\theta)$ be a twice differentiable, ℓ-periodic function whose exact form is to be specified later (see (2.25)). We define $v(x, z)$ by the relation

$$u(s, z) = v(\,x\,,\,z), \quad x = s - f(\varepsilon z). \tag{2.7}$$

We want to express (2.4) in terms of these new coordinates. We compute

$$u_s = v_x, \quad u_{ss} = v_{xx}, \tag{2.8}$$

$$u_z = v_x(-f)_z + v_z, \tag{2.9}$$

$$u_{zz} = v_{xx}|f_z|^2 + 2v_{xz}(-f)_z + v_x\,(-f)_{zz} + v_{zz}. \tag{2.10}$$

In order to write down the equation it is also convenient to expand

$$a(\varepsilon s, \varepsilon z) = a(0, \varepsilon z) + a_t(0, \varepsilon z)\varepsilon s + \frac{1}{2}a_{tt}(0, \varepsilon z)\varepsilon^2 s^2 + a_4(\varepsilon s, \varepsilon z)\varepsilon^3 s^3 \tag{2.11}$$

for a smooth function $a_4(t, \theta)$. It turns out that u solves (2.4) if and only if v defined by (2.7) solves

$$S(v) \equiv v_{zz} + v_{xx} + B_3(v) + B_4(v) + v - v^3 = 0, \tag{2.12}$$

where $B_3(v)$ is a linear differential operator defined by

$$\begin{aligned} B_3(v) = &\left[\varepsilon k - \varepsilon^2 k^2\,(\,x + f\,)\right]\, v_x \\ &+ \left[\varepsilon^2\,|f'|^2\, v_{xx} - 2\varepsilon f' v_{xz} - \varepsilon^2 f'' v_x\right] \\ &+ B_5(v), \end{aligned}$$

with

$$B_5(v) = B_0(u) - a_4(\varepsilon s, \varepsilon z)\varepsilon^3 s^3(1 - v^2) \tag{2.13}$$

and

$$B_4(v) = -\left[\varepsilon a_t\,(x + f) + \frac{\varepsilon^2}{2}a_{tt}\,(x + f)^2\right]\,(1 - v^2). \tag{2.14}$$

$B_0(u)$ is the operator in (2.6), where the derivatives are expressed in terms of the formulas (2.8)–(2.10), a_4 is given by (2.11), and s is replaced with $x + f$.

Let $H(x)$ denote the unique positive solution of (1.10). Then, taking $H(x)$ as a first approximation, the error produced is of ε times a function with exponential decay. Let us be more precise. We need to identify both the terms of order ε and those of order ε^2:

$$\begin{aligned} S(H) = B_3(H) + B_4(H) = &\left[\varepsilon k - \varepsilon^2 k^2\,(\,x + f\,)\right]\, H_x \\ &+ \left[\,\varepsilon^2 |f'|^2\, H_{xx} - \varepsilon^2 f'' H_x\,\right] \\ &- \left[\varepsilon a_t\,(x + f) + \frac{\varepsilon^2}{2}a_{tt}\,(x + f)^2\right]\,(1 - H^2) + B_5(H), \end{aligned}$$

where

$$B_5(H) = B_0(H) - \varepsilon^3 s^3 a_4(\varepsilon s, \varepsilon z)(1 - H^2).$$

Gathering terms of order ε and ε^2 we get

$$\begin{aligned} S(H) = &-\varepsilon a_t x(1-H^2) \\ &+\varepsilon\left[kH_x - a_t f(1-H^2)\right] \\ &-\varepsilon^2\left[k^2(xH_x) - |f'|^2 H_{xx} + a_{tt} f x(1-H^2)\right] \\ &-\varepsilon^2\left[k^2 f H_x + f'' H_x + \frac{a_{tt}}{2}[x^2+f^2](1-H^2)\right] \\ &\qquad + B_6(H) \\ = &\,\varepsilon S_1 + \varepsilon S_2 + \varepsilon^2 S_3 + \varepsilon^2 S_4 + B_6(H)\,. \end{aligned}$$

Let us observe that, grouped this way, the quantities S_1, S_3 are odd functions of x while S_2, S_4 are even. In addition, $B_6(H)$ is a term of order ε^3 times an exponentially decaying function. We want now to construct a further approximation to a solution which eliminates the terms of order ε in the error. If ϕ represents such an approximation, then we see that

$$S(H+\phi) = S(H) + L_0(\phi) + B_7(\phi) + N_0(\phi),$$

where

$$L_0(\phi) = \phi_{zz} + \phi_{xx} + (1-3H^2)\phi, \tag{2.15}$$

$$B_7(\phi) = B_3(H+\phi) + B_4(H+\phi) - B_3(H) - B_4(H), \tag{2.16}$$

and

$$N_0(\phi) = -3H\phi^2 - \phi^3. \tag{2.17}$$

We write

$$\begin{aligned} S(H+\phi) = &\,[\varepsilon(S_1+S_2) + \phi_{xx} + (1-3H^2)\phi] \\ &+\varepsilon^2 S_3 + \varepsilon^2 S_4 + B_6(H) + \phi_{zz} + B_7(\phi) + N_0(\phi). \end{aligned} \tag{2.18}$$

We choose $\phi = \phi_1$ in order to eliminate the term between brackets in the above expression. Namely, for fixed z, we need a solution of

$$-\phi_{xx} + (3H^2-1)\phi = \varepsilon(S_1+S_2), \quad \phi(\pm\infty) = 0.$$

As it is well known, this problem is solvable provided that

$$\int_{-\infty}^{\infty} (S_1+S_2)H_x\,dx = 0. \tag{2.19}$$

Furthermore, the solution is unique under the constraint

$$\int_{-\infty}^{\infty} \phi H_x\,dx = 0. \tag{2.20}$$

We compute

$$\int_{-\infty}^{\infty} (S_1+S_2)H_x\,dx = \int_{-\infty}^{\infty} S_2 H_x\,dx = k\int_{-\infty}^{\infty} H_x^2 - a_t f \int_{-\infty}^{\infty} (1-H^2)H_x,$$

where

$$\int_{-\infty}^{\infty}(1-H^2)H_x=\frac{4}{3}.$$

Since $a_t(0,\theta)\neq 0$, we have that the first approximation of f should be

$$f_0(\theta)=c_0\frac{k(\theta)}{a_t(0,\theta)},\quad \text{where } c_0=\frac{3\int_{\mathbb{R}}H_x^2}{4}.$$

The solution has the form

$$\phi_1=\phi_{11}+\phi_{12}, \tag{2.21}$$

where

$$\phi_{11}=\varepsilon a_{11}(\varepsilon z)H_1(x),\quad \phi_{12}=\varepsilon f_0(\varepsilon z)a_{12}(\varepsilon z)H_2(x), \tag{2.22}$$

$$a_{11}=a_t(0,\theta),\quad a_{12}=k(\theta),$$

H_1 is the unique odd function satisfying

$$-H_{1,xx}-H_1+3H^2H_1=x(1-H^2), \tag{2.23}$$

and H_2 is the unique even solution satisfying

$$-H_{2,xx}-H_2+3H^2H_2=H_x-c_0(1-H^2),\quad \int_{\mathbb{R}}H_2H_x\,dx=0. \tag{2.24}$$

Let us now choose f:

$$f(\theta)=f_0(\theta)+\mathbf{f}(\theta). \tag{2.25}$$

In all what follows, we will assume the validity of the following constraints on the parameter $\mathbf{f}$:

$$\|\mathbf{f}\|\equiv\varepsilon\|\mathbf{f}^{''}\|_{L^2(0,\ell)}+\sqrt{\varepsilon}\|\mathbf{f}'\|_{L^2(0,\ell)}+\|\mathbf{f}\|_{L^\infty(0,\ell)}\leq\varepsilon, \tag{2.26}$$

so that

$$\|\mathbf{f}\|_{L^\infty(0,\ell)}\leq\varepsilon,\quad \|\mathbf{f}'\|_{L^2(0,\ell)}\leq\sqrt{\varepsilon},\quad \|\mathbf{f}''\|_{L^2(0,\ell)}\leq 1. \tag{2.27}$$

By interpolation, it also holds that

$$\|\mathbf{f}'\|_{L^\infty(0,\ell)}\leq\sqrt{\varepsilon}. \tag{2.28}$$

We now take our basic approximation to a solution to the problem near the curve Γ_ε to be

$$f(\theta)=f_0(\theta)+\mathbf{f}(\theta),\qquad \mathtt{H}=H+\phi_1. \tag{2.29}$$

Substituting $\phi=\phi_1$ in (2.18), we can compute the new error:

$$\begin{aligned}E_1&=S(\mathtt{H})=S(H+\phi_1)\\&=\varepsilon(S_1+S_2)+\phi_{1,xx}+(1-3H^2)\phi_1\\&\quad+\varepsilon^2S_3+\varepsilon^2S_4+B_6(H)+\phi_{1,zz}+B_7(\phi_1)+N_0(\phi_1)\\&=-\varepsilon a_t\mathbf{f}(1-H^2)+\varepsilon^2S_3+\varepsilon^2S_4+B_6(H)+\phi_{1,zz}+B_7(\phi_1)+N_0(\phi_1).\end{aligned} \tag{2.30}$$

Observe that since ϕ_1 and $\mathbf{f}$ are of size $O(\varepsilon)$, all terms above carry ε^2 in front. Observe also that all functions involved are expressed in (x,z) variables, and the natural domain for those variables is the infinite strip

$$\mathcal{S} = \{-\infty < x < \infty, \quad 0 < z < \ell/\varepsilon\}.$$

We now want to measure the size of the error in the $L^2(\mathcal{S})$ norm.

Note that

$$\| - \varepsilon a_t \mathbf{f}(1-H^2) + \varepsilon^2 S_3 + \varepsilon^2 S_4\|_{L^2(\mathcal{S})} \leq C\varepsilon^{\frac{3}{2}}. \tag{2.31}$$

A rather delicate term in the cubic remainder $B_6(H)$ is the one carrying $\mathbf{f}''$ since in reality we shall only assume a uniform bound on $\|\mathbf{f}''\|_{L^2(0,\ell)}$. For instance, one term arising from $B_6(H)$ can be written as

$$R = \varepsilon^3 (x+f) f''(\varepsilon z) a_2(\varepsilon(x+f), \varepsilon z) H_x(x), \quad f = f_0 + \mathbf{f},$$

with a_2 smooth (see (2.6)). Observe that

$$\int_{\mathcal{S}} |R|^2 \leq C\varepsilon^6 \int_0^{\frac{\ell}{\varepsilon}} |f''(\varepsilon z)|^2 dz = \varepsilon^5 \|f''\|^2_{L^2(0,\ell)}.$$

Hence

$$\|R\|_{L^2(\mathcal{S})} \leq C\varepsilon^{\frac{5}{2}} \|f''\|_{L^2(0,\ell)}.$$

Since ϕ_1 can be bounded by $C\varepsilon|x|^2 e^{-c|x|}$ for large $|x|$, we obtain that

$$\|B_7(\phi_1)\|_{L^2(\mathcal{S})} \leq C\varepsilon^{\frac{3}{2}} \ .$$

A similar bound holds for the term $N_0(\phi_1)$:

$$\|N_0(\phi_1)\|_{L^2(\mathcal{S})} \leq C\varepsilon^{\frac{3}{2}}. \tag{2.32}$$

In summary, we have

$$\|S(H+\phi_1)\|_{L^2(\mathcal{S})} \leq C\varepsilon^{\frac{3}{2}}. \tag{2.33}$$

We set up the full problem in the form $S(\mathtt{H}+\phi) = 0$, which takes the form near the curve,

$$S(\mathtt{H}+\phi) = L_0(\phi) + B_8(\phi) + E_1 + N_1(\phi) = 0, \tag{2.34}$$

where $E_1 = S(\mathtt{H})$ and

$$L_0(\phi) = \phi_{xx} + \phi_{zz} + (1 - 3\mathtt{H}^2)\phi \ , \tag{2.35}$$

$$B_8(\phi) = B_7(\phi+\phi_1) - B_7(\phi_1), \tag{2.36}$$

$$N_1(\phi) = N_0(\phi+\phi_1) - N_0(\phi_1). \tag{2.37}$$

We recall that the description made here is only local. However, we will be able to reduce the problem to one qualitatively similar to that of the above form in the infinite strip.

3. The matching procedure. We follow [12] to perform a procedure that we refer to as an infinite-dimensional Liapunov–Schmidt reduction (see the explanations at the end of this section). Since it is quite similar to that of [12], we shall only sketch the proofs.

First, we need to match solutions near and outside Γ. The idea is to solve the problem outside a tubular neighborhood of Γ and then to reduce the problem to an infinite strip.

Let $\mathtt{H}(y)$ denote the first approximation constructed near the curve in the coordinate $y = (y_1, y_2)$ in $\mathbb{R}^2$. Let $\delta < \delta_0/100$ be a fixed number. We consider a smooth cut-off function $\eta_\delta(t)$ such that $\eta_\delta(t) = 1$ if $|t| < \delta$ and $= 0$ if $|t| > 2\delta$. Denote as well $\eta^\varepsilon_\delta(s) = \eta_\delta(\varepsilon|s|)$, where s is the normal coordinate to Γ_ε. We define our first global approximation to be simply

$$\mathbf{H}(y) = \begin{cases} \eta^\varepsilon_{3\delta}(s)(\mathtt{H}+1) - 1 \text{ if } y \in \Omega_+, \\ \eta^\varepsilon_{3\delta}(s)(\mathtt{H}-1) + 1 \text{ if } y \in \mathbb{R}^2 \setminus \Omega_+. \end{cases}$$

Denote $S(u) = \Delta u + (u - a(\varepsilon s, \varepsilon z))(1 - u^2)$ for $u = \mathbf{H} + \tilde\phi$. Then $S(\mathbf{H} + \tilde\phi) = 0$ if and only if

$$\tilde L(\tilde\phi) = \tilde E + \tilde N(\tilde\phi), \tag{3.1}$$

where

$$\begin{aligned} \tilde E &= -S(\mathbf{H}), \\ \tilde L(\tilde\phi) &= \Delta\tilde\phi + [1 - 3\mathbf{H}^2 + 2a(\varepsilon y)\mathbf{H}]\tilde\phi, \end{aligned}$$

and

$$\tilde N(\tilde\phi) = -3\mathbf{H}(\tilde\phi)^2 - (\tilde\phi)^3 + a(\varepsilon y)(\tilde\phi)^2.$$

We further separate $\tilde\phi$ in the following form:

$$\tilde\phi = \eta^\varepsilon_{3\delta}\phi + \psi,$$

where, in coordinates (x, z), we assume that ϕ is defined in the whole strip $\mathcal{S}$. We want

$$\tilde L(\eta^\varepsilon_{3\delta}\phi) + \tilde L(\psi) = \tilde E + \tilde N(\eta^\varepsilon_{3\delta}\phi + \psi).$$

We achieve this if the pair (ψ, ϕ) satisfies the following nonlinear coupled system:

$$\eta^\varepsilon_{3\delta}\tilde L(\phi) = \eta^\varepsilon_\delta\tilde E + \eta^\varepsilon_\delta\tilde N(\eta^\varepsilon_{3\delta}\phi + \psi) - 3\eta^\varepsilon_\delta(1 - \mathbf{H}^2)\psi, \tag{3.2}$$

$$\begin{aligned} &\Delta\psi - 2(1 - a\mathbf{H})\psi + 3(1 - \eta^\varepsilon_\delta)(1 - \mathbf{H}^2)\,\psi = (1 - \eta^\varepsilon_\delta)\,\tilde E - 2\varepsilon\nabla\eta^\varepsilon_{3\delta}\,\nabla\phi \\ &\quad - 2\varepsilon^2(\Delta\eta^\varepsilon_{3\delta})\,\phi + (1 - \eta^\varepsilon_\delta)\,\tilde N(\,\eta^\varepsilon_{3\delta}\,\phi + \psi\,), \end{aligned} \tag{3.3}$$

where ϕ is defined globally on $\mathcal{S}$ and ψ is defined in Ω_ε and is required to satisfy the Neumann boundary condition.

Notice that the operator $\tilde L$ in the strip $\mathcal{S}$ may be taken as any compatible extension outside the $6\delta/\varepsilon$-neighborhood of the curve.

What we want to do next is to reduce the problem to one in the strip. To do this, we solve, given a small ϕ, problem (3.3) for ψ. This can be done in an elementary way: Let us observe first that since $|a(x)| < 1$, we have

$$\gamma_0^2 = \min_{x\in\bar{\Omega}} 2(1 - |a(x)|) > 0. \tag{3.4}$$

Since $1-\mathbf{H}^2$ is exponentially small for $|s| > \delta\varepsilon^{-1}$, where s is the normal coordinate to Γ_ε, then the problem

$$\Delta\psi - 2(1 - a(\varepsilon y)\mathbf{H})\psi + 3(1 - \eta_\delta^\varepsilon)(1 - \mathbf{H}^2)\psi = h, \text{ in } \Omega, \quad \frac{\partial\psi}{\partial\nu} = 0 \text{ on } \partial\Omega_\varepsilon, \tag{3.5}$$

has a unique bounded solution ψ whenever $\|h\|_\infty < +\infty$. Moreover,

$$\|\psi\|_\infty \le C\|h\|_\infty.$$

Assume now that ϕ satisfies the following decay condition:

$$|\nabla\phi(y)| + |\phi(y)| \le e^{-\frac{\gamma}{\varepsilon}} \quad \text{for } |s| > \frac{\delta}{\varepsilon}. \tag{3.6}$$

Since $\tilde{N}$ has a power-like behavior with power greater than one, a direct application of the contraction mapping principle yields that problem (3.3) has a unique (small) solution $\psi = \psi(\phi)$ with

$$\|\psi(\phi)\|_\infty \le Ce^{-\delta/\varepsilon} + C\varepsilon[\,\|\phi\|_{L^\infty(|s|>\delta\varepsilon^{-1})} + \|\nabla\phi\|_{L^\infty(|s|>\delta\varepsilon^{-1})}\,],$$

where with some abuse of notation by $\{|s| > \delta/\varepsilon\}$ we denote the complement of the δ/ε-neighborhood of Γ_ε. The nonlinear operator ψ satisfies a Lipschitz condition of the form

$$\|\psi(\phi_1) - \psi(\phi_2)\|_\infty \le C\varepsilon[\,\|\phi_1 - \phi_2\|_{L^\infty(|s|>\delta\varepsilon^{-1})} + \|\nabla(\phi_1 - \phi_2)\|_{L^\infty(|s|>\delta\varepsilon^{-1})}\,].$$

The full problem has been reduced to solving the (nonlocal) problem in the infinite strip $\mathcal{S}$,

$$L_2(\phi) = \eta_\delta^\varepsilon\tilde{E} + \eta_\delta^\varepsilon\tilde{N}(\eta_{3\delta}^\varepsilon\phi + \psi(\phi)) - 3\eta_\delta^\varepsilon(1 - \mathbf{H}^2)\psi(\phi), \tag{3.7}$$

for a $\phi \in H^2(\mathcal{S})$ satisfying condition (3.6). Here L_2 denotes a linear operator that coincides with $\tilde{L}$ on the region $|s| < 10\delta/\varepsilon$.

We shall define this operator next. The operator $\tilde{L}$ for $|s| > 20\delta/\varepsilon$ is given in coordinates (x, z) by

$$L_1(\phi) = \phi_{xx} + \phi_{zz} + (1 - 3\mathtt{H}^2)\phi.$$

We extend it for functions ϕ defined in the entire strip $\mathcal{S}$, in terms of (x, z), as follows:

$$L_2(\phi) = L_1(\phi) + 2\chi\,(\varepsilon|x|)a(\varepsilon s, \varepsilon z)\mathtt{H}\phi + \chi\,(\varepsilon|x|)B_1(\phi), \tag{3.8}$$

where $\chi\,(r)$ is a smooth cut-off function which equals 1 for $r < 10\delta$ and vanishes identically for $r > 20\delta$.

Rather than solving problem (3.1) directly, we shall do it in steps. We consider the following projected problem in $H^2(\mathcal{S})$: Given $f = f_0 + \mathbf{f}$, with $\mathbf{f}$ satisfying bounds (2.26), find functions $\phi \in H^2(\mathcal{S})$, c such that

$$L_2(\phi) = \eta_\delta^\varepsilon \tilde{E} + N_2(\phi) + c(\varepsilon z)\chi_\delta^\varepsilon H_x \quad \text{in } \mathcal{S}, \tag{3.9}$$

$$\phi(x,0) = \phi(x,\ell/\varepsilon), \quad \phi_z(x,0) = \phi_z(x,\ell/\varepsilon), \quad -\infty < x < +\infty, \tag{3.10}$$

$$\int_{-\infty}^{\infty} \phi(x,z)\, H_x(x)\, dx = 0, \quad 0 < z < \frac{\ell}{\varepsilon}. \tag{3.11}$$

Here $N_2(\phi) = \eta_\delta^\varepsilon \tilde{N}(\eta_{3\delta}^\varepsilon \phi + \psi(\phi)) - 3\eta_\delta^\varepsilon(1 - \mathtt{H}^2)\psi(\phi)$ and $\chi_\delta^\varepsilon(x) = \chi_1(\varepsilon|x|/\delta)$, where $\chi_1(t)$ is a cut-off function equal to 1 for $|t| < 1/2$ and equal to 0 for $|t| > 1$.

We will prove that this problem has a unique solution whose norm is controlled by the L^2 norm of $\eta_\delta^\varepsilon \tilde{E} = E_1 = S(\mathtt{H})$. The main step here is to show bounded invertibility of a suitable perturbation of the operator L_2. The proof of this fact is a combination of an a priori estimate (Lemma 4.1) with an application of the Fredholm alternative (Lemma 4.2). After this first step, our task is to adjust the parameter $\mathbf{f}$ in such a way that c is identically zero. As we will see, this turns out to be equivalent to solving a nonlocal, nonlinear, second order differential equation for $\mathbf{f}$ under periodic boundary conditions. This system is solvable in a region where the bound (2.26) holds.

We call the entire procedure described above as *infinite-dimensional Lyapunov–Schmidt reduction* because of its analogy to a method devised by Floer and Weinstein [16] in a finite-dimensional context for a related problem. In a finite-dimensional setting, the main step in this method, which corresponds to adjustment of a parameter to make $c = 0$, is also known as *quasi-invariant manifold reduction*. The whole scheme has been refined and widely used in singular perturbation elliptic problems.

We will carry out the outlined program in the next sections. To solve (3.9)–(3.11) we need to investigate invertibility of L_2 in the L^2-H^2 setting under periodic boundary conditions and orthogonality conditions.

4. Invertibility of L_2. Let L_2 be the operator defined in $H^2(\mathcal{S})$ by (3.8). In this section we study the linear problem

$$L_2(\phi) = h + c(\varepsilon z)\chi_\delta^\varepsilon H_x \quad \text{in } \mathcal{S}, \tag{4.1}$$

$$\phi(x,0) = \phi(x,\ell/\varepsilon), \quad \phi_z(x,0) = \phi_z(x,\ell/\varepsilon), \quad -\infty < x < +\infty, \tag{4.2}$$

$$\int_{-\infty}^{\infty} \phi(x,z)\, H_x(x)\, dx = 0, \quad 0 < z < \frac{\ell}{\varepsilon} \tag{4.3}$$

for a given $h \in L^2(\mathcal{S})$. Our main result in this section is the following.

PROPOSITION 4.1. *If δ in the definition of L_2 is chosen sufficiently small, then there exists a constant $C > 0$, independent of ε, such that for all small ε, problem* (4.1)–(4.3) *has a unique solution $\phi = T(h)$, which satisfies the estimate*

$$\|\phi\|_{H^2(\mathcal{S})} \le C\|h\|_{L^2(\mathcal{S})}.$$

For the proof of this result we need the validity of the corresponding assertion for a simpler operator which does not depend on δ. Let us consider the problem

$$\mathtt{L}(\phi) = \phi_{ss} + \phi_{zz} + (1 - 3H^2)\phi = h \quad \text{in } \mathcal{S}, \tag{4.4}$$

$$\phi(x,0) = \phi(x,\ell/\varepsilon), \quad \phi_z(x,0) = \phi_z(x,\ell/\varepsilon), \quad -\infty < x < +\infty, \tag{4.5}$$

$$\int_{-\infty}^{\infty} \phi(x,z)\, H_x(x)\, dx = 0, \quad 0 < z < \frac{\ell}{\varepsilon}. \tag{4.6}$$

LEMMA 4.1. *There exists a constant $C > 0$, independent of ε such that the solutions of* (4.4)–(4.6) *satisfy the a priori estimate*

$$\|\phi\|_{H^2(\mathcal{S})} \le C\|h\|_{L^2(\mathcal{S})}.$$

Proof. Let us consider Fourier series decompositions for h and ϕ of the form

$$\phi(x,z) = \sum_{k=0}^{\infty} \left[\phi_{1k}(x)\cos\left(\frac{2\pi k}{\ell}\varepsilon z\right) + \phi_{2k}(x)\sin\left(\frac{2\pi k}{\ell}\varepsilon z\right)\right],$$

$$h(x,z) = \sum_{k=0}^{\infty} \left[h_{1k}(x)\cos\left(\frac{2\pi k}{\ell}\varepsilon z\right) + h_{2k}(x)\sin\left(\frac{2\pi k}{\ell}\varepsilon z\right)\right].$$

Then we have the validity of the equations

$$-\frac{4\pi^2 k^2 \varepsilon^2}{l^2}\phi_{lk} + \mathtt{L}_0(\phi_{lk}) = h_{lk}, \quad x \in \mathbb{R}, \tag{4.7}$$

with orthogonality conditions

$$\int_{-\infty}^{\infty} \phi_{lk}\, H_x\, dx = 0. \tag{4.8}$$

We have denoted here

$$\mathtt{L}_0(\phi_{lk}) = \phi_{lk,xx} + (1-3H^2)\phi_{lk}.$$

Let us consider the bilinear form in $H^1(\mathbb{R})$ associated with the operator $\mathtt{L}_0$, namely,

$$B(\psi,\psi) = \int_{-\infty}^{\infty} [|\psi_x|^2 + (3H^2-1)|\psi|^2]\, dx\ .$$

Since (4.8) holds, we conclude that

$$C[\|\phi_{lk}\|^2_{L^2(\mathbb{R})} + \|\phi_{lk,x}\|^2_{L^2(\mathbb{R})}] \le B(\phi_{lk},\phi_{lk}) \tag{4.9}$$

for a constant $C > 0$ independent of l, k. Using this fact and (4.7) we conclude with the estimate

$$(1+k^4\varepsilon^4)\|\phi_{lk}\|^2_{L^2(\mathbb{R})} + \|\phi_{lk,x}\|^2_{L^2(\mathbb{R})} \le C\|h_{lk}\|^2_{L^2(\mathbb{R})}.$$

In particular, we see from (4.7) that ϕ_{lk} satisfies an equation of the form

$$-\phi_{lk,xx} + 2\phi_{lk} = \tilde{h}_{lk}, \quad x \in \mathbb{R},$$

where $\|\tilde{h}_{lk}\|_{L^2(\mathbb{R})} \le C\|h_{lk}\|_{L^2(\mathbb{R})}$. Hence it follows that additionally we have the estimate

$$\|\phi_{lk,xx}\|^2_{L^2(\mathbb{R})} \le C\|h_{lk}\|^2_{L^2(\mathbb{R})}. \tag{4.10}$$

Adding up estimates (4.9), (4.10) in k and l we conclude that

$$\|D^2\phi\|^2_{L^2(\mathcal{S})} + \|D\phi\|^2_{L^2(\mathcal{S})} + \|\phi\|^2_{L^2(\mathcal{S})} \le C\|h\|^2_{L^2(\mathcal{S})},$$

which ends the proof. ∎

We consider now the following problem: Given $h \in L^2(\mathcal{S})$, find functions $\phi \in H^2(\mathcal{S})$, $c \in L^2(0,\ell)$ such that

$$\mathtt{L}(\phi) \;=\; h \;+\; c(\varepsilon z)\chi_\delta^\varepsilon H_x \quad \text{in } \mathcal{S}, \tag{4.11}$$

$$\phi(x,0) = \phi(x,\ell/\varepsilon), \quad \phi_z(x,0) = \phi_z(x,\ell/\varepsilon), \quad -\infty < x < +\infty, \tag{4.12}$$

$$\int_{-\infty}^{\infty} \phi(x,z)\, H_x(x)\, dx = 0, \quad 0 < z < \frac{\ell}{\varepsilon}. \tag{4.13}$$

LEMMA 4.2. *Problem* (4.11)–(4.13) *possesses a unique solution. Moreover,*

$$\|\phi\|_{H^2(\mathcal{S})} \le C\|h\|_{L^2(\mathcal{S})}.$$

Proof. To establish existence, we assume that

$$h(x,z) = \sum_{k=0}^{\infty} \left[h_{1k}(x)\cos\left(\frac{2\pi k}{\ell}\varepsilon z\right) \;+\; h_{2k}(x)\sin\left(\frac{2\pi k}{\ell}\varepsilon z\right)\right]$$

and consider the problem of finding $\phi_{lk} \in H^1(\mathbb{R})$, and constants c_{lk}, such that

$$-\frac{4\pi^2 k^2 \varepsilon^2}{l^2}\phi_{lk} + L_0(\phi_{lk}) = h_{lk} + c_{lk}\chi_\delta^\varepsilon H_x, \quad x \in \mathbb{R},$$

and

$$\int_{-\infty}^{\infty} \phi_{lk}\, H_x\, dx = 0.$$

Fredholm's alternative yields that this problem is solvable with the choices

$$c_{lk} = -\frac{\int_{-\infty}^{\infty} h_{lk} H_x dx}{\int_{-\infty}^{\infty} H_x^2 \chi_\delta^\varepsilon dx}.$$

Observe in particular that

$$\sum_{k=0}^{\infty} |c_{lk}|^2 \le C\varepsilon\|h\|^2_{L^2(\mathcal{S})}. \tag{4.14}$$

Finally, define

$$\phi(x,z) = \sum_{k=0}^{\infty} \left[\phi_{1k}(x)\cos\left(\frac{2\pi k}{\ell}\varepsilon z\right) \;+\; \phi_{2k}(x)\sin\left(\frac{2\pi k}{\ell}\varepsilon z\right)\right],$$

and correspondingly

$$c(z) = \sum_{k=0}^{\infty} \left[c_{1k}\cos\left(\frac{2\pi k}{\ell} z\right) \;+\; c_{2k}\sin\left(\frac{2\pi k}{\ell} z\right)\right].$$

The estimate (4.14) gives that $c(\varepsilon z)\chi_\delta^\varepsilon H_x$ has the $L^2(\mathcal{S})$ norms controlled by that of h. The a priori estimates of the previous lemma tell us that the series for ϕ is convergent in $H^2(\mathcal{S})$ and defines a unique solution for the problem with the desired bounds. ∎

Proof of Proposition 4.1. Problem (4.1)–(4.3) can be reduced to a small perturbation of a problem of the form (4.11)–(4.13) in which Lemma 4.2 is applicable. In fact, we have

$$L_2(\phi) = \mathtt{L}(\phi) + \tilde{B}(\phi), \tag{4.15}$$

where

$$\tilde{B}(\phi) = 3(\mathtt{H}^2 - 3H^2)\phi + 2\chi(\varepsilon|x|)a(\varepsilon s, \varepsilon z)\phi + \chi(\varepsilon|x|)B_1(\phi).$$

In the operator $B_1(\phi)$, consider for instance the following term involving f'':

$$B_f(\phi) = \varepsilon^2 f''(\varepsilon z)\phi_x.$$

Then we have

$$\|B_f(\phi)\|^2_{L^2(\mathcal{S})} \le \varepsilon^3 \int_0^\ell |f''(\theta)|^2 d\theta \left(\sup_z \int_{-\infty}^{\infty} |\phi_x(x,z)|^2 \, dx \right).$$

Let $\varphi(z) = \int_{-\infty}^{\infty} |\phi_x(x,z)|^2 dx$. Then

$$\begin{aligned} \sup_z \varphi(z) &\le \varepsilon \int_{\mathcal{S}} |\phi_x|^2 + 2 \int_{\mathcal{S}} |\phi_x||\phi_{xz}| \\ &\le \frac{1}{2} \sup_z \varphi(z) + 4\varepsilon^{-1} \int_{\mathcal{S}} |\phi_{xz}|^2 + \varepsilon \int_{\mathcal{S}} |\phi_x|^2. \end{aligned}$$

Hence

$$\varphi(z) \le C\varepsilon^{-1} \|\phi\|^2_{H^2(\mathcal{S})}, \tag{4.16}$$

so that finally

$$\|B_f(\phi)\|_{L^2(\mathcal{S})} \le C\varepsilon \|f''\|_{L^2(0,\ell)}.$$

For other terms the analysis follows in a simpler way. In fact we get

$$\|\tilde{B}(\phi)\|_{L^2(\mathcal{S})} \le C\delta \|\phi\|_{H^2(\mathcal{S})}.$$

This last estimate is a rather straightforward consequence of the fact that $|\varepsilon s| < 20\delta$ wherever the operator $\chi(\varepsilon|x|)B_1$ is supported, and $|a(\varepsilon s, \varepsilon z)| \le C\delta$ in $\mathcal{S}$. Thus, by reducing δ if necessary, we apply the invertibility result of Lemma 4.2. This concludes the proof. □

5. Solving the nonlinear intermediate problem. In this section we will solve problem (3.9)–(3.11). For brevity we let $E_2 = \eta^\varepsilon_\delta \tilde{E}$.

Notice that

$$\|E_2\|_{L^2(\mathcal{S})} \le C\varepsilon^{\frac{3}{2}}.$$

For further reference, it is useful to point out the Lipschitz dependence of the term of error E_2 on the parameters $\mathbf{f}$ for the norms defined in (2.26). We have the validity of the estimate

$$\|E_2(\mathbf{f}_1) - E_2(\mathbf{f}_2)\|_{L^2(\mathcal{S})} \le C\varepsilon^{\frac{1}{2}} [\|\mathbf{f}_1 - \mathbf{f}_2\|]. \tag{5.1}$$

Let T be the operator defined by Proposition 4.1. Then the equation is equivalent to the fixed point problem

$$\phi = T(E_2 + N_2(\phi)) \equiv \mathcal{A}(\phi). \tag{5.2}$$

The operator T has a useful property: Assume h has support contained in $|x| \leq \frac{\delta}{\varepsilon}$. Then by elliptic estimates, $\phi = T(h)$ satisfies the estimate

$$|\phi(x,z)| + |\nabla\phi(x,z)| \leq \|\phi\|_\infty e^{-\frac{\gamma_0\delta}{\varepsilon}} \quad \text{for } |x| > \frac{\delta}{\varepsilon}. \tag{5.3}$$

Now we recall that the operator $\psi(\phi)$ satisfies, as seen directly from its definition,

$$\|\psi(\phi)\|_{L^\infty} \leq C\left[\,\|\,|\nabla\phi| + |\phi|\,\|_{L^\infty(|s|>\frac{\delta}{\varepsilon})} + e^{-\frac{\gamma_0\delta}{\varepsilon}}\right], \tag{5.4}$$

and also the Lipschitz condition

$$\|\psi(\phi_1) - \psi(\phi_2)\|_{L^\infty} \leq C\left[\,\|\,|\nabla(\phi_1-\phi_2)| + |\phi_1-\phi_2|\,\|_{L^\infty(|s|>\frac{\delta}{\varepsilon})}\right]; \tag{5.5}$$

here $s = x + f$. These facts will allow us to construct a region where the contraction mapping principle applies. As we have said,

$$\|E_2\|_{L^2(\mathcal{S})} \leq C_* \varepsilon^{\frac{3}{2}}$$

for certain constant $C_* > 0$. We consider the following closed, bounded subset of $H^2(\mathcal{S})$:

$$\mathcal{B} = \left\{ \phi \in H^2(\mathcal{S}) \;\middle|\; \begin{array}{l} \|\phi\|_{H^2(\mathcal{S})} \leq D\varepsilon^{\frac{3}{2}}, \\ \|\,|\phi| + |\nabla\phi|\,\|_{L^\infty(|s|>\frac{\delta}{\varepsilon})} \leq \|\phi\|_{H^2(\mathcal{S})} e^{-\frac{\gamma_0\delta}{2\varepsilon}} \end{array} \right\}.$$

We claim that if the constant D is fixed sufficiently large, then the map $\mathcal{A}$ defined in (5.2) is a contraction from B into itself.

Let us analyze the Lipschitz character of the nonlinear operator $N_2(\phi)$, involved in $\mathcal{A}$ for functions in $\mathcal{B}$. Arguing as in [12], we have the following Lipschitz estimates for $N_2(\phi)$:

$$\|N_2(\phi_1) - N_2(\phi_2)\|_{L^2(\mathcal{S})} \leq C\varepsilon^{\frac{3}{2}} \|\phi_1 - \phi_2\|_{H^2(\mathcal{S})}. \tag{5.6}$$

Now let $\phi \in \mathcal{B}$; then $\varphi = \mathcal{A}(\phi)$ satisfies

$$\|\varphi\|_{H^2(\mathcal{S})} \leq C_* \varepsilon^{\frac{3}{2}} \|T\|.$$

Choosing any number $D > C_*\|T\|$ we get that for small ε

$$\|\varphi\|_{H^2(\mathcal{S})} \leq D\varepsilon^{\frac{3}{2}}.$$

On the other hand we have

$$\|\varphi\|_{L^\infty(\mathcal{S})} \leq C\|\varphi\|_{H^2(\mathcal{S})}.$$

But φ satisfies an equation of the form $L_2(\varphi) = h$ with h compactly supported. Hence φ belongs to $\mathcal{B}$ thanks to the discussion above. $\mathcal{A}$ is clearly a contraction mapping thanks to (5.6). We conclude that $\mathcal{A}$ has a unique fixed point in $\mathcal{B}$.

We recall that the error E_2 and the operator T themselves carry the function $\mathbf{f}$ as a parameter. A tedious but straightforward analysis of all terms involved in the differential operator and in the error yield that this dependence is indeed Lipschitz with respect to the H^2 norm (for each fixed ε). Indeed, emphasizing now the dependence of L_2 on $\mathbf{f}$ we can write

$$L_{2,\mathbf{f}_1}(\phi(\mathbf{f}_1)) - L_{2,\mathbf{f}_2}(\phi(\mathbf{f}_2)) = L_{2,\mathbf{f}_1}[\phi(\mathbf{f}_1) - \phi(\mathbf{f}_2)] + [L_{2,\mathbf{f}_1} - L_{2,\mathbf{f}_2}](\phi(\mathbf{f}_1))$$

and use the theory just developed to estimate $\phi(\mathbf{f}_1) - \phi(\mathbf{f}_2)$. Taking advantage of the Lipschitz character of the error term $E_2(\mathbf{f})$, we can show the Lipschitz character of T, and we find

$$\|T_{\mathbf{f}_1} - T_{\mathbf{f}_2}\| \le C\varepsilon\|\mathbf{f}_1 - \mathbf{f}_2\|.$$

Hence

$$\|\phi(\mathbf{f}_1) - \phi(\mathbf{f}_2)\|_{H^2(\mathcal{S})} \le C\varepsilon\|\mathbf{f}_1 - \mathbf{f}_2\|. \tag{5.7}$$

We summarize the result we have obtained in the following.

PROPOSITION 5.1. *There is a number $D > 0$ such that for all sufficiently small ε and all $\mathbf{f}$ satisfying* (2.26)*, problem* (3.9)–(3.11) *has a unique solution $\phi = \phi(\mathbf{f})$ which satisfies*

$$\|\phi\|_{H^2(\mathcal{S})} \le D\varepsilon^{\frac{3}{2}},$$
$$\| \, |\phi| + |\nabla\phi| \, \|_{L^\infty(|s|>\frac{\delta}{\varepsilon})} \le \|\phi\|_{H^2(\mathcal{S})} e^{-\frac{\gamma_0\delta}{2\varepsilon}}.$$

Besides, ϕ depends Lipschitz continuously on $\mathbf{f}$ in the sense of estimate (5.7).

Next we carry out the second part of the program, which is to set up an equation for $\mathbf{f}$, which is equivalent to making c identically zero. The equation is obtained by simply integrating the equation (only in x) against H_x. It is therefore of crucial importance to carry out computations of the terms $\int_{\mathbb{R}} E_2 H_x \, dx$. We do that in the next section.

6. Estimates for projections of the error. In this section we carry out some estimates for the terms $\int_{\mathbb{R}} E_2 H_x \, dx$, where $E_2 = \eta_\delta^\varepsilon E_1$ and E_1 was defined as in (2.30). Observe that it suffices to evaluate $\int_{\mathbb{R}} E_1 H_x \, dx$ instead since the difference $E_2 - E_1$ is exponentially small in ε. Notice that the odd terms in x in E_1 do not contribute to the value of the integral since H_x is an even function.

We recall

$$S(H+\phi_1) = -\varepsilon a_t \mathbf{f}(1-H^2) + \varepsilon^2 S_3 + \varepsilon^2 S_4 + B_6(H) + \phi_{1,zz} + B_7(\phi_1) + N_0(\phi_1),$$

where S_3 is an odd function, S_4 is an even function, and $B_6(H)$ is of order ε^3. Thus, we see that

$$\begin{aligned}
&\int_{\mathbb{R}} S(H+\phi_1)H_x \\
&\quad = -\varepsilon a_t \mathbf{f} \int_{\mathbb{R}} (1-H^2)H_x \\
&\qquad -\varepsilon^2 \left\{ \mathbf{f}'' \int_{\mathbb{R}} H_x^2 + \mathbf{f}\left[k^2 \int_{\mathbb{R}} H_x^2 + a_{tt} f_0 \int_{\mathbb{R}} (1-H^2)H_x \right] \right. \\
&\qquad\qquad \left. + \frac{\mathbf{f}^2}{2} a_{tt} \int_{\mathbb{R}} (1-H^2)H_x dx \right\} \\
&\qquad + \int_{\mathbb{R}} N_0(\phi_1)H_x + \int_{\mathbb{R}} B_7(\phi_1)H_x + \varepsilon^2 \gamma_0(\varepsilon z) + \varepsilon^3 b_{1\varepsilon}\mathbf{f}'' + \varepsilon^3 b_{2\varepsilon}\,.
\end{aligned}$$

Here and below we denote by $b_{l\varepsilon}$, $l = 1, 2$, generic, uniformly bounded continuous functions of the form

$$b_{l\varepsilon} = b_{l\varepsilon}(z, \mathbf{f}(\varepsilon z), \mathbf{f}'(\varepsilon z)),$$

where additionally $b_{1\varepsilon}$ is uniformly Lipschitz in its last two arguments. Here and below, functions $\gamma_j(\theta), j = 0, 1, 2, \ldots$, are C^2 smooth in its argument $\theta \in (0, \ell)$.

Next we estimate $\int_{\mathbb{R}} N_0(\phi_1)H_x$. This term is to main order of the form $\int_{\mathbb{R}} H\phi_1^2 H_x$. Since ϕ_1 doesn't depend on $\mathbf{f}$, we have

$$\int_{\mathbb{R}} N_0(\phi_1)H_x = \varepsilon^2\gamma_1(\varepsilon z).$$

Now, let us consider $\int_{\mathbb{R}} B_7(\phi_1)H_x$. All terms in this expression, with the exception of the terms of size ε in B_7, carry in the L^2 norm as functions of $\theta = \varepsilon z$ powers 3 or higher. Thus, we find

$$\begin{aligned}\int_{\mathbb{R}} B_7(\phi_1)H_x &= \varepsilon \int_{\mathbb{R}} \left[k\phi_{1,x} - a_t(0, \varepsilon z)(x + f)\phi_1\right] H_x \ + O(\varepsilon^3) \\ &= -\varepsilon^2 \mathbf{f} a_t(0, \varepsilon z) \int_{\mathbb{R}} \phi_1 H_x \, dx + \varepsilon^2\gamma_2(\varepsilon z) + \varepsilon^3 b_{3\varepsilon} f^{''} + \varepsilon^3 b_{4\varepsilon},\end{aligned}$$

where $b_{3\varepsilon}$ is uniformly Lipschitz in $\mathbf{f}$ and $\mathbf{f}^{'}$.

In summary, we have established that

$$\begin{aligned}\int_{\mathbb{R}} S(H + \phi_1)H_x \, dx = -\left[\varepsilon^2(\mathbf{f}''(\varepsilon z) + \gamma_3(\varepsilon z)\mathbf{f}\ ) + \varepsilon \mathbf{f}\gamma_4(\varepsilon z)\right] \int_{\mathbb{R}} H_x^2 \\ + \varepsilon^2\gamma_5(\varepsilon z) + \varepsilon^3[b_{5\varepsilon}\mathbf{f}'' + b_{6\varepsilon}],\end{aligned} \tag{6.1}$$

where γ_4 is given by

$$\gamma_4(\theta) = \frac{a_t(0, \theta) \int_{\mathbb{R}} (1 - H^2)H_x}{\int_{\mathbb{R}} H_x^2}, \tag{6.2}$$

and $b_{5\varepsilon}$ is uniformly Lipschitz in $\mathbf{f}$ and $\mathbf{f}^{'}$.

7. Projections of terms involving ϕ. We will estimate next the terms that involve ϕ in (3.9)–(3.11) integrated against H_x. We call the sum of them $\varphi(\phi)$:

$$\begin{aligned}\varphi = -2 \int_{\mathbb{R}} \chi(\varepsilon|x|)a(\varepsilon s, \varepsilon z)\phi H_x \, dx \\ - \int_{\mathbb{R}} \chi(\varepsilon|x|)B_8(\phi)H_x \, dx + \int_{\mathbb{R}} N_2(\phi)H_x \, dx \\ + 3 \int_{\mathbb{R}} [\mathtt{H}^2 - H^2]\phi H_x \, dx = \sum_{i=1}^{4} \varphi_i.\end{aligned}$$

Let $\varphi_1(\varepsilon z) = -2 \int_{\mathbb{R}} a(\varepsilon s, \varepsilon z)\chi(\varepsilon|x|)\phi H_x$. Then it is easy to see that

$$\int_0^{\ell} |\varphi_1(\theta)|^2 d\theta \le C\varepsilon^3 \|\phi\|^2_{H^2(S)},$$

and hence

$$\|\varphi_1\|_{L^2(0,\ell)} \le C\varepsilon^3.$$

The Lipschitz continuity of φ_1 follows from the Lipschitz continuity of ϕ.

Next we let $\varphi_2(\varepsilon z) = \int_{\mathbb{R}} B_1(\phi)\chi(\varepsilon|x|)H_x$. We make the following observation: All terms in $B_1(\phi)$ carry ε and involve powers of x times derivatives of powers of $0, 1$ or two orders of ϕ. The conclusion is that since H_x has exponential decay, then

$$\int_0^\ell |\varphi_2(\theta)|^2 d\theta \le C\varepsilon^3 \|\phi\|^2_{H^2(S)}.$$

Hence

$$\|\varphi_2\|_{L^2(0,\ell)} \le C\varepsilon^3.$$

To prove the Lipschitz regularity of φ_2, we single out one less regular terms in $B_8(\phi)$. The one whose coefficient depends on $\mathbf{f}''$ explicitly has the form

$$\varphi_{2*} = \varepsilon^2 \mathbf{f}'' \int_{\mathbb{R}} \phi_x H_x = -\varepsilon^2 \mathbf{f}'' \int_{\mathbb{R}} \phi H_{xx}.$$

Since ϕ has Lipschitz dependence on $\mathbf{f}$ in the form (5.7), we see that this is transmitted from Sobolev's embedding into

$$\|\phi(\mathbf{f}_1) - \phi(\mathbf{f}_2)\|_{L^\infty(\mathcal{S})} \le C\varepsilon^{\frac{3}{2}} \|\mathbf{f}_1 - \mathbf{f}_2\|,$$

from where it follows

$$\|\varphi_{2*}(\mathbf{f}_1) - \varphi_{2*}(\mathbf{f}_2)\|_{L^2(0,\ell)} \le C\varepsilon^{1+\alpha} \|\mathbf{f}_1 - \mathbf{f}_2\|.$$

The remainder $\varphi_2 - \varphi_{2*}$ actually defines for fixed ε a compact operator for $\mathbf{f}$ in $L^2(0,\ell)$. This is a consequence of the fact that weak convergence in $H^2(\mathcal{S})$ implies local strong convergence in $H^1(\mathcal{S})$, and the same is the case for $H^2(0,\ell)$ and $C^1[0,\ell]$. If $\mathbf{f}_j$ are weakly convergent sequences in $H^2(0,\ell)$, then clearly the functions $\phi(\mathbf{f}_j)$ constitute a bounded sequence in $H^1(\mathcal{S})$. In the above remainder one can integrate by parts, if necessary, once in x. Averaging against H_x, which decays exponentially, localizes the situation, and the desired fact follows.

We observe also that $\varphi_3(\varepsilon z) = \int_{\mathbb{R}} N_2(\phi) H_x$ can be estimated similarly. Using the definition of $N_2(\phi)$ and the exponential decay of H_x we obtain

$$\|\varphi_3\|_{L^2(0,\ell)} \le C\varepsilon^{\frac{1}{2}} \|\phi\|^2_{H^2(\mathcal{S})} \le C\varepsilon^3.$$

Let us consider now

$$\varphi_4(\varepsilon z) = \int_{\mathbb{R}} 3[\mathtt{H}^2 - H^2]\phi H_x.$$

Since $\mathtt{H} = H + \phi_1$ and ϕ_1 can be estimated as

$$|\phi_1(x,z)| \le C\varepsilon(|x|^2 + 1)\, e^{-c|x|},$$

we easily see that

$$\|\varphi_4\|_{L^2(0,\ell)} \le C\varepsilon^{\frac{3}{2}} \|\phi\|_{H^2(\mathcal{S})} \le C\varepsilon^3 .$$

These terms define compact operators similarly as before.

In summary, we have

$$\|\varphi(\phi)\|_{L^2(0,\ell)} \le C\varepsilon^3. \tag{7.1}$$

8. The reduced equation for f: Proof of the theorem. In this section we set up an equation relating $\mathbf{f}$ such that for the solution ϕ of (3.9)–(3.10) obtained via Proposition 5.1 one has that the coefficient $c(\varepsilon z)$ is identically zero. To achieve this we multiply first the equation against H_x and integrate only in x. The equation $c = 0$ is then equivalent to the relation

$$\int_{\mathbb{R}} E_2 H_x dx + \varphi(\phi) = 0.$$

Using the estimates in the previous sections we then find that these relations are equivalent to the following nonlinear, nonlocal, differential equation for $\mathbf{f}$:

$$\mathcal{L}(\mathbf{f}) \equiv \varepsilon \mathbf{f}'' + (\varepsilon\gamma_3 + \gamma_4)\mathbf{f} = \varepsilon\gamma_5(\varepsilon z) + \varepsilon^2 M_\varepsilon. \tag{8.1}$$

We further set

$$\mathbf{f} = \varepsilon \frac{\gamma_5}{\gamma_4 + \varepsilon\gamma_3} + \hat{\mathbf{f}}.$$

Then (8.1) becomes a nonlocal equation for $\hat{\mathbf{f}}$,

$$\mathcal{L}(\hat{\mathbf{f}}) \equiv \varepsilon \hat{\mathbf{f}}'' + (\varepsilon\gamma_3 + \gamma_4)\hat{\mathbf{f}} = \varepsilon^2 M_\varepsilon. \tag{8.2}$$

The operators $M_\varepsilon = M_\varepsilon(\hat{\mathbf{f}})$ can be decomposed into the following form:

$$M_\varepsilon(\hat{\mathbf{f}}) = A_\varepsilon(\hat{\mathbf{f}}) + K_\varepsilon(\hat{\mathbf{f}}),$$

where K_ε is uniformly bounded in $L^2(0,\ell)$ for $\hat{\mathbf{f}}$ satisfying constraints (2.26) and is also compact. The operator A_ε is Lipschitz in this region:

$$\|A_\varepsilon(\hat{\mathbf{f}}_1) - A_\varepsilon(\hat{\mathbf{f}}_2)\|_{L^2(0,\ell)} \le C\varepsilon \|\hat{\mathbf{f}}_1 - \hat{\mathbf{f}}_2\|.$$

The functions $\gamma_i, i = 1, 2$, are smooth. Furthermore, we have

$$\gamma_4 = \frac{4}{3}\left(\int_{\mathbb{R}} H_x^2\right)^{-1} a_t(0,\theta) > 0.$$

We will solve now (8.2). First we need to use assumption (1.8) to deal with the invertibility of $\mathcal{L}$. We have the following lemma.

LEMMA 8.1. *Assume that condition* (1.8) *holds. If $d \in L^2(0,\ell)$, then there is a unique solution $\hat{\mathbf{f}} \in H^2(0,\ell)$ of $\mathcal{L}(\hat{\mathbf{f}}) = d$ which is ℓ-periodic and satisfies*

$$\varepsilon\|\hat{\mathbf{f}}''\|_{L^2(0,\ell)} + \sqrt{\varepsilon}\|\hat{\mathbf{f}}'\|_{L^2(0,\ell)} + \|\hat{\mathbf{f}}\|_{L^\infty(0,\ell)} \le C\varepsilon^{-1/2}\|d\|_{L^2(0,\ell)}.$$

Moreover, if d is in $H^2(0,\ell)$, then

$$\varepsilon\|\hat{\mathbf{f}}''\|_{L^2(0,\ell)} + \|\hat{\mathbf{f}}'\|_{L^2(0,\ell)} + \|\hat{\mathbf{f}}\|_{L^\infty(0,\ell)} \le C[\|d''\|_{L^2(0,\ell)} + \|d'\|_{L^2(0,\ell)}] + C\|d\|_{L^2(0,\ell)}.$$

Let us accept for the moment the validity of this result and let us conclude the proof of the theorem. From the contraction mapping principle, the equation

$$\mathcal{L}\hat{\mathbf{f}} = g$$

is uniquely solvable for $\hat{\mathbf{f}}$ satisfying (2.26) if $\|g\|_2 < \varepsilon^{\frac{3}{2}+\rho}$ for some $\rho > 0$. The desired result for the full problem (8.2) then follows directly from Schauder's fixed point theorem. In fact, refining the fixed point region, we can actually get $\|\hat{\mathbf{f}}\| = O(\varepsilon^{3/2})$ for the solution.

Proof of Lemma 8.1. We consider the boundary value problem

$$\mathcal{L}(\hat{\mathbf{f}}) = d, \quad \hat{\mathbf{f}}(0) = \hat{\mathbf{f}}(\ell), \qquad \hat{\mathbf{f}}'(0) = \hat{\mathbf{f}}'(\ell). \tag{8.3}$$

We notice that it suffices to show Lemma 8.1 with

$$\mathcal{L}_1(\hat{\mathbf{f}}) = \varepsilon\beta^{-2}\hat{\mathbf{f}}'' + \hat{\mathbf{f}},$$

where $\beta = \sqrt{\gamma_4}$. We make the following *Liouville transformation* (cf. [20]):

$$\begin{aligned}
&\ell_0 = \int_0^\ell \beta(\theta)d\theta, \quad t = \frac{\int_0^\theta \beta(\theta)d\theta}{\ell_0}\pi, \quad \lambda_0 = \frac{\ell_0^2}{\pi^2},\\
&\Psi(\theta) = (\beta(\theta))^{-\frac{1}{2}}, \quad y(t) = \Psi^{-1}(\theta)\hat{\mathbf{f}}(\theta), \quad q(t) = \frac{\ell_0^2}{\pi^2}\frac{\Psi_{\theta\theta}}{\beta^2\Psi},\\
&\tilde{d}(t) = \Psi^{-1}(\theta)d(\theta).
\end{aligned}$$

Then (8.3) with $\mathcal{L}$ replaced by $\mathcal{L}_1$ is transformed into

$$\tilde{\mathcal{L}}_2(y) = \varepsilon(y'' + q(t)y) + \lambda_0 y = \tilde{d}, \quad y(0) = y(\pi), \qquad y'(0) = y'(\pi), \tag{8.4}$$

and it then suffices to establish the estimates in Lemma 8.1 for the solution of this problem in terms of the corresponding norms of $\tilde{d}$. It is standard that the eigenvalue problem

$$y'' + q(t)y + \lambda y = 0, \quad y(0) = y(\pi), \qquad y'(0) = y'(\pi) \tag{8.5}$$

has an infinite sequence of eigenvalues λ_k, $k \geq 0$, with an associated orthonormal basis in $L^2(0,\pi)$, $\{y_k\}$, constituted by eigenfunctions. A result in [20] provides asymptotic expressions as $k \to +\infty$ for these eigenvalues and eigenfunctions, which turn out to correspond to those for $q \equiv 0$. We have

$$\sqrt{\lambda_k} = 2k + O\Big(\frac{1}{k^3}\Big), \quad k \to \infty. \tag{8.6}$$

Problem (8.4) is then solvable if and only if $\lambda_k\varepsilon \neq \lambda_0$ for all $k \geq 1$. In such a case, the solution to (8.3) then can be written as

$$y(t) = \sum_{k=1}^{\infty} \frac{\tilde{d}_k}{\lambda_0 - \lambda_k\varepsilon} y_k(t)$$

with this series convergent in L^2. Hence

$$\|y\|^2_{L^2(0,\pi)} = \sum_{k=0}^{\infty} \frac{|\tilde{d}_k|^2}{(\lambda_0 - \lambda_k\varepsilon)^2}.$$

We then choose ε such that

$$|4k^2\varepsilon - \lambda_0| \;\geq\; c\sqrt{\varepsilon} \tag{8.7}$$

for all k, where c is small. This corresponds precisely to condition (1.8). From (8.6) we then find that $|\lambda_0 - \lambda_k\varepsilon| \geq \frac{c}{2}\sqrt{\varepsilon}$ if ε is also sufficiently small. It follows that $\|y\|_{L^2(0,\pi)} \leq C\varepsilon^{-\frac{1}{2}}\|\tilde{d}\|_{L^2(0,\pi)}$. Next we notice that

$$\begin{aligned}|y(t)| &\leq \sum_{k=1}^{\infty}\left|\frac{\tilde{d}_k y_k(t)}{\lambda_0 - \lambda_k\varepsilon}\right| \\ &\leq \left(\sum_{k=1}^{\infty}\tilde{d}_k^2 y_k^2(t)\right)^{1/2}\left(\sum_{k=1}^{\infty}\frac{1}{(\lambda_0 - \lambda_k\varepsilon)^2}\right)^{1/2} \\ &\leq \frac{C}{\sqrt{\varepsilon}}\|\tilde{d}\|_{L^2(0,\pi)};\end{aligned}$$

hence the L^∞ estimate for y follows, and thus we get

$$\varepsilon\|y'\|_{L^2(0,\pi)} + \|y\|_{L^\infty(0,\pi)} \leq C\varepsilon^{-\frac{1}{2}}\|\tilde{d}\|_{L^2(0,\pi)}.$$

Observe also that

$$\|y'\|^2_{L^2(0,\pi)} \leq C\sum_{k=0}^{\infty}|\tilde{d}_k|^2\frac{1+|\lambda_k|}{(\lambda_0 - \lambda_k\varepsilon)^2} \leq C\sum_{k=0}^{\infty}(1+k^4)|\tilde{d}_k|^2.$$

Besides, if d is in $H^2(0,\pi)$ with $d(0) = d(\pi)$, $d'(0) = d'(\pi)$, then the sum $\sum_k k^4 d_k^2$ is finite and bounded by the H^2 norm of d. This and the equation automatically imply

$$\varepsilon\|y''\|_{L^2(0,\pi)} + \|y'\|_{L^2(0,\pi)} + \|y\|_{L^\infty(0,\pi)} \leq C\|\tilde{d}\|_{H^2(0,\pi)},$$

and the proof is complete. □

Remark 8.1. In section 3 of [26], an equivalent form of (8.2) was also derived for a system of singularly perturbed elliptic equations on N-dimensional domains ($N \geq 2$). There it was assumed that $\gamma_4(\theta) < 0$ (condition (A7) in [26]). It was also observed that when $\gamma_4 > 0$, there is a resonance of eigenvalues hitting 0.

Acknowledgments. We thank one of the referees for pointing out [26] to us. We also thank the second referee for valuable comments regarding the bibliography and the history of the problem as well as for a careful and insightful revision of the manuscript.

REFERENCES

[1] N. D. Alikakos and P. W. Bates, *On the singular limit in a phase field model of phase transitions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1988), pp. 141–178.

[2] N. D. Alikakos, P. W. Bates, and G. Fusco, *Solutions to the nonautonomous bistable equation with specified Morse index.* I. *Existence*, Trans. Amer. Math. Soc., 340 (1993), pp. 641–654.

[3] N. D. Alikakos, P. W. Bates, and X. Chen, *Periodic traveling waves and locating oscillating patterns in multidimensional domains*, Trans. Amer. Math. Soc., 351 (1999), pp. 2777–2805.

[4] N. D. Alikakos, X. Chen, and G. Fusco, *Motion of a droplet by surface tension along the boundary*, Calc. Var. Partial Differential Equations, 11, (2000), pp. 233–306.

[5] N. D. Alikakos and H. C. Simpson, *A variational approach for a class of singular perturbation problems and applications*, Proc. Roy. Soc. Edinburgh Sect. A, 107 (1987), pp. 27–42.

[6] S. Allen and J. W. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metall., 27 (1979), pp. 1084–1095.

[7] S. Angenent, J. Mallet-Paret, and L.A. Peletier, *Stable transition layers in a semilinear boundary value problem*, J. Differential Equations, 67 (1987), pp. 212–242.
[8] E. N. Dancer and S. Yan, *Multi-layer solutions for an elliptic problem*, J. Differential Equations, 194 (2003), pp. 382–405.
[9] E. N. Dancer and S. Yan, *Construction of various types of solutions for an elliptic problem*, Calc. Var. Partial Differential Equations, 20 (2004), pp. 93–118.
[10] M. del Pino, *Layers with nonsmooth interface in a semilinear elliptic problem*, Comm. Partial Differential Equations, 17 (1992), pp. 1695–1708.
[11] M. del Pino, *Radially symmetric internal layers in a semilinear elliptic system*, Trans. Amer. Math. Soc., 347 (1995), pp. 4807–4837.
[12] M. del Pino, M. Kowalczyk, and J. Wei, *Concentration on curves for nonlinear Schrödinger equations*, Comm. Pure Appl. Math., 60 (2007), pp. 113–146.
[13] A. S. do Nascimento, *Stable transition layers in a semilinear diffusion equation with spatial inhomogeneities in N-dimensional domains*, J. Differential Equations, 190 (2003), pp. 16–38.
[14] P. C. Fife, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497–521.
[15] P. Fife and W. M. Greenlee, *Interior transition layers for elliptic boundary value problems with a small parameter*, Russian Math. Surveys, 29 (1974), pp. 103–131.
[16] A. Floer and A. Weinstein, *Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential*, J. Funct. Anal., 69 (1986), pp. 397–408.
[17] M. Kowalczyk, *On the existence and Morse index of solutions to the Allen-Cahn equation in two dimensions*, Ann. Mat. Pura Appl. (4), 184 (2005), pp. 17–52.
[18] R. V. Kohn and P. Sternberg, *Local minimizers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 11 (1989), pp. 69–84.
[19] J. Hale and K. Sakamoto, *Existence and stability of transition layers*, Japan J. Appl. Math., 5 (1988), pp. 367–405.
[20] B. M. Levitan and I. S. Sargsjan, *Sturm-Liouville and Dirac Operators*, Math. Appl. (Soviet Ser.) 59, Kluwer Academic Publishers Group, Dordrecht, 1991.
[21] A. Malchiodi, *Adiabatic limits of closed orbits for some Newtonian systems in $\mathbb{R}^n$*, Asymptot. Anal., 25 (2001), pp. 149–181.
[22] A. Malchiodi and M. Montenegro, *Boundary concentration phenomena for a singularly perturbed elliptic problem*, Commun. Pure Appl. Math., 55 (2002), pp. 1507–1568.
[23] A. Malchiodi and M. Montenegro, *Multidimensional boundary layers for a singularly perturbed Neumann problem*, Duke Math. J., 124 (2004), pp. 105–143.
[24] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 357–383.
[25] Y. Nishiura and H. Fujii, *Stability of singularly perturbed solutions to systems of reaction–diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 1726–1770.
[26] K. Sakamoto, *Internal layers in high-dimensional domains*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 59–401.
[27] K. Sakamoto, *Infinitely many fine modes bifurcating from radially symmetric internal layers*, Asymptot. Anal., 42 (2005), pp. 55–104.
[28] K. Sakamoto, *Construction and stability analysis of transition layer solutions in reaction-diffusion systems*, Tohoku Math. J. (2), 42 (1990), pp. 17–44.
[29] J. Shatah and C. Zeng, *Periodic solutions for Halmitonian systems under strong constraining forces*, J. Differential Equations, 186 (2002), pp. 572–585.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1565–1587

EXPONENTIAL HOMOGENIZATION OF LINEAR SECOND ORDER ELLIPTIC PDEs WITH PERIODIC COEFFICIENTS*

VLADIMIR KAMOTSKI†, KARSTEN MATTHIES‡, AND VALERY P. SMYSHLYAEV‡

Abstract. A problem of homogenization of a divergence-type second order uniformly elliptic operator is considered with arbitrary bounded rapidly oscillating periodic coefficients, either with periodic "outer" boundary conditions or in the whole space. It is proved that if the right-hand side is Gevrey regular (in particular, analytic), then by optimally truncating the full two-scale asymptotic expansion for the solution one obtains an approximation with an exponentially small error. The optimality of the exponential error bound is established for a one-dimensional example by proving the analogous lower bound.

Key words. homogenization, elliptic equation, exponential asymptotics, Gevrey regularity, and analyticity

AMS subject classifications. 35B27, 35C20, 35J25, 74Q05

DOI. 10.1137/060651045

1. Introduction.

Classical homogenization theory describes the relation of solutions $u^\varepsilon(x)$ of boundary value problems with rapidly oscillating coefficients to solutions $u_0(x)$ of a homogenized problem, i.e., a problem without rapidly oscillating coefficients. In appropriate function spaces convergence can be established as $\varepsilon \to 0$ (with ε describing the period or wavelength of the coefficients' oscillations); see, e.g., [1, 2, 3, 4] and the references therein. For particular homogenization problems, e.g., for those described by linear second order elliptic PDEs with periodic coefficients, the rate of convergence with respect to ε can often also be determined, see, e.g., [4, 5, 6, 7]. The order of convergence can sometimes be improved further by constructing higher order correctors. The presence of a boundary creates additional "boundary layers," which substantially complexifies the problem of constructing the higher order terms; see, e.g., [5, 3, 4, 8, 9]. However, in the absence of the boundary, either for a problem with outer periodicity conditions or in the whole space (away from the spectrum), higher order terms can often be explicitly constructed. In particular, under the assumptions of sufficient regularity of the coefficients and the right-hand side of the equation, it is possible to construct and rigorously justify a full two-scale asymptotic expansion for $u^\varepsilon(x)$, i.e., to establish the error bounds both for linear problems (e.g., [3, 10]) and even for appropriate nonlinear ones [11].

The above can be referred to, in the context of homogenization, as "homogenization in all orders," by analogy with "asymptotics in all orders": by appropriately

*Received by the editors January 27, 2006; accepted for publication (in revised form) August 21, 2006; published electronically January 12, 2007. This work was partly supported by Bath Institute for Complex Systems (EPSRC grant GR/S86525/01) and by the Institute for Mathematics and Its Applications (IMA), University of Minnesota. Part of this work was done during the authors' visits to IMA during February–June 2005.

http://www.siam.org/journals/sima/38-5/65104.html

†Corresponding author. Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (vk209@maths.bath.ac.uk). This author's research was partially supported by grant RFBR 04-01-00522a.

‡Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (km230@maths.bath.ac.uk, vps@maths.bath.ac.uk). The second author's work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Schwerpunktprogramm "Modellierung, Analysis und effektive Simulation von Mehrskalenproblemen."

truncating the infinite asymptotic series one arrives at an asymptotic approximation to the actual solution u^ε with accuracy of any desirable polynomial order in ε as $\varepsilon \to 0$. We address in this paper the question of homogenization "beyond all orders," i.e., with an exponentially small error, via an optimal truncation of the (generally divergent) asymptotic expansion. The ideas of exponential asymptotics have been intensively developed in the recent literature (see, e.g., [12] and the references therein); however, not that much progress has been achieved in this direction specifically for problems of averaging and more specifically of homogenization. An exponential averaging technique was developed for ODEs by Neishtadt [13] and recently adjusted to PDEs with a temporal [14] and then *one-dimensional spatial* oscillations [15].

To the best of our knowledge, the present work represents the first example of a rigorous analytic exponential averaging for truly *multidimensional spatial oscillations*, i.e., for multidimensional homogenization. On the other hand, exponentially accurate approximations are potentially relevant to the problem of achieving exponentially convergent numerical schemes for homogenization; see, e.g., [16].

We consider the abovementioned "classical" elliptic homogenization problems with periodic coefficients, both for the case of periodic boundary conditions and in the whole space. We will assume that the right-hand side is sufficiently regular (not only infinitely smooth as required for constructing the full asymptotic expansion, but additionally "Gevrey regular," in particular, analytic). Then we show that one obtains an approximation with an exponentially small error by optimally truncating the full two-scale asymptotic series for the solution. Importantly, the above exponential bounds are sharp in the sense that we establish analogous lower bounds for the error in an explicit but rather generic one-dimensional example.

The Gevrey regularity techniques have proved useful in exploring exponentially small effects in different problems, for example, in diffraction/scattering for describing the wave field in the shadow [17] and the asymptotic distribution of resonances [18], and in the one-dimensional exponential averaging [14, 15] for controlling the effect of Galerkin approximation of PDEs via ODEs. In the present work, however, the Gevrey regularity allows us to control the error of the truncation of a full asymptotic expansion both with respect to the short period or wavelength of the oscillations ε and the large number n of the terms in the truncated asymptotic series.

The next section gives a precise formulation of the problems and the statements of the main results, which are Theorems 1 and 1′, and specifically the exponential error bounds (17) and (19). The rest of the paper is devoted to the proof of the theorems, as well as of the optimality of the estimates (17) and (19) for an explicit one-dimensional example; see Theorem 5. In particular, for analytic right-hand sides $f(x)$, in Theorems 1 and 5, the exponential bounds (17) and (79) hold with $\beta = 1$, with the "rate" of decay (the constants C_2 and $\tilde{C}_2$) related to the imaginary part of the "nearest" singularity in the analytic continuation of $f(x)$ for complex x; see Remark 7.

2. Statement of the problem and main results. We consider a family of differential operators with rapidly oscillating periodic coefficients:

$$(L^\varepsilon u)(x) := -\nabla \cdot \left(A\left(\frac{x}{\varepsilon}\right) \nabla u \right)(x). \tag{1}$$

The matrix $A(y) = (A_{ij}(y))_{ij} \in L^\infty(\mathbb{T})$, $i, j = 1, \dots, d$, where $\mathbb{T} = \mathbb{R}^d/\mathbb{Z}^d$, $d \geq 1$, is a d-dimensional torus, is assumed to be symmetric[1] and uniformly elliptic; i.e.,

[1]The assumption of the symmetry of matrix $A(y)$ holds in most physically relevant examples, but could be waived for the purposes of this paper: the stated results would still hold at the expense of a slightly more complicated algebra in the exposition.

$A_{ij}(y) = A_{ji}(y)$ for any i, j and $y \in \mathbb{T}$ and there exists $\nu_0 > 0$ such that for all $\xi \in \mathbb{R}^d$ and $y \in \mathbb{T}$

$$A_{ij}(y)\xi_i\xi_j \geq \nu_0|\xi|^2. \tag{2}$$

Here and throughout the paper we use the Einstein summation convention with respect to repeated indices.

The main problem considered in this paper is for the right-hand side f being infinitely smooth and periodic with a "fixed" period chosen to be equal to unity and having zero mean, with the solution also required to have zero mean and to satisfy the periodic boundary conditions; cf. [3, 10]. Namely, assuming $\varepsilon^{-1} \in \mathbb{N}$ to be a large integer, we address the following homogenization problem: for a given f with zero-mean value

$$\langle f \rangle := \int_{\mathbb{T}} f(x)\,dx = 0, \tag{3}$$

we seek a solution to the problem

$$(L^\varepsilon u^\varepsilon)(x) = f(x) \text{ in } \mathbb{T}, \tag{4}$$

$$\langle u^\varepsilon \rangle := \int_{\mathbb{T}} u^\varepsilon(x)\,dx = 0. \tag{5}$$

Equation (4) is a "classical" model of periodic homogenization, physically corresponding to, e.g., stationary heat conduction, electric conductivity, linear elasticity in anti-plane shear, etc.

For a special class of functions f, namely for Gevrey regular functions, we will construct an exponentially accurate asymptotic approximation to u^ε. Thus, we adopt the following definition (cf., e.g., [20], [21]).

DEFINITION 1. *We say that a $C^\infty(\mathbb{T})$ function f is β-Gevrey regular, where $\beta \geq 1$, if there exists $B > 0$ such that for all $l \in \mathbb{N}$*

$$\|f\ ;\ H^l(\mathbb{T})\| \leq B^l(l!)^\beta, \tag{6}$$

where B may depend on f but is independent of l. We use notation $f \in \mathcal{G}^\beta(\mathbb{T})$.

Here and below we use the scale of Sobolev spaces $H^l(\mathbb{X})$, $l \in \mathbb{N}$, on a Riemannian manifold $\mathbb{X}$, with the norm

$$\|f\ ;\ H^l(\mathbb{X})\| = \sum_{|k|=l} \|D^k f\ ;\ L^2(\mathbb{X})\| + \|f\ ;\ L^2(\mathbb{X})\|, \tag{7}$$

where $\|\cdot\ ;\ L^2(\mathbb{X})\|$ is the standard L^2 norm on $\mathbb{X}$, and we adopt the following conventional multi-index notation: $k = (k_1, \ldots, k_d) \in \mathbb{Z}_+^d$, where $\mathbb{Z}_+ := \mathbb{N} \cup \{0\}$ is the set of nonnegative integers; $|k| := k_1 + \cdots + k_d$; and $D^k := \partial^{|k|}/\partial x_1^{k_1} \ldots \partial x_d^{k_d}$. We will also deal with $H^{-1}(\mathbb{X})$ norms, defined as duals to the space $H_0^1(\mathbb{X})$ of functions from $H^1(\mathbb{X})$ with zero mean.

Definition 1 gives one of several equivalent definitions of the Gevrey "extreme regularity" class $\mathcal{G}^\beta$ (see also [19]). In particular, for $\beta = 1$ the functions are from $\mathcal{G}^\beta$ if and only if they are real analytic; for $\beta > 1$ the functions are infinitely smooth but not necessarily analytic. A conventional way of clarifying these relations is by reformulating them in the Fourier space. For the above $\mathbb{T}$-periodic functions f, when represented by their Fourier series

$$f(x) = \sum_{p \in \mathbb{Z}^d} f_p \exp(2i\pi p \cdot x), \tag{8}$$

a sufficient condition for f to belong to $\mathcal{G}^\beta$ is for its Fourier coefficients f_p to decay exponentially with the "rate" $|p|^{1/\beta}$; i.e.,

$$|f_p| \le c_1 \exp\left(-c_2|p|^{1/\beta}\right) \tag{9}$$

with some p-independent positive constants c_1 and c_2. The latter is well known and can be seen, for example, by applying the Plancherel theorem to (7), using (9), then replacing the resulting series by "asymptotically equivalent" integrals, and finally employing the Stirling asymptotic formulae; see, e.g., [29, (6.1.37)]. Throughout the paper we will use various minor modifications of the direct implication of the Stirling formula for

$$\Gamma(z) := \int_0^\infty \exp(-s)s^{z-1}ds,\ z>0;\ \Gamma(l+1)=l!,\ l\in\mathbb{N}, \tag{10}$$

which we display below for the reader's convenience:

$$A_1\left(\frac{z}{e}\right)^{z-1/2} \le \Gamma(z) \le A_2\left(\frac{z}{e}\right)^{z-1/2},\quad z\ge 1, \tag{11}$$

with some "universal" constants A_1 and A_2.

Notice that for real-analytic f (9) holds with $\beta=1$, and the rate of exponential decay c_2 is determined by the absolute value of the imaginary part of the "nearest" singularity in the analytic continuation.

For any fixed $\varepsilon>0$ the problem (4)–(5) is a well-posed elliptic problem which has a unique solution $u^\varepsilon\in H^1(\mathbb{T})$. Given $n\in\mathbb{Z}_+$ we seek an approximation to this solution in the standard form of the appropriately truncated two-scale asymptotic series (cf., e.g., [3]):

$$u^{\varepsilon,n}(x) = \sum_{m=0}^{n+2}\varepsilon^m u^{(m)}\left(x,\frac{x}{\varepsilon}\right), \tag{12}$$

where the functions $u^{(m)}(x,y)$ are required to be periodic in the "fast" variable y. It is known that for the present problem one can construct in this way a full asymptotic expansion with $u^{(m)}$ adopting the following form (see, e.g., [3, 10]):

$$u^{(m)}(x,y) = \sum_{l=0}^{m}\sum_{|k|=l} N_k(y)D_x^k\, v_{m-l}(x), \tag{13}$$

where $N_0(y)\equiv 1$ and $N_k(y)$ are periodic solutions of the "main" ($|k|=1$) and "higher order" ($|k|>1$) "canonical" unit cell problems in the "fast" variable y. The functions $v_s(x)$, $s\ge 0$, solve certain recurrent systems of equations in the "slow" variable x (see [3]), which are briefly reviewed in the next section.

Before formulating the main result, for convenience of the future referencing, we combine (12) and (13) to give

$$u^{\varepsilon,n}(x) = \sum_{l=0}^{n+2}\varepsilon^l\sum_{|k|=l} N_k\left(\frac{x}{\varepsilon}\right)D_x^k\, V^{(n-l+2)}(x,\varepsilon). \tag{14}$$

The slowly varying part in (14) is a partial sum of the formal asymptotic series $V^{(\infty)}(x,\varepsilon)$ (see (29)):

$$V^{(M)}(x,\varepsilon) := \sum_{s=0}^{M} \varepsilon^s v_s(x). \tag{15}$$

The main results of the present paper are the following.

THEOREM 1. *Suppose $A \in L^\infty(\mathbb{T})$ and satisfies (2), and $f \in \mathcal{G}^\beta(\mathbb{T}), \beta \geq 1, \langle f \rangle = 0$. Let u^ε be the unique solution of (4)–(5). Then there exist ε-independent constants $C_1 > 0$, $C_2 > 0$, $\kappa_1 > 0$, and $\kappa_2 > \kappa_1$, such that for any n satisfying*

$$\kappa_1 \varepsilon^{-1/\beta} \leq n \leq \kappa_2 \varepsilon^{-1/\beta} \tag{16}$$

the approximation (14) has the error bound

$$\|u^\varepsilon - u^{\varepsilon,n} \ ; \ H^1(\mathbb{T})\| \leq C_1 \exp(-C_2 \varepsilon^{-\frac{1}{\beta}}). \tag{17}$$

The above result may be interpreted in the sense that if the (generally divergent) asymptotic series (12) is, for sufficiently small ε, "optimally" truncated by choosing $n = n(\varepsilon)$ according to (16), for example, $n(\varepsilon) = \left[\kappa_2 \varepsilon^{-1/\beta}\right]$ with the square brackets denoting the entire part, then this produces an exponentially small error in the sense of (17).

Note also that for less regular f the earlier results on the polynomial rather than exponential error (see, e.g., [5, Thm. 11.1], [3, section 4.2, Thm. 2]) will be a by-product of our analysis: if, e.g., f has a finite regularity in the scale of Sobolev spaces, say f belongs to $H^M(\mathbb{T})$ but does not belong to $H^{M+1}(\mathbb{T})$ for some M, one can construct only finitely many terms in the expansion (12). As a result one obtains only an error bound of polynomial order ε^n with a finite n related to M. On the other hand, if one assumes $f \in C^\infty$ but makes no assumption on the "rate" of growth of its H^l norms when $l \to \infty$, one does reproduce the "homogenization in all orders" with an error bound $C_n \varepsilon^n$ for any n. However, in the latter case one has no control on the growth of C_n as $n \to \infty$, which disallows any possible further "a priori" improvement of the error bound.

Let us also note that Theorem 1 can be generalized further in a number of ways. Assuming higher regularity of the coefficients A_{ij}, one can get in (17) the same rate of convergence but in stronger norms. One can also consider another case without the boundary for a "shifted" operator $L^\varepsilon + 1$ in entire $\mathbb{R}^d$ rather than for L^ε in a fixed domain with periodic boundary condition. Then the same exponential estimate holds; i.e., the following theorem can be obtained adapting the proof of Theorem 1 with minor changes.

THEOREM 1′. *Suppose $A \in L^\infty(\mathbb{T})$ and satisfies (2); $f \in \mathcal{G}^\beta(\mathbb{R}^d), \beta \geq 1$, i.e., $f \in C^\infty(\mathbb{R}^d)$ and there exists $B > 0$ such that for all $l \in \mathbb{N}, \|f \ ; \ H^l(\mathbb{R}^d)\| \leq B^l (l!)^\beta$. Let $u^\varepsilon \in H^1(\mathbb{R}^d)$ be the unique solution of*

$$(L^\varepsilon + 1) u^\varepsilon = f. \tag{18}$$

Then there exist ε-independent constants $C_1 > 0$, $C_2 > 0$, $\kappa_1 > 0$, and $\kappa_2 > \kappa_1$, such that for any n satisfying $\kappa_1 \varepsilon^{-1/\beta} \leq n \leq \kappa_2 \varepsilon^{-1/\beta}$ the corresponding asymptotic approximation of the form (12) has the error bound

$$\|u^\varepsilon - u^{\varepsilon,n} \ ; \ H^1(\mathbb{R}^d)\| \leq C_1 \exp(-C_2 \varepsilon^{-\frac{1}{\beta}}). \tag{19}$$

Note that in the latter case the explicit structure of the two-scale asymptotics (12) is slightly different from that of (14); see (57).

We expect similar results to be valid also for *nonlinear* elliptic divergence operators (cf. [11]). Accounting for the presence of a boundary is in general a difficult open problem; cf. [8, 9].

The proof of the theorems will be divided into three steps. First we derive a priori estimates on appropriate norms of the coefficients N_k and v_s in suitable functional spaces for fixed k and s in section 3. Then, in section 4, we estimate the right-hand side error term $L^\varepsilon u^{\varepsilon,n} - f$ for fixed n and ε (in the H^{-1} norm). In section 5 we translate this into the error estimates for $u^{\varepsilon,n(\varepsilon)} - u^\varepsilon$ via analysis of the "mean" and standard ellipticity estimates, and finally "minimize" the error by an optimal choice of $n(\varepsilon)$ dependence on ε. This establishes the desired exponential error bound and hence proves Theorem 1. Proof of Theorem 1′ follows the same strategy with minor technical alterations listed in Remark 2 immediately following the proof of Theorem 1.

Optimality of the exponential error bound (17) is proved in section 6 for an explicit one-dimensional example; see Theorem 5.

3. Recurrent relations and a priori estimates. We briefly describe the procedure for determining the coefficients in (14) and (15) (see, e.g., [3, 10] for the detailed derivation in a slightly different notation). Below, we give one possible way to summarize it.

First, the infinite series version of (14),

$$u^\varepsilon(x) \sim \sum_{l=0}^{\infty} \varepsilon^l \sum_{|k|=l} N_k\left(\frac{x}{\varepsilon}\right) D_x^k V^\infty(x,\varepsilon), \tag{20}$$

is formally substituted into (4). After appropriate differentiations and re-grouping the terms with equal powers of ε (treating at this stage $V^\infty(x,\varepsilon)$ as a "whole") we arrive at

$$\sum_{l=0}^{\infty} \varepsilon^{l-2} \sum_{|k|=l} \left\{L_y^1 N_k(y) - T_k(y)\right\}_{y=x/\varepsilon} D^k V^\infty(x,\varepsilon) \sim f(x), \tag{21}$$

where

$$N_0(y) \equiv 1, \quad T_0(y) \equiv 0, \tag{22}$$

and

$$|k| = 1 : T_k(y) = A_{ij,j}(y), \ k = e_i, \tag{23}$$

$$|k| \geq 2 : T_k(y) = \sum_{\substack{i,j=1,\dots,d \\ k'=k-e_i\geq 0}} \left((A_{ij}N_{k'})_{,j} + A_{ij}N_{k',j}\right)(y) + \sum_{\substack{i,j=1,\dots,d \\ k''=k-e_i-e_j\geq 0}} A_{ij}(y)N_{k''}(y). \tag{24}$$

Here we denote by e_i the unit ith axis vector in $\mathbb{Z}^d$ and adopt the standard convention denoting derivatives by the indices following the comma in the subscript; $k \geq 0$ for a multi-index k means $k_i \geq 0$ for any $1 \leq i \leq d$ ($k > 0$ will mean $k \geq 0$ and $k_i > 0$ for some i, with $k' < k$ meaning $k - k' > 0$, etc.).

We then require the "coefficients" $\left\{L_y^1 N_k(y) - T_k(y)\right\}$ in (21) to be independent of the fast variable y, i.e., to be equal to constants which are denoted by $-h_k$. This

implies that N_k are solutions to the following "cell problems" for $|k| \geq 1$:

$$L_y^1 N_k = T_k(y) - h_k \text{ in } \mathbb{T}, \tag{25}$$

$$\langle N_k \rangle = 0, \tag{26}$$

with periodic conditions for N_k.

The solvability condition for (25)–(26) implies, necessarily, that h_k are the mean values of T_k over the periodicity cell:

$$h_k = \langle T_k \rangle. \tag{27}$$

Combining (21) with (25) yields an infinite order formal asymptotic equation for the "slow" part $V^\infty(x,\varepsilon)$:

$$-\sum_{l=0}^{\infty} \varepsilon^{l-2} \sum_{|k|=l} h_k D^k V^\infty(x,\varepsilon) \sim f(x). \tag{28}$$

A formal asymptotic solution of (28) is in turn sought in the form of an "infinite order" version of (15):

$$V^\infty(x,\varepsilon) \sim \sum_{s=0}^{\infty} \varepsilon^s v_s(x). \tag{29}$$

The substitution of (29) into (28) with subsequent rearrangements and equating terms with the same powers of ε yields

$$-A^{\text{hom}}_{i,j=1} v_{s,ij}(x) = f_s(x), \tag{30}$$

$$\langle v_s \rangle = 0, \tag{31}$$

with the right-hand sides

$$f_0 = f, \tag{32}$$

$$f_s = \sum_{l=3}^{s+2} \sum_{|k|=l} h_k D_x^k v_{s-l+2}, \ s \geq 1. \tag{33}$$

In (30) $A^{\text{hom}} = \left(A^{\text{hom}}_{ij}\right)_{i,j=1}^d$ is a "classical" homogenized matrix, which is known to be positive definite (with the same ellipticity constant ν_0 as in (2)) and symmetric:

$$A^{\text{hom}}_{ij} = \langle A_{ij} \rangle + \langle A_{is} N_{e_j,s} \rangle = \begin{cases} h_{e_i+e_j}, & i = j, \\ \frac{1}{2} h_{e_i+e_j}, & i \neq j. \end{cases}$$

Notice that (30) is uniquely solvable for any $s \geq 0$: a necessary and sufficient condition for the solvability is $\langle f_s \rangle = 0$ which does hold for $s = 0$ by assumption (3) and for $s \geq 1$ by (33). The slowly varying terms v_s are hence found recurrently as solutions to *homogenized* equations (30) with constant coefficients on a torus $\mathbb{T}$.

The relations (22)–(27) and (30)–(33) are hence sufficient for uniquely identifying all N_k and v_s, respectively. Then the so-defined asymptotic "double series" (20), (29) provides a full asymptotic expansion of the solution $u^\varepsilon(x)$ "in all orders": in particular, its truncation $u^{\varepsilon,n}$ produces an error of polynomial order in ε (see, e.g., [3, section 4.2, Thm. 2] or section 4 below).

We next aim at estimating the quantities

$$\mathcal{N}^{(l)} := \max_{|k|=l} \|N_k \ ; \ H^1(\mathbb{T})\|, \tag{34}$$

$$\mathcal{V}^{(s,m)} := \|v_s \ ; \ H^m(\mathbb{T})\|. \tag{35}$$

We will prove the following lemma.

LEMMA 2. *Under the assumptions of Theorem* 1 *the following estimates hold for all* $l, s, m \in \mathbb{N} \cup \{0\}$: (i)

$$\mathcal{N}^{(l)} \le \left(M_{\mathcal{N}}\right)^l, \tag{36}$$

(ii)

$$\mathcal{V}^{(s,m)} \le \sum_{k=0}^{s} \left(M_{\mathcal{V}}\right)^{k+1} \|f \ ; \ H^{m+k}(\mathbb{T})\| \tag{37}$$

for appropriate constants $M_{\mathcal{N}}$ *and* $M_{\mathcal{V}}$, *depending only on* $\|A \ ; \ L^\infty(\mathbb{T})\|$ *and the ellipticity constant* ν_0 *(see* (2)*).*

Proof. Further on we will use the abbreviated notation $|\cdot|_l := \|\cdot \ ; \ H^l(\mathbb{T})\|$, $l \ge -1$ ($|\cdot|_0 := \|\cdot \ ; \ L^2(\mathbb{T})\|$) and denote by C, M_1, M_2, etc. various positive constants whose precise values are insignificant and can change during the proof.

(i) Due to the standard ellipticity estimates we have

$$|v|_1 \le C(\nu_0)|G|_0 \tag{38}$$

for a solution of $L^1 v = \nabla \cdot G$, $\langle v \rangle = 0$ with arbitrary $G \in \left(L^2(\mathbb{T})\right)^d$. So we deduce from (22)–(24), (25)–(26), and (27) for $|k| \ge 2$ that

$$|N_k|_1 \le C(\nu_0)\|A \ ; L^\infty(\mathbb{T})\| \left(\sum_{k'<k, \, |k-k'|=1} |N_{k'}|_1 + \sum_{k''<k, |k-k''|=2} |N_{k''}|_0 \right). \tag{39}$$

The latter reads in terms of (34) as

$$\mathcal{N}^{(l)} \le M_1 \mathcal{N}^{(l-1)} + M_2 \mathcal{N}^{(l-2)}, \quad l \ge 2, \tag{40}$$

and implies (36) by induction: from (22) we have $\mathcal{N}^{(0)} = 1$ and, due to (23) and (38), $\mathcal{N}^{(1)} \le C(\nu_0)\|A \ ; L^\infty(\mathbb{T})\| \le M_1$. Therefore choosing $M_{\mathcal{N}} > \max\{1, M_1 + (M_2)^{1/2}\}$ we arrive at (36).

(ii) Now turning to v_s, due to (30)–(33) we estimate for $s \ge 1$ and $m \ge 1$

$$|v_s|_m \le C(A^{\mathrm{hom}}) \sum_{l=3}^{s+2} \sum_{|k|=l} |h_k| |D_x^k v_{s-l+2}|_{m-2}. \tag{41}$$

This can be established, e.g., using again the ellipticity estimates applied to (30), which being an elliptic equation with constant coefficients can be differentiated m times. Applying also a version of the Poincaré inequality, which in our choice of the domain and the norms (see (7)) is the obvious estimate

$$\|g; H^k(\mathbb{T})\| \le \|g; H^l(\mathbb{T})\|, \ k \le l, \tag{42}$$

we conclude that $C(A^{\mathrm{hom}})$ can be chosen independently of m.

Since (24) and (27) obviously imply that $\max_{|k|=l}|h_k| \le C\left(\mathcal{N}^{(l-1)}+\mathcal{N}^{(l-2)}\right)$, from (41) we arrive after a straightforward manipulation at

$$\mathcal{V}^{(s,m)} \le C\sum_{r=1}^{s}\mathcal{N}^{(r+1)}r^{d-1}\mathcal{V}^{(s-r,m+r)}. \tag{43}$$

The latter in turn, combined with (36), implies that with large enough M_0

$$\mathcal{V}^{(s,m)} \le \sum_{r=1}^{s}M_0^r\mathcal{V}^{(s-r,m+r)}. \tag{44}$$

Let us finally show by induction in s that the latter is sufficient to deduce (37) with some $M_{\mathcal{V}} > 2M_0$. Indeed for $s=0$, due to (30) for all $m \ge 0$, we have $|v_0|_m \le |v_0|_{m+2} \le M_3|f|_m$ with M_3 independent of m, implying (37). Now we proceed with the induction step: suppose (37) holds for $s=0,\dots,S$ with a constant $M_{\mathcal{V}} > \max\{2M_0, M_3\}$. Then due to (44) we have

$$\begin{aligned}\mathcal{V}^{(S+1,m)} &\le \sum_{r=1}^{S+1}M_0^r\mathcal{V}^{(S+1-r,m+r)} \le \sum_{r=1}^{S+1}\sum_{k=0}^{S+1-r}M_0^rM_{\mathcal{V}}^{k+1}|f|_{m+r+k}\\ &\le \sum_{q=1}^{S+1}M_{\mathcal{V}}^{q+1}|f|_{m+q}\sum_{r=1}^{q}\left(\frac{M_0}{M_{\mathcal{V}}}\right)^r,\end{aligned}$$

which by our choice of $M_{\mathcal{V}}$ implies (37) for $s=S+1$. ∎

4. Remainder estimates. Next we derive estimates for the error in the right-hand side of the original equation (4) as a result of substitution into its left-hand side of the truncated asymptotic ansatz $u^{\varepsilon,n}$; see (12)–(15). The following lemma is in effect an implication of the above described formal asymptotic construction: it is supplemented by a more accurate bookkeeping of the structure of the remainder term $R^{\varepsilon,n}$ (as needed for purposes of this work), which is bound, by the construction, to contain only the terms of orders ε^{n+1} and ε^{n+2} for fixed n and small ε; cf. [3, 10].

LEMMA 3. *Under the assumptions of Theorem* 1 *one has* $L^\varepsilon u^{\varepsilon,n} = f + R^{\varepsilon,n}$ *with* $R^{\varepsilon,n} \in H^{-1}(\mathbb{T})$, *and*

$$\begin{aligned}R^{\varepsilon,n} = -\varepsilon^{n+1}\Bigg(&\sum_{l=0}^{n+2}\sum_{|k|=l}\left(\left(A_{ij}N_k\right)_{,j}+A_{ij}N_{k,j}\right)D_{x_i}D_x^kv_{n-l+2}\\ &+\sum_{l=0}^{n+1}\sum_{|k|=l}A_{ij}N_kD_{x_ix_j}D_x^k(v_{n-l+1}+\varepsilon v_{n-l+2})\\ &+\varepsilon\sum_{|k|=n+2}A_{ij}N_kD_{x_ix_j}D_x^kv_0\Bigg)\end{aligned} \tag{45}$$

(denoting $D_{x_i} := \partial/\partial x_i$, $D_{x_ix_j} := \partial^2/(\partial x_i\partial x_j)$*).*

Proof. The proof is a straightforward calculation by substituting the expansion (14), (15) into (4). We notice that since $A_{ij} \in L^\infty(\mathbb{T})$, $N_k \in H^1(\mathbb{T})$, and $v_s \in C^\infty(\mathbb{T})$,

all the "product" terms in (45) are in $H^{-1}(\mathbb{T})$. For example,

$$\begin{aligned}(\,(A_{ij}N_k)_{,j}\,)\left(\frac{x}{\varepsilon}\right)D_{x_i}D_x^kV^{(n-l+2)}(x,\varepsilon) &= \varepsilon\frac{\partial}{\partial x_j}\left((A_{ij}N_k)\left(\frac{x}{\varepsilon}\right)D_{x_i}D_x^kV^{(n-l+2)}(x,\varepsilon)\right)\\ &\quad -\varepsilon\left((A_{ij}N_k)\left(\frac{x}{\varepsilon}\right)D_{x_ix_j}D_x^kV^{(n-l+2)}(x,\varepsilon)\right),\end{aligned}\tag{46}$$

with the first term in the latter expression being a derivative of an L^2 function and the last one an L^2 function itself.

The terms up to order $O(\varepsilon^n)$ equal f by (14), (15), (30), (32), and (33). Via direct inspection,

$$\begin{aligned}(L^\varepsilon u^{\varepsilon,n})(x) &= -\nabla\cdot\left(A\left(\frac{x}{\varepsilon}\right)\nabla u^{\varepsilon,n}\right)(x)\\ &= -\nabla\cdot\left(A\left(\frac{x}{\varepsilon}\right)\nabla\sum_{l=0}^{n+2}\varepsilon^l\sum_{|k|=l}N_k\left(\frac{x}{\varepsilon}\right)D_x^k\,V^{(n-l+2)}(x,\varepsilon)\right)\\ &= \sum_{l=0}^{n+2}\varepsilon^{l-2}\sum_{|k|=l}(L_y^1N_k)\left(\frac{x}{\varepsilon}\right)D_x^kV^{(n-l+2)}(x,\varepsilon)\\ &\quad -\sum_{l=0}^{n+2}\varepsilon^{l-1}\sum_{|k|=l}(\,(A_{ij}N_k)_{,j}+A_{ij}N_{k,j}\,)\left(\frac{x}{\varepsilon}\right)D_{x_i}D_x^kV^{(n-l+2)}(x,\varepsilon)\\ &\quad -\sum_{l=0}^{n+2}\varepsilon^{l}\sum_{|k|=l}(A_{ij}N_k)\left(\frac{x}{\varepsilon}\right)D_{x_ix_j}D_x^kV^{(n-l+2)}(x,\varepsilon).\end{aligned}$$

Now we replace $V^{(n-l+2)}(x,\varepsilon)$ by $V^{(n-l+1)}(x,\varepsilon)+\varepsilon^{n-l+2}v_{n-l+2}$ in the second term and by $V^{(n-l)}(x,\varepsilon)+\varepsilon^{n-l+1}v_{n-l+1}+\varepsilon^{n-l+2}v_{n-l+2}$ in the last term; see (15). These "remainders" containing v_{n-l+1} and v_{n-l+2} as well as the term corresponding to $l=n+2$ in the last sum, all being of order ε^{n+1} and ε^{n+2}, produce exactly $R^{\varepsilon,n}$. Therefore we have

$$\begin{aligned}(L^\varepsilon u^{\varepsilon,n})(x) &= R^{\varepsilon,n}+\sum_{l=0}^{n+2}\varepsilon^{l-2}\sum_{|k|=l}(L_y^1N_k)\left(\frac{x}{\varepsilon}\right)D_x^kV^{(n-l+2)}(x,\varepsilon)\\ &\quad -\sum_{l=0}^{n+1}\varepsilon^{l-1}\sum_{|k|=l}(\,(A_{ij}N_k)_{,j}+A_{ij}N_{k,j}\,)\left(\frac{x}{\varepsilon}\right)D_{x_i}D_x^kV^{(n-l+1)}(x,\varepsilon)\\ &\quad -\sum_{l=0}^{n}\varepsilon^{l}\sum_{|k|=l}(A_{ij}N_k)\left(\frac{x}{\varepsilon}\right)D_{x_ix_j}D_x^kV^{(n-l)}(x,\varepsilon).\end{aligned}$$

Now we change the summation indices in the two latter terms to obtain the first $n+2$ terms of the series (21)

$$\begin{aligned}(L^\varepsilon u^{\varepsilon,n})(x) &= R^{\varepsilon,n}+\sum_{l=2}^{n+2}(L_y^1N_k-T_k)D^kV^{(n-l+2)}\\ &= R^{\varepsilon,n}-\sum_{l=2}^{n+2}\varepsilon^{l-2}\sum_{|k|=l}h_kD^kV^{(n-l+2)}=f+R^{\varepsilon,n},\end{aligned}$$

having used in the last equality (30)–(33). ∎

Using the above formula (45) for the remainder term, we estimate $R^{\varepsilon,n}$ with an explicit dependence on both ε and n as in the following lemma ($[\cdot]$ denotes the entire part).

LEMMA 4. *Under the assumptions of Theorem* 1, *there exist* C_3, ε_0 *such that for all* n *and all* $0 < \varepsilon < \varepsilon_0$ *the remainder term* $R^{\varepsilon,n}$ *can be estimated as follows:*

$$\|R^{\varepsilon,n};\ H^{-1}(\mathbb{T})\| \le C_3^n \varepsilon^{n+1} \|f;\ H^{n+5+[d/2]}(\mathbb{T})\|. \tag{47}$$

Proof. Here we combine the formula for the error term in Lemma 3 with the estimates in Lemma 2. To estimate the H^{-1} norm of the aggregates like $((A_{ij}N_k)D^{|k|+1}v)_{,j}$ (see (46)) and $A_{ij}N_kD^{|k|+2}v$ we need to ensure that $D^{|k|+2}v$ is in L^∞. By the Sobolev embedding theorems this holds if $v \in H^{|k|+3+[d/2]}$. Therefore

$$\begin{aligned}
\|R^{\varepsilon,n};\ H^{-1}(\mathbb{T})\| = \Bigg\| & \varepsilon^{n+1} \Bigg(\sum_{l=0}^{n+2} \sum_{|k|=l} \big((A_{ij}N_k)_{,j} + A_{ij}N_{k,j} \big) D_{x_i} D_x^k v_{n-l+2} \\
& + \sum_{l=0}^{n+1} \sum_{|k|=l} A_{ij}N_k D_{x_i,x_j} D_x^k (v_{n-l+1} + \varepsilon v_{n-l+2}) \\
& + \varepsilon \sum_{|k|=n+2} A_{ij}N_k D_{x_i,x_j} D_x^k v_0 \Bigg);\ H^{-1}(\mathbb{T}) \Bigg\| \\
\le \varepsilon^{n+1} \Bigg(& C \sum_{l=0}^{n+2} l^{d-1} \|A; L^\infty\| \mathcal{N}^{(l)} \mathcal{V}^{(n-l+2,l+3+[d/2])} \\
& + C \sum_{l=0}^{n+1} l^{d-1} \|A; L^\infty\| \mathcal{N}^{(l)} \Big(\mathcal{V}^{(n-l+1,l+3+[d/2])} \\
& \qquad + \varepsilon \mathcal{V}^{(n-l+2,l+3+[d/2])} \Big) \\
& + C\varepsilon n^{d-1} \|A; L^\infty\| \mathcal{N}^{(n+2)} \mathcal{V}^{(0,n+5+[d/2])} \Bigg) \\
\le C\varepsilon^{n+1} & \Big(\max\{M_{\mathcal{N}}, M_{\mathcal{V}}\} \Big)^{n+2} n^d \|f; H^{n+5+[d/2]}\|.
\end{aligned}$$

Here we have used again the Poincaré inequality (42). An appropriate choice of C_3 yields the result. ∎

Remark 1. The last lemma could also be used to rederive results for finite regularity f, or smooth f, which are not necessarily in any Gevrey space $\mathcal{G}^\beta$. If, for example, f has finite regularity, i.e., $f \in H^M(\mathbb{T})$ and $f \notin H^{M+1}(\mathbb{T})$ for some M, then, as the above procedure demonstrates, only a finite number of terms in the asymptotic expansion can be constructed, and the H^{-1} norms can be bounded only for $n < M - 4 - d/2$. If, however, $f \in C^\infty(\mathbb{T})$, but no assumptions are made on the rate of growth of its H^l norms for large l, the estimate (47) still holds for any n, but there is no control over the growth of the Sobolev norms of f with n in the right-hand side of (47). The latter would prevent us from improving the polynomial "asymptotics in all orders" any further. This highlights the importance of the Gevrey extreme regularity of f for the exponential error bounds.

5. Proof of Theorem 1. The proof of the theorem is now essentially a corollary of Lemma 4, the estimates (6) holding due to the assumption of Gevrey regularity

of f and standard elliptic regularity theory. Let us first introduce a "normalized" approximation

$$\tilde{u}^{\varepsilon,n} := u^{\varepsilon,n} - \langle u^{\varepsilon,n} \rangle. \tag{48}$$

By the elliptic regularity theory for all n we have

$$\|\tilde{u}^{\varepsilon} - u^{\varepsilon,n} \,;\, H^1(\mathbb{T})\| \le C \|R^{\varepsilon,n};\, H^{-1}(\mathbb{T})\|. \tag{49}$$

Using Lemma 4, we obtain

$$\|\tilde{u}^{\varepsilon} - u^{\varepsilon,n} \,;\, H^1(\mathbb{T})\| \le C C_3^n \varepsilon^{n+1} \|f; H^{n+4+[d/2]}\|. \tag{50}$$

Let us next show that the mean $\langle u^{\varepsilon,n} \rangle$ can also be estimated in a similar way. Due to representations (12), (13) we have

$$\langle u^{\varepsilon,n} \rangle = \sum_{m=0}^{n+2} \varepsilon^m \sum_{l=0}^{m} \sum_{|k|=l} \left\langle N_k \left(\frac{x}{\varepsilon} \right) D_x^k \, v_{m-l}(x) \right\rangle. \tag{51}$$

Note that $\langle N_k \rangle = 0$; therefore for any $s > 0$ the functions $\left((-\Delta_y)^{-s} N_k(\cdot)\right)(y)$ and $\left((-\Delta_x)^{-s} N_k(\frac{\cdot}{\varepsilon})\right)(x)$ are correctly defined functions with zero mean, using, for example, the Fourier representation for $(-\Delta_y)$ on a torus $\mathbb{T}$. Moreover, they are linked via

$$\left((-\Delta_x)^{-s} N_k \left(\frac{\cdot}{\varepsilon} \right)\right)(x) = \varepsilon^{2s} \left((-\Delta_y)^{-s} N_k(\cdot)\right) \left(\frac{x}{\varepsilon} \right) \tag{52}$$

(recall $\varepsilon^{-1} \in \mathbb{N}$). Thus, integrating (51) by parts sufficiently many times, we get

$$\begin{aligned} &\langle u^{\varepsilon,n} \rangle \\ &= \sum_{m=0}^{n+2} \varepsilon^m \sum_{l=0}^{m} \sum_{|k|=l} \varepsilon^{n+2-m} \left\langle \left((-\Delta_y)^{-\frac{n+2-m}{2}} N_k \right) \left(\frac{x}{\varepsilon} \right) (-\Delta_x)^{\frac{n+2-m}{2}} D_x^k \, v_{m-l}(x) \right\rangle. \end{aligned} \tag{53}$$

Now, via the Cauchy–Schwartz inequality,

$$|\langle u^{\varepsilon,n} \rangle| \le \varepsilon^{n+2} \sum_{m=0}^{n+2} \sum_{l=0}^{m} \sum_{|k|=l} \|(-\Delta_y)^{-\frac{n+2-m}{2}} N_k;\, L^2(\mathbb{T})\| \, \|v_{m-l};\, H^{n+2-m+l}(\mathbb{T})\|. \tag{54}$$

Therefore, applying the Poincaré inequality (42) to the first norm and then using estimates (36), (37) of Lemma 2, we have

$$\begin{aligned} |\langle u^{\varepsilon,n} \rangle| &\le \varepsilon^{n+2} \sum_{m=0}^{n+2} \sum_{l=0}^{m} C l^{d-1} \left(\sup_{|k|=l} \|N_k;\, H^1\| \right) \sum_{p=0}^{m-l} M_{\mathcal{V}}^{p+1} \|f; H^{n+2-m+l+p}\| \\ &\le \varepsilon^{n+2} \|f; H^{n+2}\| \sum_{m=0}^{n+2} \sum_{l=0}^{m} C l^d M_{\mathcal{N}}^l \sum_{p=0}^{m-l} M_{\mathcal{V}}^{p+1} \\ &\le \varepsilon^{n+2} \|f; H^{n+2}\| \sum_{m=0}^{n+2} \sum_{l=0}^{m} C l^d M_{\mathcal{N}}^l M_{\mathcal{V}}^{m-l}, \end{aligned}$$

and therefore

$$|\langle u^{\varepsilon,n}\rangle| \le \varepsilon^{n+2} C C_4^n \|f; H^{n+2}\|. \tag{55}$$

Combining the latter with (50) we finally get

$$\|u^\varepsilon - u^{\varepsilon,n}\ ;\ H^1(\mathbb{T})\| \le C C_5^n \varepsilon^{n+1} \|f; H^{n+5+[d/2]}\| \tag{56}$$

for small enough ε with an appropriate constant $C_5 > 0$.

Further, by (6), $\|f; H^{n+5+[d/2]}\| \le B^{n+5+[d/2]}((n+5+[d/2])!)^\beta$. Using the Stirling formula (11) for the factorial (implying $M! = \Gamma(M+1) \le C M^{M+1/2} e^{-M}$ for any $M \in \mathbb{N}$ with some $C > 0$), we obtain

$$\begin{aligned}
&\|u^\varepsilon - u^{\varepsilon,n}\ ;\ H^1(\mathbb{T})\| \\
&\le C C_5^n \varepsilon^{n+1} B^{n+5+[d/2]} (n+5+[d/2])^{(n+5+[d/2])\beta} e^{-\beta(n+5+[d/2])} n^{1/2} \\
&\le C \varepsilon^{n+1} (C_6 B)^n n^{n\beta} = C\varepsilon \exp(n \ln(n^\beta C_6 B \varepsilon)).
\end{aligned}$$

Thus, we get the desired decay of this norm if the logarithm in the latter exponent is uniformly negative. The latter can be assured by choosing $n(\varepsilon) \in (\kappa_1 \varepsilon^{-1/\beta}, \kappa_2 \varepsilon^{-1/\beta})$ with any choice of constants κ_1 and κ_2 such that $0 < \kappa_1 < \kappa_2 < (C_6 B)^{-1/\beta}$. Indeed, we then estimate

$$\|u^{\varepsilon,n(\varepsilon)} - u^\varepsilon\ ;\ H^1(\mathbb{T})\| \le C\varepsilon \exp[(\kappa_1 \ln(\kappa_2^\beta C_6 B))\varepsilon^{-1/\beta}],$$

which implies (17) by choosing $C_1 = C$ and $C_2 = -\kappa_1 \ln(\kappa_2^\beta C_6 B) > 0$. The theorem is proved. □

Remark 2 (on the proof of Theorem 1′). The proof of Theorem 1′ conceptually follows the above proof of Theorem 1. We briefly sketch the proof emphasizing only the most significant alterations to the above argument. First note that, although we still use the asymptotic series (12) for the approximation, its precise structure slightly differs from (14); namely, (12) is now represented in the following form:

$$u^{\varepsilon,n}(x) = \sum_{l=0}^{n+2} \varepsilon^l \sum_{s=0}^{[\frac{l}{2}]} \sum_{|k|=l-2s} N_k^{(s)}\left(\frac{x}{\varepsilon}\right) D_x^k V^{(n-l+2)}(x,\varepsilon). \tag{57}$$

For $N_k^{(s)}$, analogously to (25), (26), one deduces the recurrence relations

$$L_y^1 N_k^{(s)} = T_k^{(s)}(y) - h_k^{(s)} - N_k^{(s-1)}, \qquad \langle N_k^{(s)}\rangle = 0,\ s \ge 0, \tag{58}$$

assuming henceforth that $N_k^{(-1)} \equiv 0$. If $|k| \ge 2$, then one finds

$$T_k^{(s)} = \sum_{\substack{i,j=1,\dots,d \\ k'=k-e_i\ge 0}} \left((A_{ij} N_{k'}^{(s)})_{,j} + A_{ij} N_{k',j}^{(s)}\right) + \sum_{\substack{i,j=1,\dots,d \\ k''=k-e_i-e_j\ge 0}} A_{ij} N_{k''}^{(s)}, \tag{59}$$

and otherwise

$$|k| = 1, k = e_i: \quad T_k^{(s)} = (A_{ij} N_0^{(s)})_{,j} + A_{ij} N_{0,j}^{(s)}, \tag{60}$$

$$k = 0: \quad T_k^{(s)} = 0,\ \ N_0^{(0)} \equiv 1. \tag{61}$$

Further, in all the cases $h_k^{(s)} = \langle T_k^{(s)} \rangle$, except $h_0^{(1)} = -1$. Obviously one has $N_k^{(0)} = N_k$ (see (22)–(26)), and thus by induction in s one finds all $N_k^{(s)}$. Now let us introduce $\mathfrak{N}_q = \max_{|k|+2s=q} |N_k^{(s)}|_1$. Due to (59) we obviously have $|T_k^{(s)} - h_k^{(s)}|_{-1} \le C(\mathfrak{N}_{|k|+2s-1} + \mathfrak{N}_{|k|+2s-2})$. The basic elliptic estimate (38) for the problem (58) still holds and therefore implies that $\mathfrak{N}_q \le M_1 \mathfrak{N}_{q-1} + M_2 \mathfrak{N}_{q-2}$, which gives an exponential estimate of growth of $\mathfrak{N}_q$: with large enough $M_{\mathfrak{N}}$ for all $q \ge 0$

$$\mathfrak{N}_q \le (M_{\mathfrak{N}})^q. \tag{62}$$

Turning now to evaluation of $v_s(x)$, substituting the expansion (57) into (18) we observe that $V^{(\infty)}(x,\varepsilon)$ formally satisfies

$$\left(-A_{ij}^{\hom} D_{ij} + 1 + \sum_{l=3}^{\infty} \varepsilon^{l-2} \sum_{s=0}^{[\frac{l}{2}]} \sum_{|k|=l-2s} h_k^{(s)} D_x^k \right) V^{\infty}(x,\varepsilon) = f(x) \ \text{ in } \mathbb{R}^d; \tag{63}$$

therefore we have

$$-A_{ij}^{\hom} v_{s,ij} + v_s = f_s \ \text{ in } \mathbb{R}^d, \tag{64}$$

where

$$f_0 = f, \tag{65}$$

$$f_s = \sum_{l=3}^{s+2} \sum_{r=0}^{[\frac{l}{2}]} \sum_{|k|=l-2r} h_k^{(r)} D_x^k v_{s-l+2}, \ s \ge 1. \tag{66}$$

The latter differs from (33) only by the presence of lower order derivatives, and without any significant alteration one deduces an exponential estimate (37) in very much the same way as in Lemma 2. As a result, introducing $\mathfrak{V}^{(s,m)} = \|v_s;\ H^m(\mathbb{R}^d)\|$ with large enough $M_{\mathfrak{V}}$, we get by induction in s an estimate

$$\mathfrak{V}^{(s,m)} \le \sum_{k=0}^{s} (M_{\mathfrak{V}})^{k+1} \|f\ ;\ H^{m+k}(\mathbb{R}^d)\|. \tag{67}$$

The remainder estimate is bound to be of order ε^{n+1}, and with some minor technical alteration of the argument in section 4 one also gets

$$\|R^{\varepsilon,n}; H^{-1}(\mathbb{R}^d)\| \le C\varepsilon^{n+1} M^n \|f; H^{n+5+[d/2]}(\mathbb{R}^d)\|. \tag{68}$$

Finally, repeating the argument at the beginning of this section (omitting the consideration of the mean), employing appropriate modifications of the Poincaré inequality, ellipticity estimates, Sobolev embedding, etc. from (68), and the fact that $f \in \mathcal{G}^{\beta}(\mathbb{R}^d)$, one finally deduces Theorem 1′.

Remark 3. Note that as formulated the theorems admit some further sharpening: for example, one can replace $H^1(\mathbb{T})$ norm in (17) with $W^{1,p}(\mathbb{T})$ norm, where $p \in (2, p_0(d,\nu_0))$ with some $p_0(d,\nu_0) > 2$. Indeed, as can be seen from the structure of our argument, we select a functional space according to the fundamental ellipticity estimate (38), whereas the latter (38) can be refined in the case of bounded measurable coefficients and the right-hand side being the divergence of an L^{∞} (vector-)function (see, e.g., [22, Chapter 6]).

Remark 4. The proof of the theorems has been via a straightforward "bookkeeping" of the terms in the full two-scale asymptotic expansion (20)–(29). On the other hand, this is known to be related to the so-called spectral method in homogenization and closely related "Bloch approximation" approach; see, e.g., [24, 25, 26, 27, 28, 6]. There is no doubt that these spectral methods are capable of at least reestablishing the background results on the "homogenization in all orders" (i.e., approximations with arbitrary high order polynomial error bounds). An interesting further prospect would be to interpret the results presented here on *exponential* homogenization in terms of underlying analytic spectral properties of the Floquet–Bloch operator with periodic coefficients.

6. On the optimality of the exponential error (17): An example. In this section we demonstrate that the main exponential error bound (17) of Theorem 1 for $\|u^\varepsilon - u^{\varepsilon,n}\ ;\ H^1(\mathbb{T})\|$ is "optimal" for a particular class of one-dimensional examples. Namely, we show that by whatever choice of the truncation $n(\varepsilon)$ the error bound (17) cannot be improved apart from "optimizing" the choice of constants C_1 and C_2. This is done by proving an analogous exponential lower bound for the error; see (79). The latter is obtained by an optimal truncation $n(\varepsilon)$ of lower bounds derived for each n and ε, which in turn is observed to be delivered by $n(\varepsilon)$ within the range (16). In this sense the exponential error bound (17) is sharp.

We consider the following one-dimensional example. Consider elliptic problem

$$-\frac{d}{dx}\left(a(x/\varepsilon)\frac{d}{dx}u(x)\right) = f(x), \tag{69}$$

with one-periodic boundary conditions, $\langle u\rangle = \langle f\rangle = 0$, $\varepsilon = 1/N$, $N \in \mathbb{N}$, which is the one-dimensional version of the problem (4)–(5), with unique solution $u^\varepsilon(x)$. To be specific, let us consider[2]

$$a(y) = \frac{1}{3/2 - \cos(2\pi y - \pi/4)}. \tag{70}$$

We fix arbitrary $\beta \geq 1$ and assume the right-hand side f to be an infinitely differentiable real-valued 1-periodic function with real nonnegative Fourier coefficients f_k (hence $f_{-k} = f_k$), i.e.,

$$f(x) = \sum_{k\in\mathbb{Z},k\neq 0} f_k \exp(2i\pi kx) = 2\sum_{k=1}^{\infty} f_k \cos(2\pi kx), \quad f_k \geq 0. \tag{71}$$

We further assume that f satisfies the "converse" inequality to (6) determining β-Gevrey regular functions; i.e., there exists $b > 0$ such that

$$\|f\ ;\ H^l(\mathbb{T})\| \geq b^l (l!)^\beta \quad \text{for all } l \in \mathbb{N}. \tag{72}$$

In particular, for "sharp" β-Gevrey regular functions both (6) and (72) hold simultaneously:

$$b^l (l!)^\beta \leq \|f\ ;\ H^l(\mathbb{T})\| \leq B^l (l!)^\beta, \quad 0 < b \leq B < +\infty, \quad \text{for all } l \in \mathbb{N}. \tag{73}$$

[2]The analysis of this section can be generalized in a straightforward way to more general $a(y)$, for example, $a(y) = (a_0 - a_1\cos(2\pi y) - a_2 \sin(2\pi y))^{-1}$, $a_0 > 0$, $a_1 \neq 0$, $a_2 \neq 0$, $a_1^2 + a_2^2 < a_0^2$. We do not pursue maximal generality to avoid unnecessary further algebraic complications.

A sufficient condition for f to satisfy (73) is for its Fourier coefficients f_k to decay exponentially with the "rate" $|k|^{1/\beta}$, i.e.,

$$A_1 \exp(-B_1|k|^{1/\beta}) \le f_k \le A_2 \exp(-B_2|k|^{1/\beta}) \tag{74}$$

with positive A_1, A_2, B_1, and B_2.

To see that such f satisfies (73) one can first apply the Plancherel theorem to definition (7), implying

$$\begin{aligned}(2\pi)^{2l} A_1 \sum_{k\in\mathbb{N}} \exp(-2B_1 k^{1/\beta}) k^{2l} \le (2\pi)^{2l} \sum_{k\in\mathbb{N}} k^{2l} f_k^2 \le \|f;\ H^l(\mathbb{T})\|^2 \\ \le 4(2\pi)^{2l} \sum_{k\in\mathbb{N}} k^{2l} f_k^2 \le 4(2\pi)^{2l} A_2 \sum_{k\in\mathbb{N}} \exp(-2B_2 k^{1/\beta}) k^{2l}. \end{aligned} \tag{75}$$

Then one can notice that the sums in (75) can be further bounded both from above and from below as follows: there exist l-independent positive constants D_1 and D_2 such that

$$\begin{aligned} D_1 \beta (2B_1)^{-\beta(2l+1)} \Gamma((2l+1)\beta) = D_1 \int_0^\infty \exp(-2B_1 s^{1/\beta}) s^{2l} ds \\ \le \sum_{k\in\mathbb{N}} \exp(-2B_1 k^{1/\beta}) k^{2l} \le \sum_{k\in\mathbb{N}} \exp(-2B_2 k^{1/\beta}) k^{2l} \\ \le D_2 l \int_0^\infty \exp(-2B_2 s^{1/\beta}) s^{2l} ds = D_2 l \beta (2B_2)^{-\beta(2l+1)} \Gamma((2l+1)\beta). \end{aligned} \tag{76}$$

(A way to establish (76) is by noticing that the series, asymptotically for large l, coincides to the main order with the integral.) Finally, by the application of the Stirling formula (11) we obtain $D_3^{2l}(l!)^{2\beta} \le \Gamma((2l+1)\beta) \le D_4^{2l}(l!)^{2\beta}$ with l-independent D_3, D_4, which implies (73).

An example of a function f satisfying (71) and (73) is

$$f(x) = \sum_{k\in\mathbb{Z}, k\neq 0} \exp(-|k|^{1/\beta}) \exp(2i\pi k x). \tag{77}$$

In particular, for $\beta = 1$

$$f(x) = 2\,\mathrm{Re} \sum_{k=1}^{\infty} \exp(k(2i\pi x - 1)) = \frac{2e\cos(2\pi x) - 2}{(e^2+1) - 2e\cos(2\pi x)} \tag{78}$$

is clearly analytic, with poles at $x = \pm i/(2\pi) + n$, $n \in \mathbb{Z}$.

We formulate the following optimality theorem for the above one-dimensional case.

THEOREM 5. *For any f satisfying* (74) *with $B_1 = B_2$ there exist positive constants $\tilde{C}_1$ and $\tilde{C}_2$ such that the following lower error bound for the exact solution u^ε of the problem* (69)–(70) *and its asymptotic approximation $u^{\varepsilon,n}$ holds for any $n \in \mathbb{N}$ and any $\varepsilon = 1/N$, $N \in \mathbb{N}$:*

$$\|u^\varepsilon - u^{\varepsilon,n}\ ;\ H^1(\mathbb{T})\| \ge \tilde{C}_1 \exp(-\tilde{C}_2 \varepsilon^{-\frac{1}{\beta}}). \tag{79}$$

Proof. In the one-dimensional case, the general recurrence relations (22)–(27), (30)–(33) for the correctors N_k, the "homogenized coefficients" h_k, and "slowly varying" parts v_s specialize to simple ODEs (see, e.g., [23, section 1F]), which can be

solved explicitly.[3] In particular, the equations (25)–(26) for the "main corrector" N_1 specialize to

$$\frac{d}{dy}N_1(y) = \frac{a^{-1}(y)}{\langle a^{-1}\rangle} - 1 = \frac{2}{3}\left(\frac{3}{2} - \cos\left(2\pi y - \frac{\pi}{4}\right)\right) - 1 = -\frac{2}{3}\cos\left(2\pi y - \frac{\pi}{4}\right), \tag{80}$$

implying $N_1(y) = -(3\pi)^{-1}\sin(2\pi y - \pi/4)$. All the higher order correctors N_k, $k \geq 2$, have a similar form, due to the recurrence relations

$$\frac{d}{dy}N_k = -N_{k-1} \tag{81}$$

(the latter also immediately follows by direct substitution of (20) into (69)). As a result,

$$\begin{aligned} N_{2m}(y) &= (-1)^m \frac{\cos(2\pi y - \pi/4)}{3\pi(2\pi)^{2m-1}} = (-1)^m 2^{1/2}\frac{\cos(2\pi y) + \sin(2\pi y)}{3(2\pi)^{2m}}, \quad m \geq 1, \\ N_{2m+1}(y) &= (-1)^{m+1}\frac{\sin(2\pi y - \pi/4)}{3\pi(2\pi)^{2m}} = (-1)^m 2^{1/2}\frac{\cos(2\pi y) - \sin(2\pi y)}{3(2\pi)^{2m+1}}, \quad m \geq 0. \end{aligned} \tag{82}$$

Further, v_0 is given by homogenized equation (30) ($s = 0$) specializing in the one-dimensional case to

$$-h_2\,\frac{d^2}{dx^2}v_0 = f, \tag{83}$$

where $h_2 = A^{\mathrm{hom}} = \langle a^{-1}\rangle^{-1} = 2/3$. Furthermore, in the one-dimensional case $h_k = 0$ for all $k \geq 3$ via a straightforward analysis of the recurrent relations (23)–(27) (see, e.g., [23, section 1F]). The latter immediately implies via (30) and (33) that $v_k = 0$ for all $k \geq 1$. Taking the above into account specializes the remainder term (45) in Lemma 3 to

$$\begin{aligned} R^{\varepsilon,n}(x) &= -\varepsilon^{n+1}\left(\left(aN'_{n+2} + a'N_{n+2}\right)\left(\frac{x}{\varepsilon}\right)D^{n+3}v_0(x)\right. \\ &\qquad \left. +\varepsilon a\left(\frac{x}{\varepsilon}\right)N_{n+2}\left(\frac{x}{\varepsilon}\right)D^{n+4}v_0(x)\right) \\ &= -\frac{d}{dx}\left(\varepsilon^{n+2}a\left(\frac{x}{\varepsilon}\right)N_{n+2}\left(\frac{x}{\varepsilon}\right)D^{n+3}v_0(x)\right) = -\frac{d}{dx}\Phi^{\varepsilon,n}(x), \end{aligned} \tag{84}$$

where

$$\Phi^{\varepsilon,n}(x) := \varepsilon^{n+2}a\left(\frac{x}{\varepsilon}\right)N_{n+2}\left(\frac{x}{\varepsilon}\right)D^{n+3}v_0(x). \tag{85}$$

Employing in the above the explicit solutions (82) for N_k and (83) for v_0, we arrive at

$$\begin{aligned} \Phi^{\varepsilon,n}(x) &= \frac{(-1)^{n/2}\varepsilon^{n+2}(\cos(2\pi x/\varepsilon) + \sin(2\pi x/\varepsilon))}{2^{1/2}(2\pi)^{n+2}\,[3/2 - \cos(2\pi x/\varepsilon - \pi/4)]}D^{n+1}f(x), \ n = 2m,\ m \geq 1, \\ \Phi^{\varepsilon,n}(x) &= \frac{(-1)^{(n+1)/2}\varepsilon^{n+2}(\cos(2\pi x/\varepsilon) - \sin(2\pi x/\varepsilon))}{2^{1/2}(2\pi)^{n+2}\,[3/2 - \cos(2\pi x/\varepsilon - \pi/4)]}D^{n+1}f(x), \ n = 2m+1,\ m \geq 0. \end{aligned} \tag{86}$$

(87)

[3] We remark that the present one-dimensional case is integrable, and an alternative but related approach for analyzing the error term in the asymptotics is from the exact solution; see Remark 5.

From (84), by the definition of the H^{-1} norm, we have

$$\|R^{\varepsilon,n};\ H^{-1}\|^2 \geq C\|\Phi^{\varepsilon,n} - \langle\Phi^{\varepsilon,n}\rangle;\ L^2\|^2 = C\left(\|\Phi^{\varepsilon,n};\ L^2\|^2 - \left|\int_0^1 \Phi^{\varepsilon,n}(x)dx\right|^2\right) \tag{88}$$

(recalling that C denotes constants whose precise value is insignificant).

With the aim of further bounding (88) from below, we introduce for any given $\varepsilon = 1/N$ and n functions $h_N(x)$ as follows:

$$\begin{aligned} h_N(x) = h(Nx) &= \frac{\cos(2\pi Nx) + \sin(2\pi Nx)}{[3/2 - \cos(2\pi Nx - \pi/4)]},\ N = 2m, \\ h_N(x) = h(Nx) &= \frac{\cos(2\pi Nx) - \sin(2\pi Nx)}{[3/2 - \cos(2\pi Nx - \pi/4)]},\ N = 2m+1. \end{aligned} \tag{89}$$

We prove the following lemma.

LEMMA 6. *There exists a constant C such that for any f satisfying* (71) *and for all n and N*

$$\|h_N D^{n+1} f;\ L^2\|^2 \geq C\|D^{n+1} f;\ L^2\|^2. \tag{90}$$

Proof. Choosing first n to be even, $n = 2m$, notice that

$$\begin{aligned} \|h_N(x)D^{n+1}f(x);\ L^2\|^2 &:= \int_0^1 \frac{(\cos(2\pi Nx) + \sin(2\pi Nx))^2}{[3/2 - \cos(2\pi Nx - \pi/4)]^2} \left(D^{n+1}f\right)^2 dx \\ &\geq \frac{4}{25}\int_0^1 (\cos(2\pi Nx) + \sin(2\pi Nx))^2 \left(D^{n+1}f\right)^2 dx \\ &= \frac{4}{25}\|\left(\cos(2\pi Nx) + \sin(2\pi Nx)\right) D^{n+1}f(x);\ L^2\|^2. \end{aligned}$$

We next notice that for n even and for f given by (71) $D^{n+1}f$ is represented by a sine Fourier series, implying that $\cos(2\pi Nx)D^{n+1}f(x)$ and $\sin(2\pi Nx)D^{n+1}f(x)$ are orthogonal in $L^2(0,1)$ and hence

$$\|\left(\cos(2\pi Nx) + \sin(2\pi Nx)\right) D^{n+1}f(x);\ L^2\|^2 \geq \|\cos(2\pi Nx)D^{n+1}f(x);\ L^2\|^2.$$

Further,

$$\begin{aligned} g_{Nn}(x) := \cos(2\pi Nx)D^{n+1}f(x) &= \frac{1}{2}\sum_{k\in\mathbb{Z}} \left(e^{2i\pi Nx} + e^{-2i\pi Nx}\right)(2i\pi k)^{n+1} f_k e^{2i\pi kx} \\ &= \frac{1}{2}\sum_{m\in\mathbb{Z}} e^{2i\pi mx}\left[(i2\pi(m-N))^{n+1} f_{m-N} + (i2\pi(m+N))^{n+1} f_{m+N}\right]. \end{aligned}$$

Hence, applying the Plancherel theorem,

$$\begin{aligned}\|g_{Nn}(x);\ L^2\|^2 &= \frac{(2\pi)^{2n+2}}{4}\sum_{m\in\mathbb{Z}}\left[(m-N)^{n+1}f_{m-N}+(m+N)^{n+1}f_{m+N}\right]^2\\ &\geq \frac{(2\pi)^{2n+2}}{4}\sum_{m=N+1}^{\infty}\left[(m-N)^{n+1}f_{m-N}+(m+N)^{n+1}f_{m+N}\right]^2\\ &\geq \frac{(2\pi)^{2n+2}}{4}\sum_{m=N+1}^{\infty}\left((m-N)^{n+1}f_{m-N}\right)^2 = \frac{(2\pi)^{2n+2}}{4}\sum_{m=1}^{\infty}\left(m^{n+1}f_m\right)^2\\ &= \frac{(2\pi)^{2n+2}}{8}\sum_{m\in\mathbb{Z}}\left(m^{n+1}f_m\right)^2 = \frac{1}{8}\|D^{n+1}f(x);\ L^2\|^2.\end{aligned}$$

In the latter we have used the nonnegativity of Fourier coefficients f_k, their symmetry ($f_{-k}=f_k$), and the fact that $\langle f\rangle = 0$ (hence $f_0=0$).

The above proves the lemma for even n. The proof for odd n is fully analogous, with the sign alteration between the "sine" and "cosine" terms, then noticing that $D^{n+1}f$ is represented by a cosine Fourier series, and then using the orthogonality and neglecting the sine term as before. ∎

We next aim at showing that, at least for sufficiently large n, the last term in the right-hand side of (88) can be bounded from the above as in (90) but with a smaller constant.

LEMMA 7. *For any f satisfying* (74) *with $B_1 = B_2$ there exists $n_0 > 0$ such that for all $n > n_0$ and all N*

$$\left\langle h_N(x)D^{n+1}f(x)\right\rangle^2 := \left|\int_0^1 h_N(x)D^{n+1}f(x)dx\right|^2 \leq \frac{1}{2}C\|D^{n+1}f(x);\ L^2\|^2, \tag{91}$$

where C is same as in Lemma 6.

Proof. The Fourier series of $h_N(x)$ has the form

$$h_N(x) = \sum_{\ell\in\mathbb{Z}} h_\ell \exp(i2\pi N\ell x), \tag{92}$$

with $\{h_\ell\}$ being two possible sets of (rapidly decaying) Fourier coefficients (for N even and odd, according to (89)), independent of N for all $\ell\in\mathbb{Z}$. Then,

$$\begin{aligned}\left\langle h_N(x)D^{n+1}f(x)\right\rangle^2 &= \left|\sum_{\ell\in\mathbb{Z}} h_\ell(2\pi N\ell)^{n+1}f_{-N\ell}\right|^2\\ &\leq (2\pi)^{2n+2}\max_{k\in\mathbb{N}}\left(k^{2n+2}f_k^2\right)\left(\sum_{\ell\in\mathbb{Z}}|h_\ell|\right)^2 \leq H(2\pi)^{2n+2}\max_{t\geq 0}\phi_n(t)\\ &\leq H(2\pi)^{2n+2}\left(\frac{\beta}{eB_1}(n+1)\right)^{2\beta(n+1)}, \qquad (93)\end{aligned}$$

where

$$\phi_n(t) := t^{2n+2}\exp\left(-2B_1t^{1/\beta}\right) \tag{94}$$

and H is independent of N and n.

On the other hand, we have

$$\begin{aligned}\|D^{n+1}f;\ L^2\|^2 &= (2\pi)^{2n+2}\sum_{k\in\mathbb{Z}} k^{2n+2}f_k^2 \geq A_1(2\pi)^{2n+2}\sum_{k\in\mathbb{Z}}\phi_n(|k|)\\ &\geq 2A_1(2\pi)^{2n+2}\left(\int_0^\infty \phi_n(t)dt - \max_{t\geq 0}\phi_n(t)\right)\\ &= 2A_1(2\pi)^{2n+2}\left(\int_0^\infty \phi_n(t)dt - \left(\frac{\beta}{eB_1}(n+1)\right)^{2\beta(n+1)}\right) \qquad (95)\end{aligned}$$

using the fact that $\phi_n(t)$, $t\geq 0$, has a single maximum for any n. Further, we estimate

$$\int_0^\infty \phi_n(t)dt = \beta(2B_1)^{-\beta(2n+3)}\Gamma((2n+3)\beta) \geq C\left(\frac{\beta}{eB_1}(n+1)\right)^{2\beta(n+1)} n^{\beta-1/2} \qquad (96)$$

with some $C>0$, having used

$$\Gamma((2n+3)\beta) \geq c\left(\frac{(2n+3)\beta}{e}\right)^{(2n+3)\beta-1/2} \qquad (97)$$

with some $c>0$ which is a direct implication of the Stirling formula (11), and then performing further straightforward manipulations.

Comparing finally (93) with (95) and (96), we conclude that

$$\left\langle h_N(x)D^{n+1}f(x)\right\rangle^2 \leq cn^{1/2-\beta}\|D^{n+1}f(x);\ L^2\|^2$$

with some $c>0$, and hence (91) holds with appropriate choice of n_0. □

Now we complete the proof of Theorem 5. Let C be a constant from Lemma 6 and let n_0 be as in Lemma 7. Denote

$$\begin{aligned}G_{nN} &:= \|h_N(x)D^{n+1}f(x);\ L^2\|^2 - \left\langle h_N(x)D^{n+1}f(x)\right\rangle^2\\ &= \int_0^1 \left(h_N(x)D^{n+1}f(x) - \left\langle h_N(x)D^{n+1}f(x)\right\rangle\right)^2 dx > 0. \qquad (98)\end{aligned}$$

(G_{nN} is strictly positive for any n and N since $h_N(x)D^{n+1}f(x)$ is not a constant: h_N vanishes at some points, but $D^{n+1}f(x)$ is clearly not identically zero.)

By Lemmas 6 and 7, for any N and for any $n>n_0$

$$G_{nN} \geq \frac{1}{2}C\|D^{n+1}f(x);\ L^2\|^2. \qquad (99)$$

Further, for any $0<n\leq n_0$

$$\lim_{N\to\infty} G_{nN} = C_1\|D^{n+1}f(x);\ L^2\|^2, \qquad (100)$$

where

$$C_1 := \left\langle \left(h_N - \langle h_N\rangle\right)^2\right\rangle > 0$$

is N-independent positive constant by (89). (A standard way to establish (100) is to subtract from h_N^2 and h_N in (98) their means, represent the resulting zero-mean

periodic functions as derivatives of other periodic functions, and then integrate by parts.) It follows from (100) that there exists $C_2 > 0$ such that

$$G_{nN} \geq C_2 \| D^{n+1} f(x);\ L^2 \|^2$$

for any N and for any $n \leq n_0$. Combining the latter with (99) implies that

$$G_{nN} \geq C_3 \| D^{n+1} f(x);\ L^2 \|^2 \tag{101}$$

for all n and N with $C_3 = \min(C/2, C_2)$.

Next, from (86)–(87)

$$\| \Phi^{\varepsilon,n}(x) - \langle \Phi^{\varepsilon,n} \rangle ;\ L^2 \|^2 = \frac{1}{4} \left(\frac{\varepsilon}{2\pi} \right)^{2n+4} G_{nN} \geq C_4 \left(\frac{\varepsilon}{2\pi} \right)^{2n+4} \| D^{n+1} f(x);\ L^2 \|^2. \tag{102}$$

Using again the lower bounds in (95)–(97) implies

$$\| D^{n+1} f(x);\ L^2 \|^2 \geq C_5^n n^{2\beta n}$$

with some $C_5 > 0$, which combined with (102) yields

$$\| \Phi^{\varepsilon,n}(x) - \langle \Phi^{\varepsilon,n} \rangle ;\ L^2 \| \geq C_6^n \varepsilon^{n+2} n^{\beta n}. \tag{103}$$

Now, by uniform continuity of L^ε as an operator from H^1 to H^{-1}, we have

$$\| u^{\varepsilon,n} - u^\varepsilon \ ;\ H^1(\mathbb{T}) \| \geq \| u^{\varepsilon,n} - \langle u^{\varepsilon,n} \rangle - u^\varepsilon \ ;\ H^1(\mathbb{T}) \| \geq C \| R^{\varepsilon,n}(x);\ H^{-1} \|. \tag{104}$$

Combining this with (88) and (103) implies

$$\| u^{\varepsilon,n} - u^\varepsilon \ ;\ H^1(\mathbb{T}) \| \geq C_7^n \varepsilon^{n+2} n^{\beta n}. \tag{105}$$

We can now "optimize" the lower bound (105) for any fixed small ε by choosing $n = n(\varepsilon)$ so that the right-hand side of (105) is minimized:

$$\| u^{\varepsilon,n} - u^\varepsilon \ ;\ H^1(\mathbb{T}) \| \geq \varepsilon^2 \min_{t \geq 1} \left[(C_7 \varepsilon)^t t^{\beta t} \right]. \tag{106}$$

The latter minimum is attained at

$$t = e^{-1} (C_7 \varepsilon)^{-1/\beta}; \tag{107}$$

substituting this back into (106), we finally obtain a lower bound of the form

$$\| u^{\varepsilon,n} - u^\varepsilon \ ;\ H^1(\mathbb{T}) \| \geq \varepsilon^2 \exp\left[-\beta e^{-1} (C_7 \varepsilon)^{-1/\beta} \right]. \tag{108}$$

Finally, since obviously there exists such a positive constant $\tilde{C}_1$ such that $\varepsilon^2 > \tilde{C}_1 \exp(-\varepsilon^{1/\beta})$ for any $0 < \varepsilon \leq 1$, (108) implies (79) for any $0 < \varepsilon \leq 1$, for example, with the above $\tilde{C}_1$ and $\tilde{C}_2 = \beta e^{-1} C_7^{-1/\beta} + 1$. The theorem is proved. □

We conclude from Theorem 5 that up to the choice of the constants $C_1, C_2 > 0$, the main error estimate (17) in Theorem 1 is sharp, at least for the above one-dimensional case. Note also that the above lower bound (79) was obtained by "optimizing" the lower bound (105) for a given small ε by choosing $n = n(\varepsilon)$ in the range given by

(107), which is consistent with (16) for the upper bound. In this sense, the range of truncation given by (16) is also "optimal."

Remark 5. Notice that the present one-dimensional case is "integrable" and that Theorem 5 could have been alternatively derived from the exact solution of (69):

$$u^\varepsilon(x) = \int_0^x (F(s) - A^\varepsilon)\, a^{-1}\left(\frac{s}{\varepsilon}\right) ds \; - \; B^\varepsilon, \tag{109}$$

where $F(x) := -\int_0^x f(t)dt$ is periodic, $A^\varepsilon := \langle F(\cdot)a^{-1}(\cdot/\varepsilon)\rangle\langle a^{-1}\rangle^{-1}$, and $B^\varepsilon := \int_0^1 (F(s)-A^\varepsilon)a^{-1}(s/\varepsilon)(1-s)ds$. One then employs $a^{-1}(x/\varepsilon) = \langle a^{-1}\rangle + \varepsilon\langle a^{-1}\rangle \frac{d}{dx} N_1(x/\varepsilon)$ (cf. (80)), in (109) and integrates by parts. Then employing iteratively (81) and integrating by parts n times is expected to explicitly reproduce $u^{\varepsilon,n}$, with the rest being the "error term." The latter would then still have to be analyzed in a fashion similar if not identical to that in the above proofs (the details are omitted).

Remark 6. The same arguments (Lemmas 6 and 7) can be used to obtain lower bounds of finite order in ε if $f \in H^M$ but $f \notin H^{M+1}$ for some $M \in \mathbb{N}$. Namely, on one hand, only a finite number of terms in the asymptotic expansion can be constructed. On the other hand, for each such n (from a finite set) a lower bound of the form (105) holds with appropriate choice of C_7. Optimizing finally with respect to the final set of lower bounds (105), one arrives at an unimprovable polynomial lower bound. On the other hand, if $f \in C^\infty(\mathbb{T})$ and is not from any Gevrey-type class with no other control on the growth of its H^l norms for large l (equivalently, on $\|D^{n+1}f; L^2\|$ for large n), there is no control on the "coefficients" multiplying ε^{n+2} in the error bounds for large n (cf. (104), (86)–(88)), which does not allow us to improve the homogenization in all orders any further. This indicates the importance of the Gevrey extreme regularity of f for the exponential lower bounds.

Remark 7. For analytic $f(x)$ Theorems 1 and 5, i.e., the exponential upper bound (17) and lower bound (79), respectively, both hold with $\beta = 1$. The "rate" of the exponential decay is then determined by the values of the constants C_2 and $\tilde{C}_2$ in (17) and (79). By tracing back the proofs of both theorems, one observes that these constants are dependent on the rate of the exponential decay of the Fourier coefficients of f, i.e., by the constants c_2 in (9) and B_1 in (74). On the other hand, for analytic f, the latter constants are directly related to the "width" of analytic continuation of $f(x)$ into the complex plane off the real axis, i.e., the absolute value of the imaginary part of the "first singularity"; for example, f in (78) has a pole in $x = \pm i/(2\pi)$, corresponding to $B_1 = 1$. In this sense one could argue that the rate of exponential error bound for analytic f is determined by the nearest singularity in the analytic continuation.

Acknowledgment. V. P. Smyshlyaev acknowledges stimulating discussions with Dr. Claudia Wulff of the University of Surrey, Surrey, UK.

REFERENCES

[1] A. Bensoussan, J.-L. Lions, and G. C. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, 1978.

[2] E. Sanchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Phys. 127, Springer-Verlag, Berlin, New York, 1980.

[3] N. S. Bakhvalov and G. P. Panasenko, *Homogenization: Averaging Processes in Periodic Media*, Nauka, Moscow, 1984 (in Russian); Math. Appl. (Soviet Series) 36, Kluwer Academic Publishers, Dordrecht, Boston, London, 1989 (in English).

[4] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

[5] J.-L. Lions, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Kexue Chubanshe (Science Press), Beijing, Gordon & Breach, New York, 1981.
[6] M. Sh. Birman and T. A. Suslina, *Periodic second-order differential operators. Threshhold properties and homogenization*, St. Petersburg Math. J., 15 (2004), pp. 639–714.
[7] V. V. Zhikov, *On operator estimates in homogenization theory*, Dokl. Akad. Nauk, 403 (2005), pp. 305–308 (in Russian).
[8] M. Neuss-Radu, *A result on the decay of the boundary layers in the homogenization theory*, Asymptot. Anal., 23 (2000), pp. 313–328.
[9] M. Neuss-Radu, *The failure of uniform exponential decay for boundary layers*, in Multiscale Problems in Science and Technology, Challenges to Mathematical Analysis and Perspectives (Dubrovnik, 2000), N. Antonic, C. J. van Duijn, W. Jäger, and A. Mikelic, eds., Springer-Verlag, 2002, pp. 243–250.
[10] V. P. Smyshlyaev and K. D. Cherednichenko, *On rigorous derivation of strain gradient effects in the overall behaviour of periodic heterogeneous media*, J. Mech. Phys. Solids, 48 (2000), pp. 1325–1357.
[11] K. D. Cherednichenko and V. P. Smyshlyaev, *On full two-scale expansion of the solutions of nonlinear periodic rapidly oscillating problems and higher-order homogenized variational problems*, Arch. Ration. Mech. Anal., 174 (2004), pp. 385-442.
[12] S. J. Chapman and D. B. Mortimer, *Exponential asymptotics and Stokes lines in a partial differential equation*, in Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 2385–2421.
[13] A. Neishtadt, *On the separation of motions in systems with rapidly rotating phase*, J. Appl. Math. Mech., 48 (1984), pp. 134-139.
[14] K. Matthies, *Time-averaging under fast periodic forcing of parabolic partial differential equations: Exponential estimates*, J. Differential Equations, 174 (2001), pp. 133-180.
[15] K. Matthies, *Homogenization of exponential order for elliptic systems in infinite cylinders*, Asymptot. Anal., 43 (2005), pp. 205-232.
[16] A. M. Matache, I. Babuška, and C. Schwab, *Generalized p-FEM in homogenization*, Numer. Math., 86 (2000), pp. 319–375.
[17] G. Lebeau, *Régularité Gevrey* 3 *pour la diffraction*, Comm. Partial Differential Equations, 9 (1984), pp. 1437–1494.
[18] C. Bardos, G. Lebeau, and J. Rauch, *Scattering frequencies and Gevrey* 3 *singularities*, Invent. Math., 90 (1987), pp. 77–114.
[19] C. Foias and R. Temam, *Gevrey class regularity for the solution of the Navier-Stokes equations*, J. Funct. Anal., 87 (1989), pp. 359–369.
[20] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. III, Springer-Verlag, New York, Heidelberg, 1973.
[21] L. Hörmander, *The Analysis of Linear Partial Differential Operators.* I. *Distribution Theory and Fourier Analysis*, Springer-Verlag, Berlin, 2003.
[22] E. Giusti, *Direct Methods in the Calculus of Variations*, World Scientific, River Edge, NJ, 2003.
[23] K. D. Cherednichenko, *Higher-order and Non-local Effects in Homogenisation of Periodic Media*, Ph.D. thesis, University of Bath, Bath, UK, 2001.
[24] A. Yu. Belyaev, *Compression waves in a fluid with gas bubbles*, J. Appl. Math. Mech., 52 (1988), pp. 344–348.
[25] V. V. Zhikov, *Spectral approach to asymptotic diffusion problems*, Differential Equations, 25 (1989), pp. 33–39.
[26] G. Allaire and C. Conca, *Bloch wave homogenization and spectral asymptotic analysis*, J. Math. Pures Appl. (9), 77 (1998), pp. 153–208.
[27] C. Conca, R. Orive, and M. Vanninathan, *Bloch approximation in homogenization and applications*, SIAM J. Math. Anal., 33 (2002), pp. 1166–1198.
[28] V. V. Zhikov, *On the spectral method in homogenization theory*, Tr. Mat. Inst. Steklova, 250 (2005), pp. 95–194, (in Russian); Proc. Steklov Inst. Math., 250 (2005), pp. 85–94 (in English).
[29] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1588–1612

MAGNETIC FIELD-INDUCED INSTABILITIES IN LIQUID CRYSTALS*

FANGHUA LIN† AND XING-BIN PAN‡

Abstract. We use the Landau–de Gennes model to investigate the magnetic field-induced instabilities in liquid crystals. In particular, we examine the change of weak and strong stabilities in the pure smectic states and in the pure nematic states. Motivated by de Gennes' discovery on the analogies between liquid crystals and superconductors, we introduce critical magnetic fields $H_{\rm s}$ and $H_{\rm sh}$. The pure smectic states lose their global minimality (strong stability) at $H_{\rm s}$ and lose their local minimality (weak stability) at $H_{\rm sh}$. We also examine the change of stability in the pure nematic states. We show in the case of equal elastic coefficients that a liquid crystal in a sufficiently strong magnetic field will not be in a pure nematic state, which exhibits a significant difference between the Landau–de Gennes model for liquid crystals and the Ginzburg–Landau model for superconductivity.

Key words. liquid crystal, phase transition, critical magnetic field, Landau–de Gennes model

AMS subject classifications. 82D30, 35J55, 35Q55

DOI. 10.1137/050638643

1. Introduction.

A liquid crystal configuration will change its stability under applied electric or magnetic fields. The classical mathematical description of transitions of stabilities in nematic liquid crystals under applied magnetic fields is given by introducing a magnetic energy density $-\chi(\mathbf{H}\cdot\mathbf{n})^2$ to the classical Oseen–Frank energy density $F_N(\mathbf{n},\nabla\mathbf{n})$ of nematic liquid crystals to obtain a modified energy functional (one may call it *the Oseen–Frank model with magnetic effect*):

$$\int_\Omega \{F_N(\mathbf{n},\nabla\mathbf{n}) - \chi(\mathbf{H}\cdot\mathbf{n})^2\}dx.$$

Here $\mathbf{n}:\bar\Omega\to\mathbb{S}^2$ denotes the *director field* of the nematic liquid crystals, $\mathbf{H}$ is an applied magnetic field, and χ is a positive parameter; see [E], [dGP, p. 287], [K], [HKL], and [CL]. It was widely accepted that adding the term $\chi(\mathbf{H}\cdot\mathbf{n})^2$ into the Oseen–Frank energy is only a lower order perturbation from an analysis point of view and that one does not expect substantially new phenomena. The physical phenomena, however, say otherwise (see [LL, section 2.2]).

In this paper we consider the influence of magnetic fields on the states of liquid crystals and on phase transitions from a nematic state to a smectic state. We shall work in the framework of the Landau–de Gennes theory [dGP]. According to this theory, the states and phase transitions of nematic to smectic liquid crystals can be described by the minimizers $(\psi,\mathbf{n})$ of Landau–de Gennes energy, where ψ is a complex-valued function called *order parameter* and $\mathbf{n}$ is the director field. This model was proposed by de Gennes when he discovered the analogies of liquid crystals

*Received by the editors August 22, 2005; accepted for publication July 26, 2006; published electronically January 26, 2007.

http://www.siam.org/journals/sima/38-5/63864.html

†Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (linf@cims.nyu.edu). This author was partially supported by NSF grant DMS 0201443.

‡Department of Mathematics, East China Normal University, Shanghai 200062, People's Republic of China (xbpan@math.ecnu.edu.cn). This author was partially supported by National Natural Science Foundation of China grant 10471125, the Science Foundation of the Ministry of Education of China, and Shanghai Pujiang Program 05PJ14039.

and superconductors; see [dG], [dGP]. To understand the effects of magnetic fields on liquid crystals, including nematics and smectics phase transitions, we follow the earlier work of Ericksen [E] by introducing the term $-\chi(\mathbf{H}\cdot\mathbf{n})^2$ to the Landau–de Gennes model to get a modified energy functional (which one may refer to as the *Landau–de Gennes model with magnetic effect*). We then examine the behavior of minimizers of the functional when the applied magnetic fields vary. Throughout this paper we consider the strong anchoring condition, which is simply the Dirichlet boundary condition on the director fields:

$$\mathbf{n} = \mathbf{u}_0 \quad \text{on } \partial\Omega,$$

where $\mathbf{u}_0 \in C^1(\partial\Omega, \mathbb{S}^2)$. As in [HKL], we can drop the divergence term (surface energy) in the Oseen–Frank energy. For simplicity, we also assume, as in [P2], [P4], that the bend and twist coefficients in the Oseen–Frank energy functional are equal: $K_2 = K_3$. This leads to a simplified form of the Landau–de Gennes energy functional with an applied magnetic field

(1.1)
$$\mathcal{E}[\psi,\mathbf{n}] = \int_\Omega \left\{ |\nabla_{q\mathbf{n}}\psi|^2 + \frac{\kappa^2}{2}(1-|\psi|^2)^2 + K_1|\operatorname{div}\mathbf{n}|^2 + K_2|\operatorname{curl}\mathbf{n}|^2 - \chi(\mathbf{H}\cdot\mathbf{n})^2 \right\} dx,$$

where κ, K_1, K_2, and χ are positive constants and q is a real number. Without loss of generality we shall always assume that $q \geq 0$. We may write the functional as

$$\mathcal{E}[\psi,\mathbf{n}] = \mathcal{G}[\psi,\mathbf{n}] + \mathcal{F}[\mathbf{n}] - \int_\Omega \chi(\mathbf{H}\cdot\mathbf{n})^2 dx,$$

where

$$\mathcal{F}[\mathbf{n}] = \int_\Omega \{K_1|\operatorname{div}\mathbf{n}|^2 + K_2|\operatorname{curl}\mathbf{n}|^2\} dx$$

is the simplified Oseen–Frank energy for nematics and

$$\mathcal{G}[\psi,\mathbf{n}] = \int_\Omega \left\{ |\nabla_{q\mathbf{n}}\psi|^2 + \frac{\kappa^2}{2}(1-|\psi|^2)^2 \right\} dx$$

is the Ginzburg–Landau energy of smectics. The functional $\mathcal{E}$ is well defined in the space

$$W^{1,2}(\Omega,\mathbb{C}) \times W^{1,2}(\Omega,\mathbb{S}^2,\mathbf{u}_0),$$

where $W^{1,2}(\Omega,\mathbb{C})$ is the usual Sobolev space of complex valued functions and

$$W^{1,2}(\Omega,\mathbb{S}^2,\mathbf{u}_0) = \{\mathbf{u}\in W^{1,2}(\Omega,\mathbb{R}^3) :\ |\mathbf{u}(x)| = 1 \text{ a.e. in } \Omega,\ \mathbf{u} = \mathbf{u}_0 \text{ on } \partial\Omega\}.$$

Analysis shows that the minimizers of $\mathcal{E}$ undergo complicated changes when the applied fields vary, indicating that magnetic fields have a very important influence on phase transitions of liquid crystals. To make our presentations more explicit so that one can compare our analysis with numerical computations as well as physical experiments, in this paper (except in section 2) we shall restrict ourselves to a simple situation where the applied field $\mathbf{H}$ and the boundary data $\mathbf{u}_0$ are a pair of constant vectors orthogonal to each other; i.e., we assume that

(1.2)
$$\mathbf{H} = \sigma\mathbf{h}, \qquad \mathbf{u}_0 = \mathbf{e},$$

where σ is a positive number measuring the strength of the applied magnetic field and $\mathbf{h}$ and $\mathbf{e}$ are unit vectors such that $\mathbf{h} \cdot \mathbf{e} = 0$. The energy functional can then be rewritten as

$$\mathcal{E}[\psi, \mathbf{n}] = \mathcal{G}[\psi, \mathbf{n}] + \mathcal{F}[\mathbf{n}] - \chi\sigma^2 \int_\Omega (\mathbf{h} \cdot \mathbf{n})^2 dx. \tag{1.3}$$

We shall write

$$\begin{aligned} \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}] &= \mathcal{F}[\mathbf{n}] - \chi\sigma^2 \int_\Omega (\mathbf{h} \cdot \mathbf{n})^2 dx \\ &= \int_\Omega \{K_1 |\operatorname{div} \mathbf{n}|^2 + K_2 |\operatorname{curl} \mathbf{n}|^2 - \chi\sigma^2 (\mathbf{h} \cdot \mathbf{n})^2\} dx. \end{aligned}$$

Under the above assumptions, one can easily see that the energy functional $\mathcal{E}$ has two families of trivial critical points. One is given by

$$\psi = 0, \quad \mathbf{n} = \mathbf{n}_\sigma, \tag{1.4}$$

where $\mathbf{n}_\sigma$ is a global minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$:

$$\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}_\sigma] = \inf_{\mathbf{n} \in W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e})} \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}].$$

The second family is

$$\psi = ce^{iq\mathbf{e}\cdot\mathbf{x}}, \quad \mathbf{n} = \mathbf{e}, \tag{1.5}$$

where c is an arbitrary complex number such that $|c| = 1$. Following de Gennes' view on the analogies of superconductors and liquid crystals, we may compare the family (1.4) with the normal state of superconductors and call them *pure nematic states* and compare the family (1.5) with the Meissner state of superconductors and call them *pure smectic states*. We shall see in section 5 (Lemma 5.1) that $\mathbf{n}_\sigma = \mathbf{e}$ if σ is below a critical field $H_{\mathrm{n}}(0)$. Hence for $0 \le \sigma < H_{\mathrm{n}}(0)$, the only pure nematic state is $(0, \mathbf{e})$.

In this paper we shall introduce critical magnetic fields H_{sh} and H_{s}, where the pure smectic states change their weak stability (local minimality) at H_{sh} and change their strong stability (global minimality) at H_{s}.

One of the main goals of the paper is to show that

(i) if $0 \le \sigma < H_{\mathrm{s}}(\kappa, q)$, the pure smectic states are the only global minimizers of the functional $\mathcal{E}$;

(ii) if $H_{\mathrm{s}}(\kappa, q) < \sigma < H_{\mathrm{sh}}(q)$, the pure smectic states are local minimizers but not global minimizers; and

(iii) if $\sigma > H_{\mathrm{sh}}(q)$, the pure smectic states are not local minimizers.

We shall also give a criterion to test the minimality of the pure nematic states.

One believes that the critical field H_{sh} given here is an analogy of the superheating field of superconductors.[1] One may also expect existence of a critical magnetic field for liquid crystals that is an analogy of the upper critical magnetic field H_{C_3} for type II superconductors.[2] However, we are not able to provide a formal result to verify that

[1] See recent mathematical research on the superheating field for cylindrical superconductors by Chapman [Ch2], Lin and Du [LD], and Pan and Kwek [PK] and for bulk superconductors by Bates and Pan [BP].

[2] Nucleation of superconductivity and the upper critical field for type II superconductors have been studied by many mathematicians in recent years, among them Chapman [Ch1], Bauman, Phillips, and Tang [BPT], Bernoff and Sternberg [BS], Lu and Pan [LP1], [LP2], Helffer and Pan [HP], Helffer and Morame [HM1], [HM2], and Pan [P1], [P3].

this simple analogy exists, as liquid crystals and superconductors may have different responses in strong magnetic fields. In fact, when applied magnetic fields are above H_{C_3}, superconductors will be in the normal state, i.e., the normal state is the global minimizer of the Ginzburg–Landau functional. However, a liquid crystal under very strong magnetic fields may not be in a pure nematic state. Indeed, in section 5, we will give a proof of this conclusion in the case that $K_1 = K_2$.

Besides the normal and Meissner states, a superconductor may be in the so-called mixed state or in the surface superconducting state. Liquid crystals may also have many phases. Naturally, one is interested in various intermediate states of liquid crystals that can be described by nontrivial minimizers of $\mathcal{E}$. One also wishes to understand phase transitions between these states, as applied magnetic fields vary. We leave these problems to the future.

The outline of this paper is as follows. In section 2 we recall the concept of weak stability used by Cohen and Luskin [CL]. In section 3 we examine the weak instability in pure smectic states induced by external magnetic fields, and we introduce the critical magnetic field $H_{\rm sh}$. In section 4 we introduce the critical field $H_{\rm s}$. We also introduce an auxiliary field $H_{\rm n}$ and show that $H_{\rm s} \le H_{\rm n} \le H_{\rm sh}$. We then discuss the properties of these critical fields. In section 5 we examine the minimality of the pure nematic states. We then prove in the case of equal elastic coefficients that a liquid crystal under a strong magnetic field may not be in a pure nematic state.

We should point out that there exist many mathematical research works on liquid crystals done by physicists and mathematicians; see, for instance, [dGP], [K], [B], [E], [HKL], [L1], [L2], [L3], [LL], and the references therein. The magnetic field-induced instabilities and the Freedericksz transitions of nematic liquid crystals have been investigated in some simple cases by Cohen and Luskin [CL] and Atkin and Stewart [AS1], [AS2]. For recent work on the Landau–de Gennes model we refer to [Ca] and [BCLP].

2. Weak stability of critical points. In this section we consider the functional $\mathcal{E}$ defined in (1.1) for a general magnetic field $\mathbf{H}$. Following the earlier works [H], [CT], and [CL], we define the weak stability of critical points of $\mathcal{E}$ as follows.

DEFINITION 2.1. *$(\psi_0, \mathbf{n}_0) \in W^{1,2}(\Omega, \mathbb{C}) \times W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{u}_0)$ is called a critical point of the functional $\mathcal{E}$ if for any $\phi \in W^{1,2}(\Omega, \mathbb{C})$ and $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$, we have*

$$\frac{d}{dt}\Big|_{t=0} \mathcal{E}[\psi_t, \mathbf{n}_t] = 0;$$

here

$$\psi_t = \psi_0 + t\phi, \qquad \mathbf{n}_t = \frac{\mathbf{n}_0 + t\mathbf{v}}{|\mathbf{n}_0 + t\mathbf{v}|}. \tag{2.1}$$

A critical point $(\psi_0, \mathbf{n}_0)$ is said to be weakly stable if for any $\phi \in W^{1,2}(\Omega, \mathbb{C})$ and $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$, there exists a positive number T depending on ϕ and $\mathbf{v}$ such that for all $0 < t < T$ it holds that

$$\mathcal{E}[\psi_0, \mathbf{n}_0] \le \mathcal{E}[\psi_t, \mathbf{n}_t].$$

For ψ_t and $\mathbf{n}_t$ given in (2.1), computations show that (also see [CL])

$$\begin{aligned} &\mathbf{n}_t = \mathbf{n}_0 + t\mathbf{n}_1 + t^2\mathbf{n}_2 + O(t^3), \qquad \text{where} \\ &\mathbf{n}_1 = \mathbf{v} - (\mathbf{v}\cdot\mathbf{n}_0)\mathbf{n}_0, \\ &\mathbf{n}_2 = -(\mathbf{v}\cdot\mathbf{n}_0)\mathbf{v} + \frac{1}{2}\big[3(\mathbf{v}\cdot\mathbf{n}_0)^2 - |\mathbf{v}|^2\big]\mathbf{n}_0, \end{aligned} \tag{2.2}$$

and

$$\begin{aligned}\nabla_{q\mathbf{n}_t}\psi_t &= \nabla_{q\mathbf{n}_0}\psi_0 + t\Phi_1 + t^2\Phi_2 + O(t^3), \qquad \text{where}\\ \Phi_1 &= \nabla_{q\mathbf{n}_0}\phi - iq\mathbf{n}_1\psi_0,\\ \Phi_2 &= -iq(\mathbf{n}_1\phi + \mathbf{n}_2\psi_0).\end{aligned} \tag{2.3}$$

For small t, we have

$$\begin{aligned}&\mathcal{G}[\psi_t,\mathbf{n}_t] = \mathcal{G}[\psi_0,\mathbf{n}_0]\\ &+ 2t\int_\Omega\Big\{\Re[\overline{\nabla_{q\mathbf{n}_0}\phi}\nabla_{q\mathbf{n}_0}\psi_0 - \kappa^2\bar\phi(1-|\psi_0|^2)\psi_0] - q\mathbf{n}_1\Im(\bar\psi_0\nabla_{q\mathbf{n}_0}\psi_0)\Big\}dx\\ &+ t^2\int_\Omega\Big\{|\Phi_1|^2 - \kappa^2(1-|\psi_0|^2)|\phi|^2 + 2\kappa^2(\Re(\bar\phi\psi_0))^2 - 2q\Im[(\mathbf{n}_1\bar\phi + \mathbf{n}_2\bar\psi_0)\nabla_{q\mathbf{n}_0}\psi_0]\Big\}dx\\ &+ O(t^3);\end{aligned}$$

$$\begin{aligned}&\mathcal{F}[\mathbf{n}_t] = \mathcal{F}[\mathbf{n}_0] + 2t\int_\Omega\{K_1\mathrm{div}\,\mathbf{n}_0\,\mathrm{div}\,\mathbf{n}_1 + K_2\mathrm{curl}\,\mathbf{n}_0\cdot\mathrm{curl}\,\mathbf{n}_1\}dx\\ &+ t^2\int_\Omega\big\{K_1[(\mathrm{div}\,\mathbf{n}_1)^2 + 2\mathrm{div}\,\mathbf{n}_0\,\mathrm{div}\,\mathbf{n}_2] + K_2[|\mathrm{curl}\,\mathbf{n}_1|^2 + 2\mathrm{curl}\,\mathbf{n}_0\cdot\mathrm{curl}\,\mathbf{n}_2]\big\}dx\\ &+ O(t^3);\end{aligned}$$

and

$$\begin{aligned}\int_\Omega(\mathbf{H}\cdot\mathbf{n}_t)^2dx &= \int_\Omega(\mathbf{H}\cdot\mathbf{n}_0)^2dx + 2t\int_\Omega(\mathbf{H}\cdot\mathbf{n}_0)(\mathbf{H}\cdot\mathbf{n}_1)dx\\ &\quad + t^2\int_\Omega\{(\mathbf{H}\cdot\mathbf{n}_1)^2 + 2(\mathbf{H}\cdot\mathbf{n}_0)(\mathbf{H}\cdot\mathbf{n}_2)\}dx.\end{aligned}$$

So

$$\begin{aligned}\mathcal{E}[\psi_t,\mathbf{n}_t] &= \mathcal{E}[\psi_0,\mathbf{n}_0]\\ &+ 2t\left\{\mathcal{A}(\psi_0,\mathbf{n}_0;\phi,\mathbf{v}) - \chi\int_\Omega(\mathbf{H}\cdot\mathbf{n}_0)(\mathbf{H}\cdot\mathbf{n}_1)dx\right\}\\ &+ t^2\left\{\mathcal{B}(\psi_0,\mathbf{n}_0;\phi,\mathbf{v}) - \chi\int_\Omega\{(\mathbf{H}\cdot\mathbf{n}_1)^2 + 2(\mathbf{H}\cdot\mathbf{n}_0)(\mathbf{H}\cdot\mathbf{n}_2)\}dx\right\}\\ &+ O(t^3),\end{aligned} \tag{2.4}$$

where

$$\begin{aligned}\mathcal{A}(\psi_0,\mathbf{n}_0;\phi,\mathbf{v}) &= \int_\Omega\Big\{\Re[\overline{\nabla_{q\mathbf{n}_0}\phi}\nabla_{q\mathbf{n}_0}\psi_0 - \kappa^2\bar\phi(1-|\psi_0|^2)\psi_0] - q\mathbf{n}_1\Im(\bar\psi_0\nabla_{q\mathbf{n}_0}\psi_0)\\ &\qquad + K_1\mathrm{div}\,\mathbf{n}_0\,\mathrm{div}\,\mathbf{n}_1 + K_2\mathrm{curl}\,\mathbf{n}_0\cdot\mathrm{curl}\,\mathbf{n}_1\Big\}dx\end{aligned} \tag{2.5}$$

and

$$\begin{aligned}&\mathcal{B}(\psi_0,\mathbf{n}_0;\phi,\mathbf{v})\\ &= \int_\Omega\Big\{|\nabla_{q\mathbf{n}_0}\phi - iq\mathbf{n}_1\psi_0|^2 - \kappa^2(1-|\psi_0|^2)|\phi|^2 + 2\kappa^2[\Re(\bar\phi\psi_0)]^2\\ &\quad - 2q\Im[(\mathbf{n}_1\bar\phi + \mathbf{n}_2\bar\psi_0)\nabla_{q\mathbf{n}_0}\psi_0] + K_1[(\mathrm{div}\,\mathbf{n}_1)^2 + 2\mathrm{div}\,\mathbf{n}_0\,\mathrm{div}\,\mathbf{n}_2]\\ &\quad + K_2[|\mathrm{curl}\,\mathbf{n}_1|^2 + 2\mathrm{curl}\,\mathbf{n}_0\cdot\mathrm{curl}\,\mathbf{n}_2]\Big\}dx.\end{aligned} \tag{2.6}$$

Therefore we have the following conclusions.

LEMMA 2.2. *$(\psi_0, \mathbf{n}_0) \in W^{1,2}(\Omega, \mathbb{C}) \times W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{u}_0)$ is a critical point of the functional $\mathcal{E}$ if and only if for any $\phi \in W^{1,2}(\Omega, \mathbb{C})$ and $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$ it holds that*

$$\mathcal{A}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v}) - \chi \int_\Omega (\mathbf{H} \cdot \mathbf{n}_0)(\mathbf{H} \cdot \mathbf{n}_1) dx = 0. \tag{2.7}$$

If a critical point $(\psi_0, \mathbf{n}_0) \in W^{1,2}(\Omega, \mathbb{C}) \times W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{u}_0)$ is weakly stable, then for any $\phi \in W^{1,2}(\Omega, \mathbb{C})$ and $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$ we have

$$\mathcal{B}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v}) \geq \chi \int_\Omega \{(\mathbf{H} \cdot \mathbf{n}_1)^2 + 2(\mathbf{H} \cdot \mathbf{n}_0)(\mathbf{H} \cdot \mathbf{n}_2)\} dx. \tag{2.8}$$

Remark 2.1. From (2.7) we see that if $(\psi, \mathbf{n})$ is a critical point of $\mathcal{E}$, then it satisfies the equation

$$\begin{cases} -\nabla_{q\mathbf{n}}^2 \psi = \kappa^2 (1 - |\psi|^2)\psi & \text{in } \Omega, \\ \nabla_{q\mathbf{n}} \psi \cdot \nu = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.9}$$

3. Magnetic field-induced instability: Loss of local minimality of pure smectic states. Henceforth, we assume $\mathbf{H} = \sigma \mathbf{h}$ and $\mathbf{u}_0 = \mathbf{e}$ as in (1.2), where σ is a positive number, $\mathbf{h}$ and $\mathbf{e}$ are unit vectors such that $\mathbf{h} \cdot \mathbf{e} = 0$, and we assume the functional $\mathcal{E}$ is given by (1.3). We shall examine, in this section, the weak stability of the critical points given in (1.5) that correspond with a pure smectic state. For any $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$, let ψ_t and $\mathbf{n}_t$ be defined by (2.1). In the present case we have

$$\mathbf{n}_1 = \mathbf{v} - (\mathbf{v} \cdot \mathbf{e})\mathbf{e}, \qquad \mathbf{H} \cdot \mathbf{n}_1 = \sigma(\mathbf{v} \cdot \mathbf{h}).$$

Using Lemma 2.2 we conclude that if the critical point $(\psi_0, \mathbf{n}_0) = (ce^{iq\mathbf{e}\cdot\mathbf{x}}, \mathbf{e})$ given in (1.5) is weakly stable, then for any $\phi \in W^{1,2}(\Omega, \mathbb{C})$ and $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$ there holds that

$$\mathcal{B}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v}) \geq \chi\sigma^2 \int_\Omega |\mathbf{v} \cdot \mathbf{h}|^2 dx. \tag{3.1}$$

Note that $W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$ is dense in $W_0^{1,2}(\Omega, \mathbb{R}^3)$. If the above inequality holds for all $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3) \cap L^\infty(\Omega, \mathbb{R}^3)$, then it must hold for all $\mathbf{v} \in W_0^{1,2}(\Omega, \mathbb{R}^3)$.

With the particular choice of $\mathbf{H}$ and $\mathbf{n}_0$ given in (1.2), we can simplify the expression of $\mathcal{B}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v})$. Since

$$\nabla_{q\mathbf{n}_0} \psi_0 = 0, \qquad |\psi_0| = 1,$$

we have

$$\begin{aligned} &\mathcal{B}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v}) \\ &= \int_\Omega \{|\nabla_{q\mathbf{n}_0} \phi - iq\mathbf{n}_1\psi_0|^2 + 2\kappa^2(\Re(\bar{\phi}\psi_0))^2 + K_1|\operatorname{div} \mathbf{n}_1|^2 + K_2|\operatorname{curl} \mathbf{n}_1|^2\} dx. \end{aligned}$$

For any $\phi \in W^{1,2}(\Omega, \mathbb{C})$, one can always write

$$\phi = icqe^{iq\mathbf{n}_0\cdot\mathbf{x}} u, \qquad \text{where} \quad u \in W^{1,2}(\Omega, \mathbb{C}).$$

Thus

$$\nabla_{q\mathbf{n}_0}\phi - iq\mathbf{n}_1\psi_0 = icqe^{iq\mathbf{n}_0\cdot x}(\nabla u - \mathbf{n}_1),$$
$$\Re(\bar{\phi}\psi_0) = |c|^2\Re(iqu) = -q\Im(u).$$

Here we have used the condition $|c| = 1$. Let us change the notation by writing $\mathbf{w}$ for $\mathbf{n}_1$, namely,

$$\mathbf{n}_1 = \mathbf{w} \in W_0^{1,2}(\Omega, \mathbb{R}^3), \qquad \mathbf{w}(x)\cdot\mathbf{e} = 0 \quad \text{in } \Omega.$$

Then one has

$$\mathcal{B}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v}) = \int_\Omega \{q^2|\nabla u - \mathbf{w}|^2 + 2\kappa^2 q^2(\Im(u))^2 + K_1|\operatorname{div}\mathbf{w}|^2 + K_2|\operatorname{curl}\mathbf{w}|^2\}dx.$$

Obviously, if $(\phi, \mathbf{v})$ is such that $\mathcal{B}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v})/\|\mathbf{v}\cdot\mathbf{h}\|^2_{L^2(\Omega)}$ is minimal, then

$$u = -\frac{i}{cq}e^{-iq\mathbf{n}_0\cdot\mathbf{x}}\phi$$

is real-valued. Thus it suffices to consider u as a real-valued function for our purposes, and we write

$$\mathcal{B}(\psi_0, \mathbf{n}_0; \phi, \mathbf{v}) = B(u, \mathbf{w}),$$

where

$$B(u, \mathbf{w}) = \int_\Omega \{q^2|\nabla u - \mathbf{w}|^2 + K_1|\operatorname{div}\mathbf{w}|^2 + K_2|\operatorname{curl}\mathbf{w}|^2\}dx. \tag{3.2}$$

We define now a nonnegative number $H_{\rm sh}$ as follows.

DEFINITION 3.1. *Given $q \geq 0$, $K_1 > 0$, $K_2 > 0$, and a pair of mutually orthogonal unit vectors $\mathbf{h}$ and $\mathbf{e}$, we define $H_{\rm sh} = H_{\rm sh}(q, K_1, K_2, \Omega, \mathbf{h}, \mathbf{e})$ by*

$$H_{\rm sh}^2 = \frac{1}{\chi}\inf\left\{\frac{q^2\|\nabla u - \mathbf{w}\|^2_{L^2(\Omega)} + \mathcal{F}[\mathbf{w}]}{\int_\Omega |\mathbf{h}\cdot\mathbf{w}|^2 dx} : (u, \mathbf{w}) \in W^{1,2}(\Omega)\times W_0^{1,2}(\Omega, \mathbb{R}^3),\ \mathbf{w}(x)\cdot\mathbf{e} = 0 \text{ in } \Omega\right\}. \tag{3.3}$$

From the above discussion we have the next lemma.

LEMMA 3.2. *Under condition* (1.2)*, the pure smectic states lose their weak stability when σ increases to the critical value $H_{\rm sh}$. More precisely, a pure smectic state is weakly stable if $|\sigma| < H_{\rm sh}$, and it is not weakly stable if $|\sigma| > H_{\rm sh}$.*

We may further simplify the expression in the right-hand side of (3.3). Note that if $(u, \mathbf{w})$ is a minimizer of $H_{\rm sh}$, then

$$u = \xi_{\mathbf{w}},$$

where $\xi_{\mathbf{w}}$ is a solution of

$$\Delta\xi_{\mathbf{w}} = \operatorname{div}\mathbf{w} \quad \text{in } \Omega, \qquad \frac{\partial\xi_{\mathbf{w}}}{\partial\nu} = \gamma_\nu\mathbf{w} \quad \text{on } \partial\Omega, \qquad \int_\Omega \xi_{\mathbf{w}}dx = 0, \tag{3.4}$$

where $\gamma_\nu \mathbf{w}$ is the trace of $\mathbf{w}$ on the boundary, which is equal to the restriction of $\nu \cdot \mathbf{w}$ on $\partial\Omega$ when $\mathbf{w}$ is smooth. Moreover,

$$\int_\Omega |\nabla \xi_{\mathbf{w}} - \mathbf{w}|^2 dx = \omega(\mathbf{w})|\Omega|,$$

where

$$\omega(\mathbf{w}) = \inf_{\xi \in W^{1,2}(\Omega)} \fint_\Omega |\nabla \xi - \mathbf{w}|^2 dx. \tag{3.5}$$

Write

$$B(\mathbf{w}) = B(\xi_{\mathbf{w}}, \mathbf{w}). \tag{3.6}$$

Obviously we have, for any $b > 0$,

$$\xi_{b\mathbf{w}} = b\xi_{\mathbf{w}}.$$

Thus

$$B(b\mathbf{w}) = bB(\mathbf{w}).$$

We have therefore verified

$$H_{\rm sh}^2 = \frac{1}{\chi} \inf \Big\{ B(\mathbf{w}) : \ \mathbf{w} \in W_0^{1,2}(\Omega, \mathbb{R}^3), \ \mathbf{w}(x) \cdot \mathbf{e} = 0 \text{ in } \Omega, \ \|\mathbf{h} \cdot \mathbf{w}\|_{L^2(\Omega)} = 1 \Big\}. \tag{3.7}$$

Note that the minimizers $(\psi_0, \mathbf{n}_0)$ given in (1.5) contain a complex number c. However, $B(u, \mathbf{w})$ and $B(\mathbf{w})$ are independent of c.

In the following, we write $H_{\rm sh}$ by $H_{\rm sh}(q)$ to emphasize the dependence on the parameter q. We will see in section 5 that the critical value $H_{\rm sh}(0)$ is particularly interesting:

$$H_{\rm sh}^2(0) = \frac{1}{\chi} \inf \big\{ \mathcal{F}[\mathbf{w}] : \ \mathbf{w} \in W_0^{1,2}(\Omega, \mathbb{R}^3), \ \mathbf{w}(x) \cdot \mathbf{e} = 0 \text{ in } \Omega, \ \|\mathbf{h} \cdot \mathbf{w}\|_{L^2(\Omega)} = 1 \big\}. \tag{3.8}$$

PROPOSITION 3.3. *$H_{\rm sh}(q) > 0$ and it is achieved. For fixed K_1, K_2, Ω, $\mathbf{h}$, $\mathbf{e}$, we have*

$$\lim_{q \to +\infty} H_{\rm sh}(q) = \infty. \tag{3.9}$$

Proof. Step 1. Let $\mathbf{w}_j \in W_0^{1,2}(\Omega, \mathbb{R}^3)$ and $\xi_j = \xi_{\mathbf{w}_j}$ be such that

$$\mathbf{w}_j \cdot \mathbf{e} = 0 \quad \text{in } \Omega, \quad \|\mathbf{w}_j \cdot \mathbf{h}\|_{L^2(\Omega)} = 1, \quad B(\mathbf{w}_j) \to \chi H_{\rm sh}^2 \quad \text{as } j \to \infty.$$

Since $\mathbf{w}_j \in W_0^{1,2}(\Omega, \mathbb{R}^3)$, we have

$$\int_\Omega |\nabla \mathbf{w}_j|^2 dx = \int_\Omega \{|\operatorname{div} \mathbf{w}_j|^2 + |\operatorname{curl} \mathbf{w}_j|^2\} dx \le \frac{1}{\min\{K_1, K_2\}} B(\mathbf{w}_j) \le C.$$

Hence $\{\mathbf{w}_j\}$ is bounded in $W_0^{1,2}(\Omega, \mathbb{R}^3)$. After passing to a subsequence we may assume that, as $j \to \infty$,

$$\mathbf{w}_j \to \mathbf{w}_0 \quad \text{weakly in } W_0^{1,2}(\Omega, \mathbb{R}^3) \text{ and strongly in } L^2(\Omega, \mathbb{R}^3).$$

Then $\mathbf{w}_0(x)\cdot\mathbf{e}=0$ for a.e. $x\in\Omega$. We also have

$$\int_\Omega |\nabla\xi_j-\mathbf{w}_j|^2dx\le\frac{1}{q^2}B(\mathbf{w}_j)\le C.$$

From this and the fact $\int_\Omega\xi_j dx=0$, we see that $\{\xi_j\}$ is bounded in $W^{1,2}(\Omega)$. Note that

$$\|\operatorname{div}\mathbf{w}_j\|^2_{L^2(\Omega)}\le C(\Omega)\|\mathbf{w}_j\|^2_{W^{1,2}(\Omega)}\le C,$$

and $\mathbf{w}_j=\mathbf{0}$ on $\partial\Omega$. Applying elliptic estimates to (3.4) for ξ_j with $\mathbf{w}=\mathbf{w}_j$, we find that $\{\xi_j\}$ is bounded in $W^{2,2}(\Omega)$. Passing to a subsequence we have $\xi_j\to\xi_0$ weakly in $W^{2,2}(\Omega)$ and strongly in $W^{1,2}(\Omega)$ as $j\to\infty$, and ξ_0 satisfies (3.4) for $\mathbf{w}=\mathbf{w}_0$. Thus $\xi_0=\xi_{\mathbf{w}_0}$. Hence we have proved that

$$B(\mathbf{w}_0)=B(\xi_{\mathbf{w}_0},\mathbf{w}_0)\le\liminf_{j\to\infty}B(\xi_j,\mathbf{w}_j)=\chi H^2_{\rm sh}.$$

On the other hand, since $\|\mathbf{w}_0\cdot\mathbf{h}\|_{L^2(\Omega)}=\lim_{j\to\infty}\|\mathbf{w}_j\cdot\mathbf{h}\|_{L^2(\Omega)}=1$, we have

$$B(\mathbf{w}_0)\ge\chi H^2_{\rm sh}(q).$$

Thus $\mathbf{w}_0$ is a minimizer of $B(\mathbf{w})$, i.e., $(\xi_0,\mathbf{w}_0)$ achieves $H_{\rm sh}(q)$. It in turn implies that $H_{\rm sh}(q)>0$.

Step 2. Suppose $H_{\rm sh}(q)\le c$ for all q. We may assume $\mathbf{e}=\mathbf{e}_1$. Let us choose $q_j\to+\infty$ and choose $u_j\in W^{1,2}(\Omega)$ and $\mathbf{w}_j\in W^{1,2}_0(\Omega,\mathbb{R}^3)$ such that $\int_\Omega u_j dx=0$, $\mathbf{e}_1\cdot\mathbf{w}_j=0$, and $(u_j,\mathbf{w}_j)$ achieves $H_{\rm sh}(q_j)$. Thus

$$\int_\Omega\{q_j^2|\nabla u_j-\mathbf{w}_j|^2+K_1|\operatorname{div}\mathbf{w}_j|^2+K_2|\operatorname{curl}\mathbf{w}_j|^2\}dx\le\chi c^2\int_\Omega(\mathbf{h}\cdot\mathbf{w}_j)^2dx.$$

In particular, one has $\|\operatorname{div}\mathbf{w}_j\|_{L^2(\Omega)}\le C$ and $\|\operatorname{curl}\mathbf{w}_j\|_{L^2(\Omega)}\le C$. Recall that for a bounded and simply connected domain with smooth boundary, the following inequality holds:

(3.10)
$$\|\mathbf{B}\|_{H^{k+1}(\Omega)}\le C(\Omega,k)\left\{\|\operatorname{div}\mathbf{B}\|_{H^k(\Omega)}+\|\operatorname{curl}\mathbf{B}\|_{H^k(\Omega)}+\left\|\begin{matrix}\nu\cdot\mathbf{B}\\ \nu\times\mathbf{B}\end{matrix}\right\|_{H^{k+1/2}(\partial\Omega)}\right\},$$

where $\left\|\begin{smallmatrix}\nu\cdot\mathbf{B}\\ \nu\times\mathbf{B}\end{smallmatrix}\right\|_{H^{k+1/2}(\partial\Omega)}$ means either $\|\nu\cdot\mathbf{B}\|_{H^{k+1/2}(\partial\Omega)}$ or $\|\nu\times\mathbf{B}\|_{H^{k+1/2}(\partial\Omega)}$; see Theorem 3 on p. 209 and Proposition 6 on p. 237 in [DL]. Since $\mathbf{w}_j\in W^{1,2}_0(\Omega,\mathbb{R}^3)$, from the above inequalities we see that $\{\mathbf{w}_j\}$ is bounded in $W^{1,2}_0(\Omega,\mathbb{R}^3)$. After passing to a subsequence we may assume that

$$\mathbf{w}_j\to\hat{\mathbf{w}}\quad\text{weakly in } W^{1,2}_0(\Omega,\mathbb{R}^3)\text{ and strongly in } L^4(\Omega,\mathbb{R}^3)\text{ as } j\to\infty.$$

In particular, $\|\hat{\mathbf{w}}\|_{L^2(\Omega)}=1$. Since $\|\nabla u_j-\mathbf{w}_j\|_{L^2(\Omega)}=O(q_j^{-1})$ and $\mathbf{w}_j\to\hat{\mathbf{w}}$ strongly in $L^2(\Omega,\mathbb{R}^3)$, we conclude that

$$u_j\to\hat u\quad\text{strongly in } W^{1,2}(\Omega)\quad\text{and } \nabla\hat u=\hat{\mathbf{w}}.$$

In particular, $\frac{\partial\hat u}{\partial x_1}=\mathbf{e}_1\cdot\nabla\hat u=\mathbf{e}_1\cdot\hat{\mathbf{w}}=0$, and hence $\hat u=\hat u(x_2,x_3)$. On the other hand, via the $W^{2,2}$-estimates for u_j's described in the first step, one concludes that

$\nabla\hat{u} = \mathbf{w} = \mathbf{0}$ on $\partial\Omega$. These two facts together imply that $\nabla\hat{u} = \mathbf{0}$ on the whole Ω, and hence $\mathbf{w} = \nabla\hat{u} = \mathbf{0}$, which contradicts the fact that $\|\hat{\mathbf{w}}\|_{L^2(\Omega)} = 1$. This completes the proof. □

Now we derive the Euler–Lagrange equation for the minimizers of $H_{\rm sh}$. We can rotate the coordinate system and assume without loss of generality that $\mathbf{h} = \mathbf{e}_1$ and $\mathbf{u}_0 = \mathbf{e}_3$. Hence

$$\mathbf{H} = \sigma\mathbf{e}_1 = (\sigma, 0, 0), \qquad \mathbf{u}_0 = \mathbf{e}_3 = (0, 0, 1),$$

and

$$\mathbf{w} = (w_1, w_2, 0), \qquad \mathbf{w} \cdot \mathbf{h} = w_1.$$

Define an operator

$$P : \mathbb{R}^3 \to \mathbb{R}^2, \quad P(x_1, x_2, x_3) = (x_1, x_2).$$

Let $\mathbf{w} = (w_1, w_2, 0)$ be a minimizer of $H_{\rm sh}$. Then $\mathbf{w}$ satisfies the following Euler–Lagrange equations:

$$\begin{cases} P(-K_1\nabla\operatorname{div}\mathbf{w} + K_2\operatorname{curl}^2\mathbf{w} + q^2(\mathbf{w} - \nabla\xi_{\mathbf{w}})) = (\chi H_{\rm sh}^2 w_1, 0) & \text{in } \Omega, \\ \mathbf{w} = \mathbf{0} \quad \text{on } \partial\Omega. \end{cases}$$

In the special case when $K_1 = K_2 \equiv K$ we have

$$\begin{cases} -K\Delta w_1 + (q^2 - \chi H_{\rm sh}^2) w_1 = q^2 \partial_1 \xi_{\mathbf{w}}, \\ -K\Delta w_2 + q^2 w_2 = q^2 \partial_2 \xi_{\mathbf{w}} & \text{in } \Omega, \\ w_1 = w_2 = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.11}$$

We would like to point out that (3.7) defines a minimization problem under the pointwise orthogonality condition

$$\mathbf{w}(x) \cdot \mathbf{e} = 0 \quad \text{for all } x \in \Omega. \tag{3.12}$$

It is the orthogonality condition (3.12) that makes the minimization problem for $H_{\rm sh}$ interesting. In particular, from (3.9) we see that, for fixed $\mathbf{h}$, $\mathbf{e}$, K_1 and $K_2 > 0$, $H_{\rm sh}(q)$ diverges as $q \to \infty$. In contrast, if we consider a similar minimization problem formulated by removing the orthogonality condition (3.12) from the definition of $H_{\rm sh}(q)$ and define $\tilde{H}(q) = \tilde{H}(q, K_1, K_2, \Omega, \mathbf{h})$ by

$$\tilde{H}^2(q) = \frac{1}{\chi} \inf \left\{ B(\mathbf{w}) : \ \mathbf{w} \in W_0^{1,2}(\Omega, \mathbb{R}^3), \ \ \|\mathbf{h} \cdot \mathbf{w}\|_{L^2(\Omega)} = 1 \right\},$$

then there exists a constant C independent of q such that

$$\tilde{H}(q) \le C \left(\frac{K_1}{\chi} \right)^{1/2} \quad \text{for all } q. \tag{3.13}$$

To prove (3.13), we assume $\mathbf{h} = \mathbf{e}_1$. For any $u \in W^{2,2}(\Omega)$ such that $\nabla u = 0$ on $\partial\Omega$, take $\mathbf{w} = \nabla u$. Then

$$B(\nabla u) = B(u, \nabla u) = K_1 \|\Delta u\|_{L^2(\Omega)}^2.$$

Therefore (3.13) holds with C defined by

$$C = \inf \left\{ \frac{\|\Delta u\|_{L^2(\Omega)}^2}{\|\partial_1 u\|_{L^2(\Omega)}^2} : \ u \in W^{2,2}(\Omega), \ \ \nabla u = 0 \text{ on } \partial\Omega \right\}.$$

4. Magnetic field-induced instabilities: Loss of global minimality of pure smectic states. In this section we examine the change of the global minimality of the pure smectic states when the applied magnetic fields vary. For a pure smectic state $(\psi_0, \mathbf{n}_0)$ given in (1.5), we have

$$\mathbf{h} \cdot \mathbf{n}_0 \equiv 0 \quad \text{in } \Omega, \qquad \mathcal{E}[\psi_0, \mathbf{n}_0] = 0.$$

Now we consider a global minimizer $(\psi, \mathbf{n})$ of the functional $\mathcal{E}$ in the space $W^{1,2}(\Omega, \mathbb{C}) \times W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{u}_0)$. A simple computation shows that if the global minimizer $(\psi, \mathbf{n})$ is not a pure smectic state, then $\mathbf{h} \cdot \mathbf{n} \not\equiv 0$ in Ω, and $\mathcal{E}[\psi, \mathbf{n}] \leq \mathcal{E}[\psi_0, \mathbf{n}_0] \leq 0$. (See Lemma 5.1 for a similar discussion for minimizers of the functional $\mathcal{F}_{\sigma \mathbf{h}}$.) Therefore

$$\mathcal{G}[\psi, \mathbf{n}] + \mathcal{F}[\mathbf{n}] \leq \chi \sigma^2 \int_\Omega (\mathbf{h} \cdot \mathbf{n})^2 dx.$$

This observation leads to the following definition.

DEFINITION 4.1. *Given $q \geq 0$, $\kappa > 0$, $K_1 > 0$, $K_2 > 0$, and given a pair of mutually orthogonal unit vectors $\mathbf{h}$ and $\mathbf{e}$, the critical magnetic field $H_{\rm s} = H_{\rm s}(q, \kappa, K_1, K_2, \Omega, \mathbf{h}, \mathbf{e})$ is defined by*

$$H_{\rm s}^2 = \frac{1}{\chi} \inf \left\{ \frac{\mathcal{G}[\psi, \mathbf{n}] + \mathcal{F}[\mathbf{n}]}{\int_\Omega (\mathbf{h} \cdot \mathbf{n})^2 dx} : \ (\psi, \mathbf{n}) \in W^{1,2}(\Omega, \mathbb{C}) \times W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e}), \ \mathbf{h} \cdot \mathbf{n} \not\equiv 0 \right\}. \tag{4.1}$$

The critical field $H_{\rm s}$ can be used to test whether the pure smectic states are global minimizers. If we fix the parameters κ, q, K_1, and K_2, then the pure smectic states remain as global minimizers for small σ, and they lose the global minimality when σ increases and reaches the critical field $H_{\rm s}$.

LEMMA 4.2. *Under the above assumptions we have the following:*

(i) *If there exists a global minimizer $(\psi, \mathbf{n})$ that is not a pure smectic state, then $\sigma \geq H_{\rm s}$.*

(ii) *If $\sigma > H_{\rm s}$, then the global minimizers are not pure smectic states.*

(iii) *If $0 \leq \sigma < H_{\rm s}$, then the only global minimizers are the pure smectic states.*

In the following, we write $H_{\rm s}$ as $H_{\rm s}(\kappa, q)$ to emphasize its dependence on the parameters κ and q. Note that the pure smectic states lose their global minimality at the critical magnetic field $H_{\rm s}(\kappa, q)$ and lose their local minimality at the critical magnetic field $H_{\rm sh}(q)$. So we always have

$$0 < H_{\rm s}(\kappa, q) \leq H_{\rm sh}(q).$$

To get a better estimate of $H_{\rm s}(q)$, we define a number $H_{\rm n}$ which is closely related to $H_{\rm s}$ as follows.

DEFINITION 4.3. *The number $H_{\rm n} = H_{\rm n}(q, K_1, K_2, \Omega, \mathbf{h}, \mathbf{e})$ is defined by*

$$H_{\rm n}^2 = \frac{1}{\chi} \inf \left\{ \frac{q^2 \|\nabla u - \mathbf{n}\|_{L^2(\Omega)}^2 + \mathcal{F}[\mathbf{n}]}{\int_\Omega (\mathbf{h} \cdot \mathbf{n})^2 dx} : \right. \\ \left. (u, \mathbf{n}) \in W^{1,2}(\Omega) \times W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e}), \ \mathbf{h} \cdot \mathbf{n} \not\equiv 0 \right\}. \tag{4.2}$$

In the following we write $H_{\rm n}$ as $H_{\rm n}(q)$ to emphasize its dependence on q. In particular,

$$H_{\rm n}^2(0) = \frac{1}{\chi} \inf \left\{ \frac{\mathcal{F}[\mathbf{n}]}{\int_\Omega (\mathbf{h} \cdot \mathbf{n})^2 dx} : \ \mathbf{n} \in W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e}), \ \mathbf{h} \cdot \mathbf{n} \not\equiv 0 \right\}. \tag{4.3}$$

Note that $H_{\rm n}(0)$ depends on K_1, K_2, Ω, $\mathbf{h}$, and $\mathbf{e}$.

LEMMA 4.4. *$H_{\rm s}(\kappa,0)$ is independent of κ. In fact we have*

$$H_{\rm s}(\kappa,0)=H_{\rm n}(0)\quad \textit{for all } \kappa. \tag{4.4}$$

For any κ and q we have

$$H_{\rm s}(\kappa,q)\geq H_{\rm n}(0). \tag{4.5}$$

Proof. To verify (4.4), we take $\psi=1$ as a test function. Equation (4.5) follows from (4.4) and the definition of $H_{\rm s}(\kappa,q)$ immediately. ∎

The following theorem gives the relations between $H_{\rm s}(\kappa,q)$, $H_{\rm n}(q)$, and $H_{\rm sh}(q)$.

THEOREM 4.5. *Assume Ω is a simply connected and bounded domain with smooth boundary.*

(i) *For all $\kappa>0$, $q\geq 0$, we have*

$$0<H_{\rm s}(\kappa,q)\leq H_{\rm n}(q)\leq H_{\rm sh}(q). \tag{4.6}$$

(ii) *For any κ and q, if $H_{\rm s}(\kappa,q)<H_{\rm sh}(q)$, then $H_{\rm s}(\kappa,q)$ is achieved.*

(iii) *For any $q\geq 0$, if $H_{\rm n}(q)<H_{\rm sh}(q)$, then $H_{\rm n}(q)$ is achieved (and hence $H_{\rm s}(\kappa,q)<H_{\rm sh}(q)$ and $H_{\rm s}(\kappa,q)$ is achieved).*

Proof of (i). The inequality $H_{\rm s}(\kappa,q)\leq H_{\rm n}(q)$ is easy to verify. In fact, for any $\phi\in W^{1,2}(\Omega)$ and $\mathbf{n}\in W^{1,2}(\Omega,\mathbb{S}^2,\mathbf{e})$ with $\mathbf{h}\cdot\mathbf{n}\not\equiv 0$, choose $(e^{iq\phi},\mathbf{n})$ as a test function for $H_{\rm s}(\kappa,q)$. It is easy to show that

$$H_{\rm s}^2(\kappa,q)\leq\frac{1}{\chi}\frac{q^2\|\nabla\phi-\mathbf{n}\|^2_{L^2(\Omega)}+\mathcal{F}[\mathbf{n}]}{\int_\Omega(\mathbf{h}\cdot\mathbf{n})^2dx}.$$

Then we take infimum for all such $(\phi,\mathbf{n})$ to get the inequality $H_{\rm s}(\kappa,q)\leq H_{\rm n}(q)$.

To prove $H_{\rm n}(q)\leq H_{\rm sh}(q)$, let $u\in W^{1,2}(\Omega)$ and $\mathbf{w}\in W_0^{1,2}(\Omega,\mathbb{R}^3)\cap L^\infty(\Omega,\mathbb{R}^3)$, with $\mathbf{e}\cdot\mathbf{w}\equiv 0$ in Ω and $\mathbf{h}\cdot\mathbf{w}\not\equiv 0$. Set

$$\phi_t=\mathbf{e}\cdot\mathbf{x}+tqu,\qquad \mathbf{n}_t=\frac{\mathbf{e}+t\mathbf{w}}{|\mathbf{e}+t\mathbf{w}|}=\mathbf{e}+t\mathbf{w}+O(t^2).$$

Then

$$\begin{aligned}
&q^2\|\nabla\phi_t-\mathbf{n}_t\|^2_{L^2(\Omega)}+\mathcal{F}[\mathbf{n}_t]\\
&=t^2\{q^2\|\nabla u-\mathbf{w}+O(t)\|^2_{L^2(\Omega)}+\mathcal{F}[\mathbf{w}+O(t)]\}\\
&=t^2\{q^2\|\nabla u-\mathbf{w}\|^2_{L^2(\Omega)}+\mathcal{F}[\mathbf{w}]\}+O(t^3),\\
&\int_\Omega(\mathbf{h}\cdot\mathbf{n}_t)^2dx=t^2\int_\Omega(\mathbf{h}\cdot\mathbf{w})^2dx+O(t^3).
\end{aligned}$$

Hence

$$\frac{q^2\|\nabla\phi_t-\mathbf{n}_t\|^2_{L^2(\Omega)}+\mathcal{F}[\mathbf{n}_t]}{\int_\Omega(\mathbf{h}\cdot\mathbf{n}_t)^2dx}=\frac{q^2\|\nabla u-\mathbf{w}\|^2_{L^2(\Omega)}+\mathcal{F}[\mathbf{w}]}{\int_\Omega(\mathbf{h}\cdot\mathbf{w})^2dx}+O(t).$$

Letting $t\to 0$ we get

$$\chi H_{\rm n}^2(q)\leq\frac{q^2\|\nabla u-\mathbf{w}\|^2_{L^2(\Omega)}+\mathcal{F}[\mathbf{w}]}{\int_\Omega(\mathbf{h}\cdot\mathbf{w})^2dx}.$$

Then we take infimum in the right-hand side to get

$$\chi H_{\rm n}^2(q) \le \inf\left\{ \frac{q^2\|\nabla u - \mathbf{w}\|^2_{L^2(\Omega)} + \mathcal{F}[\mathbf{w}]}{\int_\Omega (\mathbf{h}\cdot\mathbf{w})^2 dx} : \right.$$
$$\left. u \in W^{1,2}(\Omega),\ \mathbf{w} \in W_0^{1,2}(\Omega,\mathbb{R}^3)\cap L^\infty(\Omega,\mathbb{R}^3),\ \mathbf{e}\cdot\mathbf{w}\equiv 0,\ \mathbf{h}\cdot\mathbf{w}\not\equiv 0 \right\}.$$

Since $W_0^{1,2}(\Omega,\mathbb{R}^3)\cap L^\infty(\Omega,\mathbb{R}^3)$ is dense in $W_0^{1,2}(\Omega,\mathbb{R}^3)$, we conclude that $H_{\rm n}(q) \le H_{\rm s}(q)$.

To prove $H_{\rm s}(\kappa,q) > 0$, recall that $H_{\rm sh}(q) > 0$. If $H_{\rm s}(\kappa,q) = H_{\rm sh}(q)$, then we have $H_{\rm s}(\kappa,q) > 0$; if $H_{\rm s}(\kappa,q) < H_{\rm sh}(q)$, then from conclusion (ii), which we shall prove below, we know that $H_{\rm s}(\kappa,q)$ is achieved. Assume $(\psi,\mathbf{n})$ achieves $H_{\rm s}(\kappa,q)$ and suppose $H_{\rm s}(\kappa,q) = 0$. Then

$$\mathcal{G}[\psi,\mathbf{n}] + \mathcal{F}[\mathbf{n}] = \chi H_{\rm s}^2(\kappa,q)\int_\Omega (\mathbf{h}\cdot\mathbf{n})^2 dx = 0, \qquad \mathbf{h}\cdot\mathbf{n}\not\equiv 0.$$

Thus $\mathbf{n}$ satisfies

$$\operatorname{div}\mathbf{n} = 0, \quad \operatorname{curl}\mathbf{n} = \mathbf{0}, \quad |\mathbf{n}(x)| = 1 \quad \text{a.e. in } \Omega, \quad \mathbf{n} = \mathbf{e} \quad \text{on } \partial\Omega.$$

Since Ω is simply connected, $\mathbf{n}$ is a constant vector and hence must be equal to $\mathbf{e}$ since $\mathbf{n} = \mathbf{e}$ on $\partial\Omega$. (See step 1.1 in the proof of Theorem 3.6 in [P2].) Then $\mathbf{h}\cdot\mathbf{n} \equiv 0$, which contradicts the assumption that $\mathbf{h}\cdot\mathbf{n} \not\equiv 0$.

Proof of (ii). In the following we assume $H_{\rm s}(\kappa,q) < H_{\rm sh}(q)$ and prove that $H_{\rm s}(\kappa,q)$ is achieved.

Step 1. Let $\{(\psi_j,\mathbf{n}_j)\}$ be a minimizing sequence for $H_{\rm s}(\kappa,q)$. Then

$$\mathcal{G}[\psi_j,\mathbf{n}_j] + \mathcal{F}[\mathbf{n}_j] = (\chi H_{\rm s}^2(\kappa,q) + o(1))\int_\Omega (\mathbf{h}\cdot\mathbf{n}_j)^2 dx. \tag{4.7}$$

Since $|\mathbf{h}\cdot\mathbf{n}_j| \le 1$, the right-hand side in (4.7) is bounded. Thus $\{\operatorname{div}\mathbf{n}_j\}$ and $\{\operatorname{curl}\mathbf{n}_j\}$ are bounded in $L^2(\Omega)$. Since $|\mathbf{n}_j| = 1$ a.e. and $\mathbf{n}_j = \mathbf{e}$ on the boundary, from (3.10) we see that $\{\mathbf{n}_j\}$ is bounded in $W^{1,2}(\Omega,\mathbb{R}^3)$. Therefore we can pass to a subsequence and assume that, as $j\to\infty$,

$$\mathbf{n}_j \to \hat{\mathbf{n}} \quad \text{weakly in } W^{1,2}(\Omega,\mathbb{R}^3) \text{ and strongly in } L^2(\Omega,\mathbb{R}^3).$$

It follows that $|\hat{\mathbf{n}}(x)| = \lim_{j\to\infty}|\mathbf{n}_j(x)| = 1$ for a.e. $x\in\Omega$ and $\hat{\mathbf{n}} = \mathbf{e}$ on $\partial\Omega$. Thus $\hat{\mathbf{n}} \in W^{1,2}(\Omega,\mathbb{S}^2,\mathbf{e})$. On the other hand, (4.7) implies that $\{|\nabla_{q\mathbf{n}_j}\psi_j|\}$ is bounded in $L^2(\Omega)$ and $\{\psi_j\}$ is bounded in $L^4(\Omega)$. Then from

$$\|\nabla\psi_j\|_{L^2(\Omega)} = \|\nabla_{q\mathbf{n}_j}\psi_j + iq\mathbf{n}_j\psi_j\|_{L^2(\Omega)} \le \|\nabla_{q\mathbf{n}_j}\psi_j\|_{L^2(\Omega)} + \|q\mathbf{n}_j\psi_j\|_{L^2(\Omega)}$$

we see that $\{\psi_j\}$ is bounded in $W^{1,2}(\Omega,\mathbb{C})$. Passing to subsequence again we assume that, as $j\to\infty$,

$$\psi_j \to \hat{\psi} \quad \text{weakly in } W^{1,2}(\Omega,\mathbb{C}) \text{ and strongly in } L^4(\Omega).$$

Taking limit in (4.7) we find

$$\mathcal{G}[\hat{\psi},\hat{\mathbf{n}}] + \mathcal{F}[\hat{\mathbf{n}}] \le \liminf_{j\to\infty}\{\mathcal{G}[\psi_j,\mathbf{n}_j] + \mathcal{B}F[\mathbf{n}_j]\} = \chi H_{\rm s}^2(\kappa,q)\int_\Omega (\mathbf{h}\cdot\hat{\mathbf{n}})^2 dx. \tag{4.8}$$

If $\mathbf{h}\cdot\hat{\mathbf{n}}\not\equiv 0$, then (4.8) implies that $(\hat{\psi},\hat{\mathbf{n}})$ achieves $H_{\rm s}(\kappa,q)$.

In the next two steps we shall prove that if $H_{\rm s}(\kappa,q)<H_{\rm sh}(q)$, then we must have $\mathbf{h}\cdot\hat{\mathbf{n}}\not\equiv 0$.

Step 2. In the following we assume $\mathbf{h}\cdot\hat{\mathbf{n}}\equiv 0$. Then from (4.8) we see that $\operatorname{div}\hat{\mathbf{n}}=0$ and $\operatorname{curl}\hat{\mathbf{n}}=\mathbf{0}$ in Ω. Since Ω is simply connected, $|\hat{\mathbf{n}}(x)|=1$ a.e. in Ω, and $\hat{\mathbf{n}}=\mathbf{e}$ on $\partial\Omega$, we have $\hat{\mathbf{n}}\equiv\mathbf{e}$. (See step 1.1 in the proof of Theorem 3.6 in [P2].) From (4.8) we also see that

$$\nabla\hat{\psi}-iq\mathbf{e}\hat{\psi}=0,\quad |\hat{\psi}|=1. \tag{4.9}$$

We claim that

$$\hat{\psi}=ce^{iq\mathbf{e}\cdot\mathbf{x}}$$

for some complex number c with $|c|=1$. In fact, since $\hat{\psi}\neq 0$, for any $x_0\in\Omega$, there exists a neighborhood of x_0, say, $U(x_0)$, on which $\hat{\psi}$ has a representation $\hat{\psi}=fe^{iq\phi}$ with $f>0$. The condition $|\hat{\psi}|=1$ implies $f=1$. From the first equality in (4.9) we have $\nabla\phi=\mathbf{e}$. So on $U(x_0)$ we have $\hat{\psi}=ce^{iq\mathbf{e}\cdot\mathbf{x}}$ for some complex number c with $|c|=1$. Since Ω is connected, the constant c does not depend on x_0. So the claim is true.

In the following, for simplicity we assume $c=1$ and

$$\psi=e^{iq\mathbf{e}\cdot\mathbf{x}}.$$

Recall that $\mathbf{h}\cdot\mathbf{n}_j\not\equiv 0$, so $\mathbf{n}_j\not\equiv\mathbf{e}$. Let us write

$$\mathbf{n}_j=\mathbf{e}+\varepsilon_j\mathbf{w}_j,\qquad \psi_j=e^{iq\mathbf{e}\cdot\mathbf{x}}(1+iq\varepsilon_j g_j), \tag{4.10}$$

where

$$\begin{aligned}&\varepsilon_j=\|\mathbf{n}_j-\mathbf{e}\|_{W^{1,2}(\Omega)}>0,\\ &\mathbf{w}_j\in W_0^{1,2}(\Omega,\mathbb{R}^3),\qquad \|\mathbf{w}_j\|_{W^{1,2}(\Omega)}=1.\end{aligned}$$

We shall estimate $\mathbf{w}_j$ and g_j.

Since $\mathbf{w}_j\in W_0^{1,2}(\Omega,\mathbb{R}^3)$, we have

$$\begin{aligned}1&=\int_\Omega\{|\nabla\mathbf{w}_j|^2+|\mathbf{w}_j|^2\}dx\le C\int_\Omega|\nabla\mathbf{w}_j|^2dx\\ &=C\int_\Omega\{|\operatorname{div}\mathbf{w}_j|^2+|\operatorname{curl}\mathbf{w}_j|^2\}dx\le\frac{C}{\min\{K_1,K_2\}}\mathcal{F}[\mathbf{w}_j]\\ &=\frac{C}{\varepsilon_j^2\min\{K_1,K_2\}}\mathcal{F}[\mathbf{n}_j],\end{aligned}$$

so

$$\varepsilon_j^2\le\frac{C}{\min\{K_1,K_2\}}\mathcal{F}[\mathbf{n}_j]=o(1).$$

Since $\{\mathbf{w}_j\}$ is bounded in $W_0^{1,2}(\Omega,\mathbb{R}^3)$, after passing to a subsequence again we may assume that, as $j\to\infty$,

$$\mathbf{w}_j\to\hat{\mathbf{w}}\quad\text{weakly in } W_0^{1,2}(\Omega,\mathbb{R}^3),\quad\text{strongly in } L^4(\Omega,\mathbb{R}^3).$$

Moreover,

$$1 = |\mathbf{n}_j(x)|^2 = |\mathbf{e} + \varepsilon_j \mathbf{w}_j(x)|^2 = 1 + 2\varepsilon_j \mathbf{e}\cdot\mathbf{w}_j(x) + \varepsilon_j^2 |\mathbf{w}_j(x)|^2,$$

so

$$\mathbf{e}\cdot\mathbf{w}_j(x) = -\frac{\varepsilon_j}{2}|\mathbf{w}_j(x)|^2 \to 0 \quad \text{strongly in } L^2(\Omega).$$

Thus

$$\mathbf{e}\cdot\hat{\mathbf{w}}(x) = 0 \quad \text{for a.e. } x \in \Omega. \tag{4.11}$$

Using (4.10) we compute

$$\begin{aligned} &|\nabla_{q\mathbf{n}_j}\psi_j|^2 = q^2\varepsilon_j^2 |\nabla g_j - (1 + iq\varepsilon_j g_j)\mathbf{w}_j|^2, \\ &|\psi_j|^2 = 1 + q\varepsilon_j(-2\Im(g_j) + q\varepsilon_j |g_j|^2). \end{aligned}$$

Using these and (4.7), (4.11) we get

$$\begin{aligned} &(\chi H_{\mathrm{s}}^2(\kappa, q) + o(1)) \int_\Omega (\mathbf{h}\cdot\mathbf{w}_j)^2 dx \\ =&\frac{1}{\varepsilon_j^2}\int_\Omega (\mathbf{h}\cdot\mathbf{n}_j)^2 dx = \frac{1}{\varepsilon_j^2}\mathcal{G}[\psi_j, \mathbf{n}_j] + \frac{1}{\varepsilon_j^2}\mathcal{F}[\mathbf{n}_j] \\ =&\int_\Omega \left\{ q^2 |\nabla g_j - (1 + iq\varepsilon_j g_j)\mathbf{w}_j|^2 + \frac{\kappa^2}{2} q^2 (-2\Im(g_j) + q\varepsilon_j |g_j|^2)^2 \right\} dx + \mathcal{F}[\mathbf{w}_j]. \end{aligned} \tag{4.12}$$

In particular, we have

$$\int_\Omega |\nabla g_j - \psi_j e^{-iq\mathbf{e}\cdot\mathbf{x}}\mathbf{w}_j|^2 dx = \int_\Omega |\nabla g_j - (1 + iq\varepsilon_j g_j)\mathbf{w}_j|^2 dx \le C_1.$$

So

$$\begin{aligned} \|\nabla g_j\|_{L^2(\Omega)} \le& \|\nabla g_j - \psi_j e^{-iq\mathbf{e}\cdot\mathbf{x}}\mathbf{w}_j\|_{L^2(\Omega)} + \|\psi_j e^{-iq\mathbf{e}\cdot\mathbf{x}}\mathbf{w}_j\|_{L^2(\Omega)} \\ \le& C_1 + \|\psi_j \mathbf{w}_j\|_{L^2(\Omega)} \le C_1 + \|\psi_j\|_{W^{1,2}(\Omega)} \|\mathbf{w}_j\|_{W^{1,2}(\Omega)} \le C_2. \end{aligned}$$

We now write

$$\tilde{g}_j = g_j - b_j, \qquad b_j = \fint_\Omega g_j dx.$$

To estimate $\tilde{g}_j$, we note that $\fint_\Omega \tilde{g}_j dx = 0$ and use the Poincaré inequality to get

$$\|\tilde{g}_j\|_{L^2(\Omega)} \le C\|\nabla \tilde{g}_j\|_{L^2(\Omega)} \le C_3.$$

Then we use the Sobolev inequality to get

$$\|\tilde{g}_j\|_{L^4(\Omega)} \le C\|\tilde{g}_j\|_{W^{1,2}(\Omega)} \le C_4. \tag{4.13}$$

Thus

$$\|\tilde{g}_j \mathbf{w}_j\|_{L^2(\Omega)} \le C\|\tilde{g}_j\|_{L^4(\Omega)} \|\mathbf{w}_j\|_{W^{1,2}(\Omega)} \le C_5.$$

To estimate b_j, we note that

$$\psi_j = e^{iq\mathbf{e}\cdot\mathbf{x}}(1+iq\varepsilon_j g_j) = \hat{\psi} + iq\varepsilon_j g_j e^{iq\mathbf{e}\cdot\mathbf{x}}.$$

So we have

$$iq\varepsilon_j g_j = e^{-iq\mathbf{e}\cdot\mathbf{x}}(\psi_j - \hat{\psi}) \to 0 \quad \text{in } L^4(\Omega);$$

hence

$$\varepsilon_j g_j \to 0 \quad \text{in } L^4(\Omega),$$

and

$$\varepsilon_j b_j = \varepsilon_j g_j - \varepsilon_j \tilde{g}_j = o(1). \tag{4.14}$$

Now we have

$$\begin{aligned}
&\|\nabla g_j - (1+iq\varepsilon_j g_j)\mathbf{w}_j\|_{L^(\Omega)} \\
=&\|\nabla \tilde{g}_j - (1+iq\varepsilon_j b_j)\mathbf{w}_j - iq\varepsilon_j \tilde{g}_j \mathbf{w}_j\|_{L^2(\Omega)} \\
=&\|\nabla \tilde{g}_j - (1+iq\varepsilon_j b_j)\mathbf{w}_j\|_{L^2(\Omega)} + O(\varepsilon_j \|\tilde{g}_j \mathbf{w}_j\|_{L^2(\Omega)}) \\
=&\|\nabla \tilde{g}_j - (1+iq\varepsilon_j b_j)\mathbf{w}_j\|_{L^2(\Omega)} + O(\varepsilon_j).
\end{aligned}$$

Set

$$u_j = \frac{\tilde{g}_j}{1+iq\varepsilon_j b_j}.$$

Then we use (4.14) to get

$$\begin{aligned}
&\|\nabla g_j - (1+iq\varepsilon_j g_j)\mathbf{w}_j\|^2_{L^(\Omega)} \\
=&|1+iq\varepsilon_j b_j|^2 \|\nabla u_j - \mathbf{w}_j\|^2_{L^2(\Omega)} + O(\varepsilon_j^2) \\
=&(1+o(1))\|\nabla u_j - \mathbf{w}_j\|_{L^2(\Omega)} + o(1).
\end{aligned}$$

Since the left-side term in the above equalities is bounded, we see that

$$\|\nabla u_j - \mathbf{w}_j\|_{L^2(\Omega)} \le C_5,$$

which implies

$$\|\nabla u_j\|_{L^2(\Omega)} \le C_6.$$

Since $\int u_j dx = 0$, we apply the Poincaré inequality to conclude that $\{u_j\}$ is bounded in $W^{1,2}(\Omega, \mathbb{C})$. Passing to a subsequence again we may assume that

$$u_j \to \hat{u} \quad \text{weakly in } W^{1,2}(\Omega, \mathbb{C}) \text{ and strongly in } L^2(\Omega).$$

Now we go back to (4.12) and find

$$\begin{aligned}
&(\chi H_s^2 + o(1)) \int_\Omega (\mathbf{h}\cdot\mathbf{w}_j)^2 dx \\
=&\int_\Omega \left\{ q^2(1+o(1))|\nabla u_j - \mathbf{w}_j|^2 + \frac{\kappa^2}{2} q^2 (-2\Im(g_j) + q\varepsilon_j |g_j|^2)^2 \right\} dx + \mathcal{F}[\mathbf{w}_j].
\end{aligned}$$

Hence

$$q^2\|\nabla u_j - \mathbf{w}_j\|^2_{L^2(\Omega)} + \mathcal{F}[\mathbf{w}_j] \le (\chi H_{\rm s}^2(\kappa, q) + o(1)) \int_\Omega (\mathbf{h}\cdot\mathbf{w}_j)^2 dx. \tag{4.15}$$

Letting $j \to \infty$ we find

$$q^2\|\nabla \hat{u} - \hat{\mathbf{w}}\|^2_{L^2(\Omega)} + \mathcal{F}[\hat{\mathbf{w}}] \le \chi H_{\rm s}^2(\kappa, q) \int_\Omega (\mathbf{h}\cdot\hat{\mathbf{w}})^2 dx. \tag{4.16}$$

Step 3. We claim that

$$\mathbf{h}\cdot\hat{\mathbf{w}} \not\equiv 0.$$

Suppose the claim were false. Then $\mathbf{h}\cdot\hat{\mathbf{w}} \equiv 0$ on Ω. Since $\mathbf{w}_j \to \hat{\mathbf{w}}$ strongly in $L^2(\Omega, \mathbb{R}^3)$, we have, as $j \to \infty$,

$$\int_\Omega (\mathbf{h}\cdot\mathbf{w}_j)^2 dx \to 0.$$

From (4.15) we have

$$\|\operatorname{div}\mathbf{w}_j\|_{L^2(\Omega)} \to 0, \qquad \|\operatorname{curl}\mathbf{w}_j\|_{L^2(\Omega)} \to 0,$$

and from (4.16),

$$\operatorname{div}\hat{\mathbf{w}} = 0 \quad \text{and} \quad \hat{\mathbf{w}} = \nabla\hat{u} \quad \text{in } \Omega, \qquad \hat{\mathbf{w}} = \mathbf{0} \quad \text{on } \partial\Omega.$$

Hence

$$\Delta\hat{u} = 0 \quad \text{in } \Omega, \qquad \nabla\hat{u} = \mathbf{0} \quad \text{on } \partial\Omega.$$

The maximum principle for harmonic functions yields that $\hat{u}$ is constant on Ω, and hence $\mathbf{w} = \mathbf{0}$. Thus

$$\mathbf{w}_j \to \mathbf{0} \quad \text{strongly in } L^2(\Omega, \mathbb{R}^3).$$

Since $\mathbf{w}_j \in W_0^{1,2}(\Omega, \mathbb{R}^3)$, from (3.10) we have

$$\|\mathbf{w}_j\|_{W^{1,2}(\Omega)} \le C(\|\operatorname{div}\mathbf{w}_j\|_{L^2(\Omega)} + \|\operatorname{curl}\mathbf{w}_j\|_{L^2(\Omega)} + \|\mathbf{w}_j\|_{L^2(\Omega)}) \to 0,$$

which contradicts the assumption $\|\mathbf{w}_j\|_{W^{1,2}(\Omega)} = 1$. So the claim is true.

Using the claim, (4.11), and (4.16) we have

$$\chi H_{\rm s}^2(\kappa, q) \ge \frac{q^2\|\nabla\hat{u} - \hat{\mathbf{w}}\|^2_{L^2(\Omega)} + \mathcal{F}[\hat{\mathbf{w}}]}{\int_\Omega (\mathbf{h}\cdot\hat{\mathbf{w}})^2 dx} \ge \chi H_{\rm sh}(q).$$

Thus we must have

$$H_{\rm s}(\kappa, q) = H_{\rm sh}(q).$$

In other words, under the assumption $H_{\rm s}(\kappa, q) < H_{\rm sh}(q)$, we must have $\mathbf{h}\cdot\hat{\mathbf{n}} \not\equiv 0$, and hence we use the result proved in Step 1 to conclude that $H_{\rm s}(\kappa, q)$ is achieved. Now conclusion (ii) is proved.

The proof of conclusion (iii) is similar to the proof of (ii) and hence is omitted. □

5. Magnetic field-induced instabilities in pure nematic states. In this section we examine the local minimality as well as global minimality of the pure nematic states given in (1.4). Note that $(0, \mathbf{n})$ is a critical point of the functional $\mathcal{E}$ if and only if $\mathbf{n}$ is a critical point of the functional $\mathcal{F}_{\sigma\mathbf{h}}$. Let

$$\begin{aligned} &C(\sigma) \equiv C(\sigma, K_1, K_2, \mathbf{h}, \mathbf{e}) = \inf_{\mathbf{n}\in W^{1,2}(\Omega,\mathbb{S}^2,\mathbf{e})} \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}], \\ &\mathcal{M}(\sigma) \equiv \mathcal{M}(\sigma, K_1, K_2, \mathbf{h}, \mathbf{e}) = \{\mathbf{n} \in W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e}) :\ \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}] = C(\sigma)\}. \end{aligned} \tag{5.1}$$

If $\mathbf{n} \in \mathcal{M}(\sigma)$, then $(0, \mathbf{n})$ is a critical point of $\mathcal{E}$.

We first look for a criterion for $\mathbf{n}$ to be a global minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$ in the set $W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e})$. If $\mathbf{n}$ is a minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$, then $\mathbf{n}$ satisfies the Euler–Lagrange equation

$$\begin{cases} -K_1 \nabla \operatorname{div} \mathbf{n} + K_2 \operatorname{curl}^2 \mathbf{n} = \chi\sigma^2(\mathbf{h}\cdot\mathbf{n})\mathbf{h} + \lambda(x)\mathbf{n} & \text{in } \Omega, \\ \mathbf{n} = \mathbf{e} \quad \text{on } \partial\Omega, \end{cases} \tag{5.2}$$

where $\lambda(x)$ is a function of x which is dependent on $\mathbf{n}$. By a simple computation we find

$$\begin{aligned} \lambda(x) &= \mathbf{n}\cdot[-K_1\nabla\operatorname{div}\mathbf{n} + K_2 \operatorname{curl}^2 \mathbf{n}] - \chi\sigma^2(\mathbf{h}\cdot\mathbf{n})^2 \\ &= K_2|\nabla\mathbf{n}|^2 + (K_1 - K_2)[(\operatorname{div}\mathbf{n})^2 - \operatorname{div}((\operatorname{div}\mathbf{n})\mathbf{n})] - \chi\sigma^2(\mathbf{h}\cdot\mathbf{n})^2. \end{aligned}$$

In the particular case where $K_1 = K_2 = K$, we have

$$\begin{cases} -\Delta\mathbf{n} = |\nabla\mathbf{n}|^2\mathbf{n} + \dfrac{\chi\sigma^2}{K}[(\mathbf{h}\cdot\mathbf{n})\mathbf{h} - (\mathbf{h}\cdot\mathbf{n})^2\mathbf{n}] & \text{in } \Omega, \\ \mathbf{n} = \mathbf{e} \quad \text{on } \partial\Omega. \end{cases} \tag{5.3}$$

Since $\mathbf{h}\cdot\mathbf{e} = 0$, the constant vector $\mathbf{e}$ is a critical point of $\mathcal{F}_{\sigma\mathbf{h}}$ for all σ. The following lemma gives a simple criterion for $\mathbf{n} = \mathbf{e}$ to be a global minimizer. Recall the numbers $H_{\mathrm{sh}}(0)$ and $H_{\mathrm{n}}(0)$ defined in (3.8) and (4.3), respectively, and recall that we have

$$0 < H_{\mathrm{n}}(0) \le H_{\mathrm{sh}}(0).$$

LEMMA 5.1. (i) *If $0 \le \sigma < H_{\mathrm{n}}(0)$, then $\mathbf{n} = \mathbf{e}$ is the only global minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$ in $W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e})$.*

(ii) *If $H_{\mathrm{n}}(0) < H_{\mathrm{sh}}(0)$ and if $H_{\mathrm{n}}(0) < \sigma < H_{\mathrm{sh}}(0)$, then $\mathbf{n} = \mathbf{e}$ is not a global minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$ in $W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e})$, but it is weakly stable (a local minimizer).*

(iii) *If $\sigma > H_{\mathrm{sh}}(0)$, then $\mathbf{n} = \mathbf{e}$ is not weakly stable.*

Lemma 5.1 shows that if K_1 and K_2 are fixed, the constant critical point $\mathbf{n} = \mathbf{e}$ remains as a global minimizer for small σ, and it loses the global minimality when σ increases and reaches $H_{\mathrm{n}}(0)$.

Proof of Lemma 5.1. We first note that $\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{e}] = 0$. If $\mathbf{n}$ is a global minimizer, then $\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}] \le \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{e}] = 0$. If $\mathbf{h}\cdot\mathbf{n} \equiv 0$, then from

$$\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}] = \int_\Omega \{K_1|\operatorname{div}\mathbf{n}|^2 + K_2|\operatorname{curl}\mathbf{n}|^2\}dx \le \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{e}] = 0$$

we see that $\operatorname{div}\mathbf{n} = 0$ and $\operatorname{curl}\mathbf{n} = \mathbf{0}$ on Ω. Since Ω is simply connected and $|\mathbf{n}| = 1$, $\mathbf{n}$ must be a constant vector, and it must be equal to $\mathbf{e}$ since $\mathbf{n} = \mathbf{e}$ on $\partial\Omega$.

Therefore a nonconstant global minimizer $\mathbf{n}$ must satisfy $\mathbf{h}\cdot\mathbf{n}\not\equiv 0$. From

$$0=\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{e}]\geq\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}]=\mathcal{F}[\mathbf{n}]-\chi\sigma^2\int_\Omega(\mathbf{h}\cdot\mathbf{n})^2dx,$$

we have

$$\sigma^2\geq\frac{1}{\chi}\frac{\mathcal{F}[\mathbf{n}]}{\int_\Omega(\mathbf{h}\cdot\mathbf{n})^2dx}\geq H_{\rm n}^2(0).$$

Thus if $0\leq\sigma<H_{\rm n}(0)$, then the only global minimizer is $\mathbf{n}=\mathbf{e}$.

Next, assume $\sigma>H_{\rm n}(0)$. Take $\tilde{\mathbf{n}}\in W^{1,2}(\Omega,\mathbb{S}^2,\mathbf{e})$ such that $\mathbf{h}\cdot\tilde{\mathbf{n}}\not\equiv 0$ on Ω and

$$\frac{1}{\chi}\frac{\mathcal{F}[\tilde{\mathbf{n}}]}{\int_\Omega(\mathbf{h}\cdot\tilde{\mathbf{n}})^2dx}<H_{\rm n}(0)^2+\delta<\sigma^2.$$

Then

$$\mathcal{F}_{\sigma\mathbf{h}}[\tilde{\mathbf{n}}]=\mathcal{F}[\tilde{\mathbf{n}}]-\chi\sigma^2\int_\Omega(\mathbf{h}\cdot\tilde{\mathbf{n}})^2dx<\mathcal{F}[\tilde{\mathbf{n}}]-\chi(H_{\rm n}(0)^2+\delta)\int_\Omega(\mathbf{h}\cdot\tilde{\mathbf{n}})^2dx<0.$$

Hence

$$\inf\{\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}]:\ \mathbf{n}\in W^{1,2}(\Omega,\mathbb{S}^2,\mathbf{e})\}<0,$$

and the constant map $\mathbf{n}=\mathbf{e}$ is not a global minimizer.

To finish the proof, we use the definition of weak stability given in section 2 to derive the conclusion about weak stability of $\mathbf{n}=\mathbf{e}$ for $0<\sigma<H_{\rm sh}(0)$ and instability for $\sigma>H_{\rm sh}(0)$. □

Let us consider next the following question: If $\mathbf{n}_\sigma\in\mathcal{M}(\sigma,K_1,K_2,\mathbf{h},\mathbf{e})$ is a global minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$, when is $(0,\mathbf{n}_\sigma)$ a global minimizer of the functional $\mathcal{E}$? To answer this question, let $\mu=\mu(q\mathbf{n})$ denote the lowest eigenvalue of the following equation:

$$-\nabla_{q\mathbf{n}}^2\phi=\mu\phi\quad\text{in }\Omega,\quad\nabla_{q\mathbf{n}}\phi\cdot\nu=0\quad\text{on }\partial\Omega. \tag{5.4}$$

Note that

$$\mu(q\mathbf{n})=\inf_{\phi\in W^{1,2}(\Omega,\mathbb{C})}\frac{\|\nabla_{q\mathbf{n}}\phi\|^2_{L^2(\Omega)}}{\|\phi\|^2_{L^2(\Omega)}}.$$

Define

$$\mu_*(q,\sigma)\equiv\mu_*(q,\sigma,K_1,K_2,\mathbf{h},\mathbf{e})=\inf_{\mathbf{n}\in\mathcal{M}(\sigma,K_1,K_2,\mathbf{h},\mathbf{e})}\mu(q\mathbf{n}). \tag{5.5}$$

LEMMA 5.2. (i) *If $(\psi,\mathbf{n})$ is a global minimizer of $\mathcal{E}$ which is not a pure nematic state, then $\mu(q\mathbf{n})<\kappa^2$.*

(ii) *If $\mu_*(q,\sigma)<\kappa^2$, then the pure nematic states are not global minimizers of $\mathcal{E}$.*

Proof. To prove (i), we note that if $(\psi,\mathbf{n})$ is a global minimizer, and if $\psi=0$, then we must have $\mathbf{n}\in\mathcal{M}(\sigma)$, and hence $(\psi,\mathbf{n})=(0,\mathbf{n})$ is a pure nematic state. Thus if $(\psi,\mathbf{n})$ is not a pure nematic state, then $\psi\not\equiv 0$. Choose $\mathbf{n}_\sigma\in\mathcal{M}(\sigma)$. We have

$$\mathcal{G}[\psi,\mathbf{n}]+\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}]\leq\mathcal{G}[0,\mathbf{n}_\sigma]+\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}_\sigma];$$

hence

$$\int_\Omega \left\{ |\nabla_{q\mathbf{n}}\psi|^2 - \kappa^2|\psi|^2 + \frac{\kappa^2}{2}|\psi|^4 \right\} dx \le \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}_\sigma] - \mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}] \le 0,$$

so

$$\int_\Omega \{|\nabla_{q\mathbf{n}}\psi|^2 - \kappa^2|\psi|^2\} dx \le -\frac{\kappa^2}{2}\int_\Omega |\psi|^4 dx < 0.$$

Thus

$$\mu(q\mathbf{n}) \le \frac{\|\nabla_{q\mathbf{n}}\psi\|^2_{L^2(\Omega)}}{\|\psi\|^2_{L^2(\Omega)}} < \kappa^2.$$

To prove (ii), note that for a pure smectic state $(0, \mathbf{n}_\sigma)$ we have $\mathcal{E}[0, \mathbf{n}_\sigma] = \frac{\kappa^2}{2}|\Omega|$. If $\mu_*(q,\sigma) < \kappa^2$, we choose $\mathbf{n} \in \mathcal{M}(\sigma)$ such that $\mu(q\mathbf{n}) < \kappa^2$ and let ϕ be an eigenfunction of $-\nabla^2_{q\mathbf{n}}$ associated with $\mu(q\mathbf{n})$. Then we choose $(t\phi, \mathbf{n})$ as a test function for $\mathcal{E}$ and find $\mathcal{E}[t\phi, \mathbf{n}] < \frac{\kappa^2}{2}|\Omega|$ if t is small. So a global minimizer $(\psi, \mathbf{n})$ must satisfy $\mathcal{E}[\psi, \mathbf{n}] < \frac{\kappa^2}{2}|\Omega|$, and hence it cannot be a pure nematic state. ∎

PROPOSITION 5.3. *If $0 \le \sigma \le H_{\mathrm{n}}(0)$ and $\kappa > 0$, or if $\sigma > H_{\mathrm{n}}(0)$ and $\mu_*(q,\sigma) < \kappa^2$, then the pure nematic states are not global minimizers of $\mathcal{E}$.*

Proof. We note that if $0 \le \sigma < H_{\mathrm{n}}(0)$, then from Lemma 5.1 we have $\mathcal{M}(\sigma) = \{\mathbf{e}\}$; and if $\sigma = H_{\mathrm{n}}(0)$, then we have $\mathbf{e} \in \mathcal{M}(\sigma)$. It is easy to see that $\mu(q\mathbf{e}) = 0$ with associated eigenfunction $\psi = e^{iq\mathbf{e}\cdot\mathbf{x}}$. Thus $\mu_*(q,\sigma) \le \mu(q\mathbf{e}) = 0 < \kappa^2$ for any $\kappa > 0$. Then the conclusion follows from Lemma 5.2(ii).

When $\sigma > H_{\mathrm{n}}(0)$ and $\mu_*(q,\sigma) < \kappa^2$, the conclusion was proved in Lemma 5.2. ∎

Note that the conclusion of Proposition 5.3 for the case $0 \le \sigma < H_{\mathrm{n}}(0)$ is a direct consequence of Lemma 4.2(iii). In fact, for any q we have $H_{\mathrm{s}}(\kappa, q) \ge H_{\mathrm{n}}(0)$. If $0 \le \sigma < H_{\mathrm{n}}(0)$, then $0 \le \sigma < H_{\mathrm{s}}(\kappa, q)$, and the pure smectic states are global minimizers with energy equal to zero. But the pure nematic state $(0, \mathbf{e})$ has energy $\frac{\kappa^2}{2}|\Omega|$, and so it cannot be a global minimizer.

Let us define

$$\sigma_*(\kappa, q) = \inf\{\sigma > 0 : \ \mu_*(q,\sigma) \ge \kappa^2\},$$
$$Q_*(\kappa, \sigma) = \inf\{q > 0 : \ \mu_*(q,\sigma) \ge \kappa^2\}.$$

From Proposition 5.3, when $\sigma > H_{\mathrm{n}}(0)$, we have the following statement about the global minimality of the pure nematic states.

Claim. In the following cases, the pure nematic states are not global minimizers of $\mathcal{E}$:

(i) $0 < \sigma < \sigma_*(\kappa, q)$, *for given κ and q.*
(ii) $\kappa^2 > \mu_*(q,\sigma)$, *for given q and σ.*
(iii) $0 \le q < Q_*(\kappa,\sigma)$, *for given κ and σ.*

Our next question arises in a comparison of liquid crystals with superconductivity. Recall that type II superconductors will always stay in the normal state if the applied magnetic field is sufficiently strong. Considering the analogies of liquid crystals to superconductors, one might expect that a liquid crystal would be in a pure nematic state if the applied magnetic field is sufficiently strong. However, this is not true. In fact we have the next theorem.

THEOREM 5.4. *Fix q, κ, K_1, K_2, $\mathbf{h}$, and $\mathbf{e}$, with $K_1 = K_2$. When σ is sufficiently large, the pure nematic states are not global minimizers.*

The conclusion of Theorem 5.4 reveals an *important difference in the responses to applied magnetic fields for smectic liquid crystals and for superconductors.* For superconductors, a sufficiently strong applied magnetic field will penetrate the material and destroy superconductivity. In contrast, smectic liquid crystals in an applied magnetic field also undergo phase transitions, but the applied magnetic field will not completely destroy the smectic structure, and the molecules of the liquid crystals will not be completely layered and aligned along the direction of the applied magnetic field.

To prove Theorem 5.4 we shall establish estimates of $C(\sigma)$ (see (5.1)) and $\mu_*(q,\sigma)$ (see (5.5)) for large σ.

LEMMA 5.5. (i) *For large σ we have*

$$C(\sigma) \le -\chi|\Omega|\sigma^2 + C_1\sigma, \tag{5.6}$$

where $C_1 > 0$ depends only on K_1, K_2, $\mathbf{h}$, $\mathbf{e}$, and Ω.

(ii) *Let $\mathbf{n}_\sigma$ be a global minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$. Then*

$$|\mathbf{h}\cdot\mathbf{n}_\sigma(x)| \to 1 \quad \text{in } L^2(\Omega) \text{ as } \sigma\to\infty.$$

(iii) *If $K_1 = K_2$, then $\mathbf{h}\cdot\mathbf{n}_\sigma(x)$ converges in $L^2(\Omega)$ to 1 or to -1 as $\sigma\to\infty$.*

Proof. After rotating the coordinate system we may assume that $\mathbf{h} = \mathbf{e}_3$ and $\mathbf{e} = \mathbf{e}_1$. Then for any $\mathbf{n} = (n_1, n_2, n_3) \in W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e})$ we may write

$$\mathcal{F}_{\sigma\mathbf{h}}[\mathbf{n}] = \mathcal{J}_\sigma[\mathbf{n}] - \chi|\Omega|\sigma^2,$$

where

$$\mathcal{J}_\sigma[\mathbf{n}] = \int_\Omega \{K_1|\operatorname{div}\mathbf{n}|^2 + K_2|\operatorname{curl}\mathbf{n}|^2 + \chi\sigma^2(n_1^2+n_2^2)\}dx. \tag{5.7}$$

Step 1. Proof of (i). We estimate the lower bound of $\mathcal{J}_\alpha$ for large σ. Let us consider a test map of the form

$$\mathbf{n} = (\cos\phi, 0, \sin\phi),$$

where ϕ is a smooth function and $\phi = 0$ on $\partial\Omega$. Hence $\mathbf{n} = \mathbf{e}_1$ on $\partial\Omega$. We compute

$$\operatorname{div}\mathbf{n} = \partial_1 n_1 + \partial_3 n_3 = -\sin\phi\partial_1\phi + \cos\phi\partial_3\phi,$$
$$\operatorname{curl}\mathbf{n} = (\partial_2 n_3, \partial_3 n_1 - \partial_1 n_3, -\partial_2 n_1) = (\cos\phi\partial_2\phi, -\sin\phi\partial_3\phi - \cos\phi\partial_1\phi, -\sin\phi\partial_2\phi).$$

So we have

$$\mathcal{J}_\sigma[\mathbf{n}] = J_\sigma[\phi] \equiv \int_\Omega f_\sigma(\phi)dx,$$

where

$$f_\sigma(\phi) = K_1|\sin\phi\partial_1\phi - \cos\phi\partial_3\phi|^2 + K_2|\partial_2\phi|^2 + K_2|\sin\phi\partial_3\phi + \cos\phi\partial_1\phi|^2 + \chi\sigma^2\cos^2\phi.$$

Note that

$$|\sin\phi\partial_1\phi - \cos\phi\partial_3\phi|^2 \le |\partial_1\phi|^2 + |\partial_3\phi|^2,$$
$$|\sin\phi\partial_3\phi + \cos\phi\partial_1\phi|^2 \le |\partial_1\phi|^2 + |\partial_3\phi|^2.$$

So

$$f_\sigma(\phi) \le K_1(|\partial_1\phi|^2 + |\partial_3\phi|^2) + K_2|\nabla\phi|^2 + \chi\sigma^2\cos^2\phi.$$

For small $\varepsilon > 0$, let

$$\Omega_\varepsilon = \{x \in \Omega : \ \mathrm{dist}(x, \partial\Omega) < \varepsilon\}, \qquad \Omega^\varepsilon = \{x \in \Omega : \ \mathrm{dist}(x, \partial\Omega) \ge \varepsilon\}.$$

Then we split $J_\sigma[\phi]$ into two parts:

$$J_\sigma[\phi] = J_{\sigma 1}[\phi] + J_{\sigma 2}[\phi],$$

where

$$J_{\sigma 1}[\phi] = \int_{\Omega_\varepsilon} f_\sigma(\phi)dx, \qquad J_{\sigma 2}[\phi] = \int_{\Omega^\varepsilon} f_\sigma(\phi)dx.$$

We choose ϕ such that

$$\phi = \frac{\pi}{2} \text{ in } \Omega^\varepsilon, \quad \phi = 0 \text{ on } \partial\Omega, \quad |\nabla\phi| \le \frac{C}{\varepsilon} \quad \text{on } \Omega.$$

Then $J_{\sigma 2}[\phi] = 0$, and

$$J_{\sigma 1}[\phi] \le \int_{\Omega_\varepsilon} \{(K_1 + K_2)|\nabla\phi|^2 + \chi\sigma^2\cos^2\phi\}dx \le \left[\frac{C^2(K_1 + K_2)}{\varepsilon^2} + \chi\sigma^2\right]|\Omega_\varepsilon|.$$

When $\partial\Omega$ is smooth, there exists $C_0 > 0$ only depending on $\partial\Omega$ such that for all small $\varepsilon > 0$ we have

$$|\Omega_\varepsilon| \le C_0\varepsilon.$$

Therefore

$$J_\sigma[\phi] = J_{\sigma 1}[\phi] \le \frac{C_0}{\varepsilon}[C^2(K_1 + K_2) + \chi\sigma^2\varepsilon^2].$$

For large σ we choose

$$\varepsilon = \frac{C}{\sigma}\sqrt{\frac{K_1 + K_2}{\chi}}$$

and find

$$J_\sigma[\phi] \le 2\sigma C_0 C\sqrt{\chi(K_1 + K_2)}.$$

This yields estimate (5.6).

Step 2. *Proof of* (ii). From (i) we have

$$\mathcal{J}_\sigma[\mathbf{n}_\sigma] = \int_\Omega \{K_1|\mathrm{div}\,\mathbf{n}_\sigma|^2 + K_2|\mathrm{curl}\,\mathbf{n}_\sigma|^2 + \chi\sigma^2|\mathbf{n}'_\sigma|^2\}dx \le C\sigma,$$

where $\mathbf{n}'_\sigma = (n_{\sigma 1}, n_{\sigma_2}, 0)$. Thus, as $\sigma \to \infty$,

$$\int_\Omega |\mathbf{n}'_\sigma|^2 dx \le \frac{C}{\sqrt{\sigma}} \to 0,$$

namely,

$$\int_\Omega (1-|n_{\sigma 3}|^2)dx \to 0.$$

So $|n_{\sigma 3}| \to 1$ in $L^1(\Omega)$. Since $|n_{\sigma 3}(x)| \le 1$ a.e., we use the Lebesgue dominated convergence theorem to conclude that $|n_{\sigma 3}| \to 1$ in $L^p(\Omega)$ for any $1 < p < \infty$.

Step 3. Now we assume $\mathbf{h} = \mathbf{e}_3$, $\mathbf{e} = \mathbf{e}_1$, and $K_1 = K_2 = K$. We prove that the minimizer $\mathbf{n}_\sigma$ has the following property: $\mathbf{n}_{\sigma 3}$ does not change sign, namely, either $\mathbf{n}_{\sigma 3}$ is always positive, or it is always negative, or it is identically equal to zero.

Now the Euler–Lagrange equation reads

$$\begin{cases} -\Delta \mathbf{n} = |\nabla \mathbf{n}|^2 \mathbf{n} + b^2\sigma^2[n_3\mathbf{e}_3 - n_3^2\mathbf{n}] & \text{in } \Omega, \\ \mathbf{n} = \mathbf{e}_1 & \text{on } \partial\Omega, \end{cases} \tag{5.8}$$

where

$$b^2 = \frac{\chi}{K}.$$

To investigate the behavior of the minimizers for large σ, we borrow the proof from [HL, Lemma 2.2]. For simplicity we drop the subscript σ and let $\mathbf{n} = (n_1, n_2, n_3)$ denote a minimizer. Set $\mathbf{u} = (n_1, n_2, |n_3|)$. Since $n_3 \in W^{1,2}(\Omega)$, we have $|\nabla |n_3|| = |\nabla n_3|$ a.e., and so $|n_3| \in W^{1,2}(\Omega)$. Note that $n_3 = 0$ on $\partial\Omega$. Hence $\mathbf{u} \in W^{1,2}(\Omega, \mathbb{S}^2, \mathbf{e})$, and

$$\int_\Omega |\nabla \mathbf{u}|^2 dx = \int_\Omega |\nabla \mathbf{n}|^2 dx.$$

Since $K_1 = K_2 = K$, we have

$$\mathcal{J}_\sigma[\mathbf{n}] = \int_\Omega \{K|\nabla \mathbf{n}|^2 + \chi\sigma^2(n_1^2 + n_2^2)\}dx.$$

Thus $\mathbf{u}$ is also a global minimizer of $\mathcal{F}_{\sigma\mathbf{h}}$, and hence it is a weak solution of (5.8). In particular, $u_3 = |n_3| \ge 0$, and it is a weak solution of

$$\begin{cases} -\Delta u_3 = |\nabla \mathbf{n}|^2 u_3 + b^2\sigma^2(u_3 - u_3^3) & \text{in } \Omega, \\ u_3 = 0 & \text{on } \partial\Omega. \end{cases}$$

Thus

$$-\Delta u_3 = |\nabla \mathbf{n}|^2 u_3 + b^2\sigma^2(1 - u_3^2)u_3 \ge 0.$$

So u_3 satisfies

$$\Delta u_3 \le 0 \quad \text{in } \Omega, \quad u_3 = 0 \quad \text{on } \partial\Omega.$$

Using the weak Harnack inequality for bounded nonnegative superharmonic functions, we have

$$\operatorname{ess\,inf}_{B_{\theta R}} u_3 \ge c\left(\fint_{B_R} u_3^p dx\right)^{1/p} \quad \text{for all } B_R \subset \Omega,$$

where p is some positive number. Thus if $u_3 \not\equiv 0$, then $u_3 > 0$ in Ω, namely, either $n_3 > 0$ everywhere in Ω or $n_3 < 0$ everywhere in Ω.

Step 4. *Proof of* (iii). Assume $K_1 = K_2 = K$. From Step 2 we see that, for large σ, $n_{\sigma 3} \not\equiv 0$; from Step 3, $n_{\sigma 3}$ does not change its sign. We may assume $n_{\sigma 3} > 0$. Then, as $\sigma \to \infty$, $n_{\sigma 3} \to 1$ in $L^2(\Omega)$ and $\mathbf{n}'_\sigma \to \mathbf{0}$ in $L^2(\Omega, \mathbb{R}^3)$. Thus $\mathbf{n}_\sigma \to \mathbf{e}_3$ in $L^2(\Omega, \mathbb{R}^3)$. □

Proof of Theorem 5.4. Assume $K_1 = K_2 = K$. Without loss of generality we assume $\mathbf{h} = \mathbf{e}_3$ and $\mathbf{e} = \mathbf{e}_1$. Let $\mathbf{n}_\sigma$ be a global minimizer of $\mathcal{F}_{\sigma \mathbf{h}}$. We estimate $\mu(q\mathbf{n}_\sigma)$ for large σ. From Lemma 5.5, we may pass to a subsequence and assume that $\mathbf{n}_\sigma \to \mathbf{e}_3$ strongly in $L^2(\Omega, \mathbb{R}^3)$. Take $\phi = e^{iqx_3}$. We have, as $\sigma \to \infty$,

$$\int_\Omega |\nabla_{q\mathbf{n}_\sigma} \phi|^2 dx = q^2 \int_\Omega |\mathbf{n}_\sigma - \mathbf{e}_3|^2 dx \to 0,$$

$$\mu(q\mathbf{n}_\sigma) \le \frac{\|\nabla_{q\mathbf{n}_\sigma} \phi\|^2_{L^2(\Omega)}}{\|\phi\|^2_{L^2(\Omega)}} = \frac{q^2 \|\mathbf{n}_\sigma - \mathbf{e}_3\|^2_{L^2(\Omega)}}{|\Omega|} \to 0.$$

So for any $\kappa > 0$, we have $\mu(q\mathbf{n}_\sigma) < \kappa^2$ for all large σ. Hence $\mu_*(q, \sigma) < \kappa^2$ for all large σ. From Lemma 5.2 we conclude that for any given κ, there exists $\hat{\sigma}$ depending on κ such that, for all $\sigma > \hat{\sigma}$, the pure nematic states are not global minimizers. □

Acknowledgments. The authors would like to thank the referees for valuable comments on the manuscript.

REFERENCES

[AS1] R. J. Atkin and I. W. Stewart, *Freedericksz transitions in spherical droplets of smectic C liquid crystals*, Quart. J. Mech. Appl. Math., 47 (1994), pp. 231–245.

[AS2] R. J. Atkin and I. W. Stewart, *Theoretical studies of Freedericksz transitions in SmC liquid crystals*, European J. Appl. Math., 8 (1997), pp. 253–262.

[B] H. Brezis, *Liquid crystals and energy estimates for S^2-valued maps*, in Theory and Applications of Liquid Crystals, J. L. Ericksen and D. Kinderlehrer, eds., IMA Vol. Math. Appl. 5, Springer-Verlag, Berlin, 1987, pp. 31–52.

[BCLP] P. Bauman, M. Calderer, C. Liu, and D. Phillips, *The phase transition between chiral nematic and smectic A^* liquid crystals*, Arch. Ration. Mech. Anal., 165 (2002), pp. 161–186.

[BP] P. Bates and X. B. Pan, *Nucleation of Instability in Meissner State of 3-Dimensional Superconductors*, preprint.

[BPT] P. Bauman, D. Phillips, and Q. Tang, *Stable nucleation for the Ginzburg-Landau system with an applied magnetic field*, Arch. Ration. Mech. Anal., 142 (1998), pp. 1–43.

[BS] A. Bernoff and P. Sternberg, *Onset of superconductivity in decreasing fields for general domains*, J. Math. Phys., 39 (1998), pp. 1272–1284.

[Ca] M. C. Calderer, *Studies of layering and chirality of smectic A^* liquid crystals*, Math. Comput. Modelling, 34 (2001), pp. 1273–1288.

[Ch1] S. J. Chapman, *Nucleation of superconductivity in decreasing fields.* I, II, European J. Appl. Math., 5 (1994), pp. 449–468; 469–494.

[Ch2] S. J. Chapman, *Superheating field of type* II *superconductors*, SIAM J. Appl. Math., 55 (1995), pp. 1233–1258.

[CL] R. Cohen and M. Luskin, *Field-induced instabilities in nematic liquid crystals*, in Nematics, J.-M. Coron et al., eds., NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 332, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991, pp. 261–278.

[CT] R. Cohen and M. Taylor, *Weak stability of the map $x/|x|$ for liquid crystal functionals*, Comm. Partial Differential Equations, 15 (1990), pp. 675–692.

[dG] P. G. de Gennes, *An analogy between superconductors and smectics A*, Solid State Commun., 10 (1972), pp. 753–756.

[dGP] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed., Oxford Science Publications, Oxford, UK, 1993.

[DL] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 3, Springer-Verlag, New York, 1990.

[E] J. L. Ericksen, *Hydrostatic theory of liquid crystals*, Arch. Ration. Mech. Anal., 9 (1962), pp. 371–378.

[H] F. Helein, *Minima de la fonctionelle energie libre des cristaux liquides*, C. R. Acad. Sci. Paris, 305 (1987), pp. 565–568.

[HKL] R. Hardt, D. Kinderlehrer, and F. H. Lin, *Existence and partial regularity of static liquid crystal configurations*, Comm. Math. Phys., 105 (1986), pp. 547–570.

[HL] F. B. Hang and F. H. Lin, *Static theory for planar ferromagnets and antiferromagnets*, Acta Math. Sin. (Engl. Ser.), 17 (2001), pp. 541–580.

[HM1] B. Helffer and A. Morame, *Magnetic bottles in connection with superconductivity*, J. Funct. Anal., 185 (2001), pp. 604–680.

[HM2] B. Helffer and A. Morame, *Magnetic bottles for the Neumann problem: Curvature effects in the case of dimension* 3 *(general case)*, Ann. Sci. École Norm. Sup., 37 (2004), pp. 105–170.

[HP] B. Helffer and X. B. Pan, *Upper critical field and location of surface nucleation of superconductivity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 145–181.

[K] M. Kleman, *Points, Lines, and Walls in Liquid Crystals, Magnetic Systems, and Various Ordered Media*, John Wiley, Chichester, New York, 1983.

[L1] F. H. Lin, *Nonlinear theory of defects in nematic liquid crystals, phase transitions and flow phenomena*, Comm. Pure Appl. Math., 42 (1989), pp. 789–814.

[L2] F. H. Lin, *On nematic liquid crystals with variable degree of orientation*, Comm. Pure Appl. Math, 44 (1991), pp. 453–468.

[L3] F. H. Lin, *Solutions of Ginzburg-Landau equations and critical points of the renormalized energy*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 599–622.

[LD] F.-H. Lin and Q. Du, *Ginzburg–Landau vortices: Dynamics, pinning, and hysteresis*, SIAM J. Math. Anal., 28 (1997), pp. 1265–1293.

[LL] F. H. Lin and C. Liu, *Static and dynamic theories of liquid crystals*, J. Partial Differential Equations, 14 (2001), pp. 289–330.

[LP1] K. Lu and X. B. Pan, *Estimates of the upper critical field for the Ginzburg-Landau equations of superconductivity*, Phys. D, 127 (1999), pp. 73–104.

[LP2] K. Lu and X. B. Pan, *Surface nucleation of superconductivity in* 3 *dimensions*, J. Differential Equations, 168 (2000), pp. 386–452.

[P1] X. B. Pan, *Surface superconductivity in applied magnetic fields above* H_{C_2}, Comm. Math. Phys., 228 (2002), pp. 327–370.

[P2] X. B. Pan, *Landau-de Gennes model of liquid crystals and critical wave number*, Comm. Math. Phys., 239 (2003), pp. 343–382.

[P3] X. B. Pan, *Surface superconductivity in* 3 *dimensions*, Trans. Amer. Math. Soc., 356 (2004), pp. 3899–3937.

[P4] X.-B. Pan, *Landau–De Gennes model of liquid crystals with small Ginzburg–Landau parameter*, SIAM J. Math. Anal., 37 (2006), pp. 1616–1648.

[PK] X. B. Pan and K. Kwek, *On a problem related to vortex nucleation of superconductivity*, J. Differential Equations, 182 (2002), pp. 141–168.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1613–1641

A MATHEMATICAL ANALYSIS OF THE OPTIMAL EXERCISE BOUNDARY FOR AMERICAN PUT OPTIONS*

XINFU CHEN† AND JOHN CHADAM†

Abstract. We study a free boundary problem arising from American put options. In particular we prove existence and uniqueness for this problem, and we derive and rigorously prove high order asymptotic expansions for the early exercise boundary near expiry. We provide four approximations for the boundary: one is explicit and is valid near expiry (weeks); two others are implicit involving inverse functions and are accurate for longer time to expiry (months); the fourth is an ODE initial value problem which is very accurate for all times to expiry, is extremely stable, and hence can be solved instantaneously on any computer. We further provide an ode iterative scheme which can reach its numerical fixed point in five iterations for all time to expiry. We also provide a large time (equivalent to regular expiration times but large interest rate and/or volatility) behavior of the exercise boundary. To demonstrate the accuracy of our approximations, we present the results of a numerical simulation.

Key words. early exercise boundary, existence and uniqueness, numerical and analytical approximations

AMS subject classifications. 91B28, 35R35, 45G05

DOI. 10.1137/S0036141003437708

1. Introduction.

With the Black–Scholes hypothesis of log-normal stock prices, the price $P(S,T)$ for an American put option on a share of price S at time T can be formulated as the solution to following free boundary problem (cf. [27]):

$$(P)\quad \begin{cases} P_T + \frac{1}{2}\sigma^2 S^2 P_{SS} + r\,S\,P_S - r\,P = 0 & \text{for } T < T_F,\ S > S_f(T), \\ P(S,T) = E - S,\quad P_S(S,T) = -1 & \text{for } T < T_F,\ S \le S_f(T), \\ S_f(T_F) = E,\quad P(S,T_F) = \max\{0, E-S\} & \text{for } T = T_F,\ S > 0. \end{cases}$$

Here E is the exercise (strike) price, T_F the expiration time, σ the constant volatility, r the constant risk-free interest rate, and $S = S_f(T)$ the free boundary separating regions of optimally holding and exercising.

There is a considerable literature on the optimal exercise boundary, both analytical and numerical; see, for example, [1, 2, 3, 4, 6, 14, 15, 16, 18, 20, 21, 24, 25, 26] and the references therein. A recent list of references, together with numerical approximations, can be found in [1, 8, 24].

For notational simplicity, we write problem (P) in a nondimensional form. Let

$$k = 2r/\sigma^2,\quad S = E\,e^x,\quad T = T_F - 2t/\sigma^2\ ,\quad P(S,T) = E\,p(x,t),\quad S_f(T) = E\,e^{s(t)}.$$

Then problem (P) becomes, for the transformed price $p(x,t)$ and the optimal exercise

*Received by the editors November 13, 2003; accepted for publication (in revised form) July 7, 2006; published electronically January 26, 2007.
http://www.siam.org/journals/sima/38-5/43770.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu+@pitt.edu, chadam+@pitt.edu). The first author was supported by National Science Foundation grants DMS-9971043, 0203991, and 0504691. The second author was supported by National Science Foundation grant DMS-9704567.

boundary $x = s(t)$,

$$
(1.1)\qquad \begin{cases} p_t - p_{xx} - (k-1)p_x + k\,p = 0 & \text{for } t > 0,\ x > s(t), \\ p(x,t) = 1 - e^x,\ p_x(x,t) = -e^x & \text{for } t > 0,\ x \le s(t), \\ s(0) = 0, \qquad p(x,0) = \max\{1 - e^x, 0\} & \text{for } t = 0,\ x \in \mathbb{R}, \\ p(x,t) > \max\{1 - e^x, 0\} & \text{for } t > 0,\ x > s(t). \end{cases}
$$

The last condition corresponds to the physical condition $P > E - S$ when $S > S_f(T)$. Though not necessary, we include this condition in (1.1) to make the definition of the free boundary clearer.

Unlike the American call option with dividend where $s(t) \sim -2\sqrt{t\alpha}$ with α being a constant [27] when the (continuous) dividend rate is less than the risk-free interest rate, here for the put option, α becomes unbounded as $t \searrow 0$ (i.e., $T \nearrow T_F$), leading to difficulties in the theoretical analysis, numerical simulation, and accurate pricing and strategic trading during this extremely volatile period, i.e., the period in which the relation between the asset and option prices is rapidly varying.

Although the analysis to be presented is quite technical, the high accuracy of the ensuing global estimates for the location of the early exercise boundary is important for practitioners. Knowing the location of the early exercise boundary a priori makes the pricing of American style financial derivatives amenable by Monte Carlo simulation, which is the preferred systematic method of fund managers with thousands of instruments. In addition to the practical importance of these estimates, the technical methods to obtain them are also of theoretical interest. Since the methods do not use the convexity of the free boundary (we have proven the convexity of the free boundary for problem (P) in a separate paper [7]), they serve as a prototype for problems with nonconvex free boundary problems. We expect this to be the most likely case in finance since even for the closely related problem (P) on a dividend-paying asset, numerical simulations by Detemple suggest (private communication) that the early exercise boundary may not be convex for all choices of the parameters.

In recent developments, Kuske and Keller [18], Bunch and Johnson [5], and Stamicar et al. [25] derived independently the following similar asymptotic expansions for $\alpha(t) := s^2(t)/(4t)$:

$$
\begin{aligned}
&\text{(KK)} && 9\pi k^2\, t\, \alpha^2\, e^{2\alpha} &&\sim 1\ , \\
&\text{(BJ)} && 4k^2\, t\, \alpha\, e^{2\alpha} &&\sim 1 - k^2/[2(1+k)^2]\ , \\
&\text{(SSC)} && 4\pi k^2\, t\, e^{2\alpha} &&\sim 1
\end{aligned}
$$

for all sufficiently small positive t. Regardless of their differences, all these asymptotics capture the dominant behavior $\lim_{t\searrow 0} \frac{2\alpha(t)}{|\log t|} = 1$ that was rigorously established by Barles et al. [4]. Nevertheless, any two of the asymptotics (KK), (BJ), and (SSC) cannot hold simultaneously at the next order. For example, by taking logarithms, (SSC) implies that $2\alpha(t) \sim |\log(t)| - \log(4\pi k^2)$, while (KK) implies $2\alpha(t) \sim |\log(t)| - \log(9\pi k^2)$ with additional log log corrections from the α^2 term. Similarly, (BJ) implies yet another constant term along with log log corrections.

On the other hand, due to the singularity of problem (1.1) near the origin, numerical simulations are very difficult, and typical methods such as the binomial or trinomial tree methods can hardly capture any asymptotic behavior of $\alpha(t)$ more accurately than the above approximations.

One purpose of this paper is to give a complete and rigorous mathematical justification to show that indeed (SSC) is the correct asymptotic behavior of $\alpha(t)$ as $t \searrow 0$. In addition, we shall prove rigorously that as $t \searrow 0$, $\alpha(t) = s^2(t)/(4t)$ has the more general asymptotic expansion

(1.2)
$$\alpha(t) = -\xi - \frac{1}{2\xi} + \frac{1}{8\xi^2} + \frac{17}{24\xi^3} - \frac{51}{64\xi^4} - \frac{287}{120\xi^5} + \frac{199}{32\xi^6} + O(\xi^{-7}), \quad \xi := \log\sqrt{4\pi k^2 t}.$$

Due to our particular choice of ξ, this expansion does not have a constant term and also does not depend otherwise on any parameters. This differentiates it from the behavior obtained in (KK) and (BJ) that give other constants and log log rather than inverse log corrections. Please see Remark 3.2 and section 7.2 for further discussions of this point.

Another purpose of this paper is to provide the following noniterative approximations to $s(t)$ for both small and large t:

$$\text{(expl)} \quad \alpha = -\xi - \frac{1}{2(\xi - a)} + \frac{1/8 + a/2}{(\xi - a)^2}, \qquad a = 0.96621\ldots,$$

$$\text{(imp1)} \quad \xi = -\alpha - \log\left\{1 - \frac{1}{2(\alpha+1)} - \frac{1}{2(\alpha+1)^2}\right\},$$

$$\text{(imp2)} \quad \xi = -\alpha - \log\left\{\tfrac{2}{\sqrt{\pi}}\textstyle\int_0^{\sqrt{\alpha}} e^{-z^2} dz\right\} + \log\tfrac{e^{\alpha} + 2k\log(1+1/k)\, e^{1/\alpha}}{e^{\alpha} + e^{1/\alpha}},$$

$$\text{(ODE)} \quad \tfrac{d}{dt}s(t) = \tfrac{s(t)}{2kt}\Gamma(s(t), t), \qquad \Gamma(z,t) := \tfrac{1}{2\sqrt{\pi t}} e^{-z^2/(4t) - (k-1)z/2 - (k+1)^2 t/4}.$$

There have also been many contributions to the study of early exercise boundaries for American options with dividends; see, for example, Evans et al. [11] and Knessl [17]. An earlier theoretical work using a variational approach for American options with multiple assets as well as a numerical algorithm for the pricing problem was supplied by Jaillet et al. [15]. By contrast, the main focus of this paper is to give a complete treatment, with particular attention on the singular behavior of the optimal exercise boundary near expiry, for the simplest nontrivial case of the American put without dividends. It is expected that analysis similar to ours can be carried over, with appropriate modifications, to the case with other payoffs, dividends, and/or multiple assets. On the other hand, as mentioned earlier, even in the closely related case of problem (P) on a dividend-paying asset, the dependence of the near expiring behavior, and possibly the convexity, on the choice of parameters suggests that the necessary modifications may be subtle.

The explicit approximation (expl) and the first implicit approximation (imp1) are derived directly from the asymptotic expansion (1.2); they are fourth order in the sense that for small t, the α values calculated from (expl) or (imp1) have error of order $O(|\xi|^{-4})$. Our numerical simulation (cf. [8]) shows that both (expl) and (imp1) are far better than any straightforward truncations of (1.2) (assuming $T_F - T$ is larger than 1 second) both in accuracy and in the length of interval of validity of the formulas. For our running example (cf. Figure 7.1), where $E = 1$, $r = 0.1/\text{year}$ and $\sigma = 0.25/\sqrt{\text{year}}$, the approximation (expl) is accurate for $T_F - T$ less than several weeks and (imp1) is accurate for $T_F - T$ less than several months.

The second implicit approximation (imp2) is an interpolation of the small time behavior $\alpha \approx -\xi$ and large time behavior $s(t) \approx \log[(1+k)/k]$ derived from Merton's solution for the infinite horizon problem for American put [22]. In general (imp2) is

better than (imp1). For our running example, the error of the approximation (imp2) is less than 2×10^{-3} for $T_F - T$ up to 3 years.

The ODE approximation is to be solved with an initial condition compatible with the limit $\alpha + \xi \to 0$ as $\xi \to -\infty$. In numerical implementation, it is transformed to an equation for α in the $\xi = \log\sqrt{4\pi k^2 t}$ variable and the initial condition is approximated by $\alpha|_{\xi=\xi_0} = -\xi_0 - 1/(2\xi_0)$, where ξ_0 is a large negative number, say, $\xi_0 = -10$. Numerical simulation shows that this ODE initial value problem is very stable and highly insensitive to any change of initial conditions, and hence can be solved instantaneously on any computer. The ODE approximation is better than any of the above three. For our running example, its error is less than 5×10^{-5} when $T_F - T$ is less than 2 months, 10^{-3} when $T_F - T$ less than 1 year, and 6×10^{-3} for all $T_F - T > 0$. We would like to point out that our ODE approximation has already surpassed those numerical approximations from the standard binomial or trinomial tree methods (with 1000 division points), which are typically used in literature as the "exact" solutions for comparisons; see the curve marked Bino in Figure 7.1.

The ODE approximation is derived from the following exact system:

$$\text{(1.3)} \qquad \begin{cases} \dot{s}(t) = \frac{s(t)}{2kt}\,\Gamma(s(t),t)\big\{1 + m(t)\big\}, \\ m(t) = k\displaystyle\int_0^t \left\{\frac{s(t)-s(\tau)}{t-\tau}\,\frac{2t}{s(t)} - 1\right\}\frac{\Gamma(s(t)-s(\tau),t-\tau)}{\Gamma(s(t),t)}\,d\,s(\tau). \end{cases}$$

From this system, we obtain an iterative scheme. Starting with the ODE approximation (corresponding to $m \equiv 0$), successively solve (1.3) with m evaluated at the previous iteration of s. As it turns out, this iteration converges very rapidly; a numerical fixed point (difference less than 10^{-7}) is obtained after only five iterations. The first iteration takes less than 1 minute and the total of five iterations takes less than 10 minutes (on a Sparc server). See Figure 7.1 for the error estimate of the first three iterations.

Note that $t = \sigma^2(T_F - T)/2 = r(T_F - T)/k$ is large when σ and/or r are large. Hence, to include cases where r and/or σ are large, we also provide a long time behavior of s. For large t,

$$\text{(long)} \qquad s(t) \sim s_\infty \exp\left\{\hat{m}\int_{(k+1)^2 t/4}^{\infty} \rho^{-3/2} e^{-\rho}\,d\rho\right\},$$

$$s_\infty = s(\infty) = \log[k/(1+k)],$$

$$\hat{m} = \frac{k+1}{4\sqrt{\pi}}\int_0^\infty \frac{s(\tau)}{s_\infty}\exp\left\{\frac{k-1}{2}(s(\tau) - s_\infty) + \frac{(k+1)^2}{4}\tau\right\} d\,s(\tau).$$

Here $\hat{m}$ can be calculated approximately by using the ODE approximation for s, which is instantaneous since we can do so by solving ODE. When (long) is incorporated with our noniterative schemes such as ODE, we can instantaneously obtain reliable approximate values of $s(t)$ for any t and any parameters r and σ; see Figure 7.1 for $r = 0.1/\text{year}$ and $\sigma = 0.25/\sqrt{\text{year}}$, and see [8] for other values of the parameters.

This paper is organized as follows. In section 2, we briefly establish, for mathematical completeness, the well-posedness of problem (1.1) via a classical variational approach [12]. We show that the solution (p, s) to (1.1) exists and is unique, that $s(t)$ is continuous and nondecreasing, and as $t \to \infty$, $(s(t), p(\cdot, t)) \to (s_\infty, p_\infty(\cdot))$, the solution to the infinite horizon problem [22]. During the review and revision of this manuscript, an alternative proof of the existence and uniqueness appeared in the literature [23].

In section 3 we derive several integral and integrodifferential equations for s by using the fundamental solution Γ for the linear parabolic PDE for p; in particular we derive (1.3).

Sections 4 through 6 are devoted to showing that (1.3) has a solution $s(\cdot)$ with $\alpha := s^2(t)/(4t)$ satisfying the asymptotic behavior (1.2). In section 4, we transform (1.3) into an equation of the form

$$(\mathbf{I} + \mathbf{L})[u'] + G(u, \xi) = F[u], \tag{1.4}$$

where $u = u(\xi) = \alpha(t) = s^2(t)/(4t)$, $\xi = \log\sqrt{4\pi k^2 t}$, $\mathbf{I}$ is the identity operator, $\mathbf{L}$ a linear operator defined in (4.5), $F[u]$ a small nonlinear operator, and G a function. In section 5 we show that the operator $(\mathbf{I}+\mathbf{L})$ is invertible from C^0 to C^0 and that $\mathbf{L}[\phi]$ is always 1/2 more differentiable than ϕ, although $\|\phi - \mathbf{L}[\phi]\|_{C^0((-\infty,\xi])} \to 0$ as $\xi \to -\infty$ for any uniformly continuous function ϕ. In section 6, we first establish the existence of a unique solution to (1.4), with $F[u]$ replaced by any known small function, in a finite interval $\xi \in (-j, \xi_0]$ for any j and some fixed negative large constant ξ_0. To take the limit $j \to \infty$, we show that (1.4), in a finite interval $[-j, \xi_0]$ or half finite interval $(-\infty, \xi_0]$, possesses a comparison principle, which allows us to construct sub and super solutions to sandwich the solutions. We let $j \to \infty$ to obtain solutions of (1.4) with given known F. A Schauder's fixed point theorem then can be used to establish the existence of a solution to (1.4). Uniqueness of the solution follows from the well-posedness result of section 3. The asymptotic expansion (1.2) is proved by the comparison principle and construction of sub and super solutions.

Finally, in section 7 we derive our approximation formula mentioned earlier and, for the purpose of illustration, provide a numerical simulation to support the advantages of our new approximations.

We repeat that recently we have shown [7] that the optimal boundary is convex; see also [10]. Using this property, many of the proofs here can be greatly simplified. Nevertheless, the method provided here is general enough to be extended to many similar option problems where the optimal exercise boundaries may not be convex.

2. Well-posedness of problem (P). In this section we establish the well–posedness of the free boundary problem (1.1). Though many techniques and results are standard and may be cited from references, we still provide a certain degree of detail for completeness of the paper.

For convenience, we denote by $\mathfrak{L}$ the operator

$$\mathfrak{L}[p] = p_{xx} + (k-1)p_x - k\,p.$$

LEMMA 2.1. *Let $p(x,t)$, together with a free boundary $x = s(t)$, be a solution to* (1.1). *Then*

$$\begin{cases} \min\{p - p_0, p_t - \mathfrak{L}[p]\} = 0 & \text{in } \ \mathbb{R} \times (0,\infty), \\ p(x,0) = p_0(x) := \max\{1 - e^x, 0\} & \forall\ x \in \mathbb{R}. \end{cases} \tag{2.1}$$

The proof follows from a straightforward verification and is omitted.

THEOREM 2.2. *There exists a unique solution p to* (2.1). *Define $s(t) = \sup\{x | p(x,t) = p_0(x)\}$ for all $t > 0$. Then* (i) *$s(\cdot)$ is a strictly decreasing continuous function on $(0,\infty)$,* (ii) *$\lim_{t\searrow 0} s(t) = 0$,* (iii) *$p(x,t) > p_0(x)$ for all $x > s(t)$ and $t > 0$, and $p(x,t) = p_0(x)$ for all $x \le s(t)$ and $t \ge 0$, and* (iv) *(p, s) solves* (1.1).

Proof. The existence of a unique solution p follows from a well-developed parabolic theory for obstacle problems; see, for example, [12, Chap. 1, sec. 8]. Here for completeness and for the existence of $s(\cdot)$, we provide the main idea of the proof.

Uniqueness. Let p_1 and p_2 be arbitrary two solutions to (2.1). Denote $\gamma_i = (\partial_t - \mathfrak{L})p_i \geq 0$. Then

$$(p_1 - p_2)\{(p_1 - p_2)_t - \mathfrak{L}[(p_1 - p_2)]\} = (p_1 - p_2)(\gamma_1 - \gamma_2) \leq 0 \quad \text{in } \mathbb{R} \times (0, \infty)$$

since $\gamma_1 = 0$ when $p_1 - p_2 > 0$ (as $p_1 > p_2 \geq p_0$) and $\gamma_2 = 0$ when $p_1 - p_2 < 0$. Integrating the above inequality over $x \in \mathbb{R}$ and using the Gronwall's inequality, one concludes that $p_1 \equiv p_2$.

Existence. For every $\varepsilon > 0$, let $q^\varepsilon(x,t)$ be the solution to the semilinear parabolic Cauchy problem

$$\begin{cases} q^\varepsilon_t - \mathfrak{L}[q^\varepsilon] = \beta_\varepsilon(q^\varepsilon - p^\varepsilon_0) & \text{in } \mathbb{R} \times (0, \infty), \\ q^\varepsilon(\cdot, 0) = p^\varepsilon_0(\cdot) := \rho^\varepsilon * p_0 & \text{on } \mathbb{R} \times \{0\}, \end{cases}$$

where $\rho^\varepsilon(z) := \varepsilon^{-1}\rho(\varepsilon^{-1}z)$ with $\rho(\cdot)$ being a smooth and nonnegative mollifier of unit integral over $\mathbb{R}^1$, and $\beta_\varepsilon(\cdot)$ is any nonnegative, bounded, and smooth function defined on $\mathbb{R}$ with the properties

$$\beta'_\varepsilon(z) \leq 0 \ \text{ for all } \ z \in \mathbb{R}, \qquad \beta_\varepsilon(0) = k, \quad \text{and } \ \beta_\varepsilon(z) = 0 \ \text{ for } \ z > \varepsilon.$$

Existence of a unique smooth solution q^ε follows from standard parabolic PDE theory; see, for example, [13]. To take the limit $\varepsilon \to 0$ to obtain a solution to (2.1), we need to establish a few ε-independent a priori estimates for q^ε.

Differentiating the differential equation with respect to t gives $(\partial_t - \mathfrak{L} - \beta'_\varepsilon)q^\varepsilon_t = 0$ in $\mathbb{R} \times (0, \infty)$. When $t = 0$, $q^\varepsilon_t(\cdot, 0) = \mathfrak{L}[p^\varepsilon_0] + \beta_\varepsilon(0) \geq 0$ in $\mathbb{R}^1$ since $\beta_\varepsilon(0) = k$ and in the distributional sense $\mathfrak{L}[p_0] \geq -k$ which implies $\mathfrak{L}[p^\varepsilon_0] = \rho^\varepsilon * \mathfrak{L}[p_0] \geq -k$. Therefore, by comparison, $q^\varepsilon_t > 0$ on $\mathbb{R} \times [0, \infty)$.

Also one can show that p^ε_0 is a subsolution and $\mathbf{1}$ is a supersolution so that $p^\varepsilon_0 < q^\varepsilon < 1$ in $\mathbb{R} \times (0, \infty)$.

Note that $p^\varepsilon_0 < q^\varepsilon$ implies $\beta^\varepsilon(q^\varepsilon - p^\varepsilon_0) \in [0, k)$ on $\mathbb{R} \times (0, \infty)$. Consequently, by local PDE regularity estimates, the set $\{q^\varepsilon\}_{0<\varepsilon<1}$ is bounded in $C^{\beta,\beta/2}(\mathbb{R} \times [0, \infty)) \cap W^{2,1}_r((-R, R) \times (0, R) \setminus (-\delta, \delta) \times (0, \delta))$ for every $\beta \in (0, 1)$, $r > 1$, $\delta > 0$, and $R > \delta$. Hence, there exist $\gamma \in L^\infty(\mathbb{R} \times (0, \infty))$ and $p \in C^{\beta,\beta/2}(\mathbb{R} \times [0, \infty)) \cap W^{2,1}_{r,\mathrm{loc}}(\mathbb{R} \times [0, \infty) \setminus (0, 0))$ such that, along some sequence $\varepsilon \searrow 0$, $\beta_\varepsilon(q^\varepsilon - p^\varepsilon_0) \to \gamma$ weakly in $L^r(B_R(0) \times (0, R))$, and $q^\varepsilon \to p$ strongly in $C^{\beta,\beta/2}([-R, R] \times [0, R])$ and weakly in $W^{2,1}_r((-R, R) \times (0, R) \setminus ((-\delta, \delta) \times (0, \delta))$ for every $r > 1, \beta \in (0, 1), \delta > 0$ and $R > \delta$. Taking the limit of the differential equation for q^ε along that convergent sequence, we conclude that $p(\cdot, 0) = p_0(\cdot)$ and

$$p_t - \mathfrak{L}[p] = \gamma \in [0, k], \quad p \geq p_0, \quad p_t \geq 0 \qquad \text{in } \mathbb{R} \times (0, \infty).$$

Since $q^\varepsilon \to p$ locally uniformly, $p(x_0, t_0) - p_0(x_0) > 0$ implies $q^\varepsilon > p^\varepsilon_0 + \varepsilon$, i.e., $\beta^\varepsilon(q^\varepsilon - p^\varepsilon) = 0$, in a ε-independent neighborhood of (x_0, t_0) for all sufficiently small positive ε in the sequence, and therefore, $\gamma = 0$ in a neighborhood of (x_0, t_0). Thus p is a solution to (2.1).

The free boundary. Let $\mathbf{C} := \{(x,t) \in \mathbb{R} \times [0, \infty) \mid p(x,t) = p_0(x)\}$ be the contact set in the obstacle problem terminology. Since $p_t \geq 0$, there exists a

(semicontinuous) function $T : \mathbb{R} \to [0, \infty) \cup \{\infty\}$, such that $\mathbf{C} = \{(x, t) \mid x \in \mathbb{R}^1, 0 \le t \le T(x)\}$.

Since $p_t - \mathfrak{L}[p] = \gamma \ge 0$ in $\mathbb{R} \times (0, \infty)$, a comparison principle implies that $p > 0$ in $\mathbb{R} \times (0, \infty)$. It then follows that $T(x) = 0$ for all $x > 0$ since $p_0(x) = 0$ for all $x \ge 0$.

Now we show that $T(\cdot)$ is nonincreasing. Indeed, if $T(x_0) > 0$, then defining a new function $\tilde{p}$ by $\tilde{p} = p$ in $[x_0, \infty) \times [0, T(x_0)]$ and $\tilde{p} = p_0$ in $(-\infty, x_0] \times [0, T(x_0)]$, one can verify that $\tilde{p}$ is a solution to (2.1) in $\mathbb{R} \times [0, T(x_0)]$, so that, by uniqueness, $p = \tilde{p}$. Consequently, $(-\infty, x_0] \times [0, T(x_0)] \in \mathbf{C}$, and therefore $T(x) \ge T(x_0)$ for all $x \le x_0$. Thus $T(\cdot)$ is nonincreasing.

Next we show that $T(x)$ is strictly decreasing for $x \le 0$. Suppose this is not true. Then for some $x_2 < x_1 \le 0$, $T(x_2) = T(x_1) < \infty$. Consequently, $p(\cdot, t_0) = p_0(\cdot)$ in $[x_2, x_1]$ and $p > p_0$ in $(x_2, \infty) \times (T(x_2), \infty)$. It then follows that $p_t - \mathfrak{L}[p] = \gamma \equiv 0$ in $(x_2, \infty) \times (T(x_2), \infty)$. Since p_0 is smooth in (x_2, x_1), so is p in $(x_2, x_1) \times [T(x_2), \infty)$. Thus $p_t(\frac{x_1+x_2}{2}, T(x_2)) = \mathfrak{L}[p_0](\frac{x_1+x_2}{2}) = -k$, contradicting $p_t \ge 0$. Hence, $T(x)$ is strictly decreasing on $(-\infty, 0]$.

It then follows that the function $t = T(x)$ for $x \le 0$ admits an inverse $x = s(t)$ defined for all $t \ge 0$ and is nondecreasing. As inverse functions of strictly monotonic functions are continuous, $s(\cdot)$ is continuous. Note that $T(x) = 0$ for $x > 0$ and $T(x) > 0$ for $x < 0$ implies that $s(0) = 0$.

Finally we verify that $s(t)$ is strictly decreasing. In fact, if $s(t)$ is a constant over an interval $[t_1, t_2]$, then $p_t - \mathfrak{L}[p] = 0$ in $(s(t_1), \infty) \times (t_1, t_2)$ and $p(s(t_1), t) = p_0(s(t_1))$ for all $t \in [t_1, t_2]$, so that $p \in C^\infty([s(t_1), \infty) \times (t_1, t_2))$. As $p_t \ge (\not\equiv) 0$ and $(\partial_t - \mathfrak{L})p_t = 0$ in $[s(t_1), \infty) \times (t_1, t_2)$, the Hopf lemma then gives $p_{tx} > 0$ on $\{s(t_1)+\} \times (t_1, t_2)$, which implies that $p_x(s(t)+, t)$ is strictly increasing for $t \in (t_1, t_2)$. On the other hand, $p \in W^{2,1}_{r,loc}(\mathbb{R} \times (0, \infty))$ for any $r > 1$ and the definition of $s(\cdot)$ implies that $p_x(s(t), t) = p_{0x}(s(t_1))$ is a constant for all $t \in (t_1, t_2)$, and we have a contradiction. Hence, $s(t)$ is strictly decreasing. This completes the proof. □

THEOREM 2.3. *There exists a unique solution (p, s) to (1.1). In addition, $p \in W^{2,1}_{r,loc}(\mathbb{R} \times [0, \infty) \setminus (0, 0))$ for any $r > 1$, and $p > 0$, $p_t \ge 0$, and $p_x < 0$ in $\mathbb{R} \times (0, \infty)$. Furthermore, $s(\cdot)$ is continuous and strictly decreasing and as $t \to \infty$,*

$$s(t) \to s_\infty := \log(1 + 1/k), \tag{2.2}$$

$$p(x, t) \to p_\infty(x) := \begin{cases} 1 - e^x & \text{if } x \le s_\infty, \\ (1 - e^{s_\infty})e^{-k(x - s_\infty)} & \text{if } x > s_\infty. \end{cases} \tag{2.3}$$

Proof. We need only show the assertions (2.2), (2.3), and $p_x < 0$ in $\mathbb{R} \times (0, \infty)$ since the rest follows from Lemma 2.1 and the proof of Theorem 2.2. As $\gamma = k$ for $x < s(t)$ and $= 0$ for $x > s(t)$, $(\partial_t - \mathfrak{L})[p_x] = \gamma_x \le 0$ in the distributional sense. A strong maximum principle then implies that $p_x < 0$ in $\mathbb{R} \times (0, \infty)$. It remains to show (2.2) and (2.3).

First of all, we can use comparison to show that $p(\cdot, t) < p_\infty(\cdot)$ for all $t \ge 0$. This inequality implies, by the definition of s, that $s(t) > s_\infty$ for all $t > 0$.

Next, as we know that $p_t(\cdot, \cdot) \ge 0$ and $s(\cdot)$ is strictly decreasing, the existence of an upper bound $p_\infty(\cdot)$ for $p(\cdot, t)$ and a lower bound s_∞ for $s(\cdot)$ then implies that the limits $p^*(\cdot) = \lim_{t\to\infty} p(\cdot, t)$ and $s^* = \lim_{t\to\infty} s(t)$ exist. From the differential equation, we can derive that $\mathfrak{L}[p^*] = 0$ in (s^*, ∞), $p^* = p_0$ in $(-\infty, s^*]$, and $p^* \in W^2_{r,loc}(\mathbb{R})$ for any $r > 1$. Solving for (p^*, s^*) from these relations we find that $s^* = s_\infty$ and $p^*(\cdot) = p_\infty(\cdot)$. This completes the proof. □

Remark 2.1. The limit $(s_\infty, p_\infty(\cdot))$ is the classical solution of Merton [22] for the infinite horizon problem for American puts.

Remark 2.2. That $s(t)$ is not differentiable at $t = 0$ is due to the nonsmoothness of the initial data $p(\cdot, 0) = p_0 = \max\{1 - e^x, 0\}$. To see this, consider the hodograph transformation: Let $x = X(z, t)$ be the inverse function of $z = p_x(x, t) + e^x$. Then $s(t) = X(0, t)$, and $X(z, t)$ solves the following initial Neumann boundary value problem for a quasi-linear parabolic PDE:

$$\begin{cases} X_t - \frac{1}{X_z^2} X_{zz} + (k-1) - kzX_z = 0 \quad \text{for } z > 0, t > 0, \\ X_z(0, t) = 1/k \quad \text{for } \ t > 0, \\ X(\cdot, 0) = \max\{0, \log(z)\} \quad \forall z > 0. \end{cases} \tag{2.4}$$

This problem is highly singular since $X(z, 0) = 0$ for $z \in (0, 1)$, which is due to the fact that $-p_{0x}(0)$ has a jump from 0 to 1.

It is important to note that if the initial data $X(z, 0) = \max\{0, \log(z)\}$ is replaced by $X^\varepsilon(z, 0)$, a smooth function with positive slope, the resulting solution satisfies $X_z^\varepsilon > 0$ on $[0, \infty) \times [0, \infty)$. Our numerical simulation (not shown here) indicates that a numerical scheme based on (2.4) is much better than the binomial or trinomial tree method.

From (2.4), we see that if for some $\beta > 0$, $s(t)$ is $C^{1+\beta}$ near some $t = t_0 > 0$, then $s(\cdot) \in C^\infty((t_0, \infty))$. Indeed, s in $C^{1+\beta}$ near t_0 implies $p(x, t)$ in $C^{2+\beta, 1+\beta/2}$ in a neighborhood of $\mathbb{R} \times \{t_0\}$, so that $X(\cdot, t_0)$ is in $C^{1+\beta}([0, \infty))$. With a more detailed a priori estimate on the previous ε problem one can show that $(p_x + e^x)_x > 0$ so $X_z(\cdot, t_0) > 0$ on $[0, \infty)$. A maximum principle for the equation for X_z can be used to show that $X_z > 0$ on $[0, \infty) \times [t_0, \infty)$. Since the Neumann boundary value $X_z(0, t) = 1/k$ is a constant function, a boot strap argument and a classical local Hölder estimate for quasi-linear parabolic equation (see for example, [19]) then implies that $X \in C^\infty([0, \infty) \times (t_0, \infty))$. In particular, $s(t) = X(0, t)$ is also in $C^\infty((t_0, \infty))$.

In the subsequent sections, we shall show (by a totally different method) that $s(\cdot) \in C^2((0, \delta))$ for some $\delta > 0$, so that $s \in C^\infty((0, \infty))$.

3. Integral representation for the free boundary $x = s(t)$. In this section, we use the Green's representation for solutions of the linear parabolic PDE in (1.1) to derive, for the free boundary $x = s(t)$, several integral and integrodifferential equations, including (1.3), which is to be solved later to establish the asymptotic behavior of $s(t)$ for small positive t.

We denote by $\Gamma(x, t)$ the fundamental solution to the operator $\partial_t - \mathfrak{L}$; more precisely,

$$\begin{aligned} \Gamma(x, t) &= \frac{1}{2\sqrt{\pi t}} \exp\left\{ -\frac{[x + (k-1)t]^2}{4t} - kt \right\} \\ &= \frac{1}{2\sqrt{\pi t}} \exp\left\{ -\frac{[x + (k+1)t]^2}{4t} + x \right\}. \end{aligned} \tag{3.1}$$

Since $(\partial_t - \mathfrak{L})[p] = \gamma$ in $L^r_{loc}(\mathbb{R} \times (0, \infty))$ for any $r > 1$ and $\gamma = k$ for $x < s(t)$ and $\gamma = 0$ for $x > s(t)$, the Green's identity gives, for the unique solution (p, s) of (1.1),

$$p(x, t) = \int_{-\infty}^{0} (1 - e^y)\Gamma(x - y, t)\, dy + k \int_0^t \int_{-\infty}^{s(t-\tau)} \Gamma(x - y, \tau)\, dy d\tau \quad x \in \mathbb{R}, t > 0. \tag{3.2}$$

It is worth mentioning that the first integral on the right-hand side is the price for the European put option because $e^{kt}\Gamma(x - y, t)dy$ is the probability that at expiry

the stock price (after scaling) is y, for which the option has value $e^{-kt}\max\{1-e^y,0\}$. Consequently, the second integral in (3.2) is the extra value (premium) of the American put option over the European put option, if the option is exercised optimally (i.e., exercise the option as soon as the (scaled) stock price s is below $s(t)$).

LEMMA 3.1. *Let $s\in C^0([0,\infty))\cap C^1((0,\infty))\cap W^{1,1}((0,1))$ be any function and $p(x,t)$ be defined as in (3.2). Set $p_0(x)=\max\{1-e^x,0\}$. Then for all $t>0$ and $x\neq 0$ and $x\neq s(t)$,*

$$
\begin{aligned}
(3.3)\quad p(x,t) &= p_0(x)+\int_0^t\left\{\Gamma(x,\tau)-k\int_{s(t-\tau)}^0\Gamma(x-y,\tau)dy\right\}d\tau,\\
p_x(x,t) &= p_{0x}(x)+\int_0^t\Big\{\Gamma_x(x,\tau)+k\,\Gamma(x,\tau)-k\,\Gamma(x-s(t-\tau),\tau)\Big\}d\tau,\\
p_t(x,t) &= \Gamma(x,t)+k\int_0^t\Gamma(x-s(t-\tau),\tau)\dot s(t-\tau)\,d\tau,\\
p_{xx}(x,t) &= p_{0xx}(x)+\Gamma(x,t)+\int_0^t\Big\{\Gamma_x(x,\tau)+k\Gamma(x,\tau)-k\,\Gamma_x(x-s(t-\tau),\tau)\Big\}d\tau,\\
p_{xt}(x,t) &= \Gamma_x(x,t)+k\int_0^t\Gamma_x(x-s(t-\tau),\tau)\dot s(t-\tau)\,d\tau.
\end{aligned}
$$

Consequently, (p,s) solves (1.1) if and only if s satisfies one of the following equations, for all $t>0$:

$$
\begin{aligned}
(3.4)\quad &\int_0^t\Gamma(s(t),\tau)d\tau=k\int_0^t\int_{s(t-\tau)}^0\Gamma(s(t)-y,\tau)dyd\tau,\\
(3.5)\quad &\int_0^t\Big\{\Gamma_x(s(t),\tau)+k\,\Gamma(s(t),\tau)\Big\}d\tau=k\int_0^t\Gamma(s(t)-s(t-\tau),\tau)\,d\tau,\\
(3.6)\quad &\Gamma(s(t),t)=-k\int_0^t\Gamma(s(t)-s(t-\tau),\tau)\dot s(t-\tau)\,d\tau,\\
(3.7)\quad &\Gamma(s(t),t)=\frac{k}{2}+k\int_0^t\Big\{\Gamma_x(s(t)-s(t-\tau),\tau)-\Gamma(s(t)-s(t-\tau),\tau)\Big\}\,d\tau,\\
(3.8)\quad &\dot s(t)=-\frac{2\Gamma_x(s(t),t)}{k}-2\int_0^t\Gamma_x(s(t)-s(t-\tau),\tau)\dot s(t-\tau)\,d\tau.
\end{aligned}
$$

THEOREM 3.2. *Let $s\in C^1((0,\infty))\cap C^0([0,\infty))$ be nonpositive and $\alpha(t)=s^2(t)/(4t)$. Assume that as $t\searrow 0$, $\alpha(t)=[-1+o(1)]\log\sqrt{t}$ and $t\,\dot\alpha(t)=O(1)$. Then s, together with p defined in (3.2), solve (1.1) if and only if s satisfies the integrodifferential equation, for all $t>0$,*

$$
(3.9)\qquad \dot s(t)=\frac{s(t)\Gamma(s(t),t)}{2kt}+\int_0^t\Big\{\frac{s(t)-s(t-\tau)}{\tau}-\frac{s(t)}{2t}\Big\}\Gamma(s(t)-s(t-\tau),\tau)\dot s(t-\tau)\,d\tau.
$$

One notices that (3.9) is exactly equivalent to (1.3).

Remark 3.1. Since $\Gamma_x(x,t)=-\frac{x+(k-1)t}{2t}\Gamma(x,t)$, adding (3.8) and (3.6) multiplied by $\frac{\lambda s(t)+2(k-1)t}{2kt}$ gives

$$
(3.10)\qquad \dot s(t)=\frac{(2-\lambda)s\Gamma(s,t)}{2kt}+\int_0^t\Big\{\frac{s(t)-s(t-\tau)}{\tau}-\frac{\lambda s(t)}{2t}\Big\}\Gamma(s(t)-s(t-\tau),\tau)\dot s(t-\tau)\,d\tau.
$$

Setting $\lambda = 1$ gives (3.9) or (1.3). We choose the particular value $\lambda = 1$ is to make the integral as small as possible, because the most significant contribution of the integral comes from small τ and when τ is small, $\frac{s(t)-s(t-\tau)}{\tau} \approx s'(t) = \frac{s(t)}{2t}[1 + \frac{t\dot{\alpha}(t)}{\alpha}] \approx \frac{s(t)}{2t}$. Indeed, the cancellation is even stronger than this. A linear combination of equations (3.6)–(3.8) shows that the integral on the right-hand side of (3.9) is equal to

$$\int_0^t \Big\{ \frac{\dot{s}(t) + \dot{s}(t-\tau)}{2} \frac{s(t) - s(t-\tau)}{\tau} - \dot{s}(t)\dot{s}(t-\tau) + \tfrac{k+1}{2}\dot{s}(t) \Big\} \Gamma(s(t) - s(t-\tau), \tau)\, d\tau.$$

Due to the strong cancellation of the first two terms in the integrand, the ratio $m(t)$ of the integral and the first term on the right-hand side of (3.9) can be expanded as $0 + 0\xi^{-1} + 0\xi^{-2} + \frac{1}{4}\xi^{-3} + O(\xi^{-4})$, where $\xi = \log[\sqrt{4\pi k^2 t}]$. Thus, we can drop the integral in (3.9) to obtain the ODE in section 1 approximating $s(t)$ accurately for small as well as large t (when t is large, $\dot{s}(t) = O(t^{-3/2}e^{-(k+1)^2 t/4})$ is exponentially small).

Remark 3.2. From (3.4) one can immediately obtain a rough estimate for $s(t)$ for small t. In fact, since $\int_{-\infty}^0 \Gamma(y,t)\, dy = \frac{1}{2}e^{-kt}$, the double integral in (3.4) can be written as $\theta(t)\, kt$ with $\theta(t) \in (0, 1/2)$. Also since $\Gamma(s(t), t) = \frac{1+o(1)}{\sqrt{4\pi t}} e^{-s(t)^2/(4t)}$ for small t, $\int_0^t \Gamma(s(t), \tau)\, d\tau = \frac{[1+o(1))]\sqrt{t\,\alpha}}{\sqrt{\pi}} \int_{\sqrt{\alpha}}^\infty \eta^{-2} e^{-\eta^2}\, d\eta$, where $\alpha = s^2/(4t)$. Thus (3.4) gives $\lim_{t\searrow 0} \alpha(t) = \infty$. Consequently, $\int_{\sqrt{\alpha}}^\infty \eta^{-2} e^{-\eta^2} d\eta = \frac{1}{2}\alpha^{-3/2} e^{-\alpha}(1 + O(\alpha^{-1}))$ and, from (3.4),

$$\alpha^{-1} e^{-\alpha}[1 + O(\alpha^{-1})] = \sqrt{4\pi k^2 t}\, \theta(t)\ .$$

Hence α is of order at least $O(|\log t|)$. One can further calculate, assuming $\alpha = [-1 + o(1)] \log \sqrt{t}$, that $\theta = \alpha^{-1}(1 + o(1))$. It then follows that $\alpha = |\log \sqrt{4\pi k^2 t}|(1 + o(1))$, a conjecture first made correctly in [25]. It is worth mentioning here that $\theta(t) \approx \alpha^{-1}$ eliminates any $\log|\log t|$ corrections (as suggested in [18] and [5]) to the leading order approximation $\alpha \approx -\log[\sqrt{t}]$ for small t.

The smallness of $\theta(t)$ results from the strong cancellation of the integral $k\int_0^t \int_{-\infty}^{s(t-\tau)} \Gamma(x - y, \tau) dy d\tau$ which represents the extra value of the American put over the European put, and the integral $-k\int_0^t \int_{-\infty}^0 \Gamma(x - y, \tau) dy d\tau = p_0 * \Gamma(\cdot, t) - \{p_0 + \int_0^t \Gamma(x, \tau) d\tau\}$ which relates to that part of the premium added on to the European put to account for the possibility that the future stock price drops below x. It seems to us that this strong cancellation was not observed in [18], resulting in $\log|\log t|$ terms appearing in their expansion of α.

The asymptotic behavior $\alpha = -\log\sqrt{4\pi k^2 t} + o(1)$ for small t can also be similarly derived from (3.5).

Remark 3.3. Equation (3.6) or (3.7) can be used to derive an interesting and highly nontrivial limit: $\lim_{t\to 0} \Gamma(s(t), t) = k$. Indeed, using the change of variable $\eta = \frac{s(t-\tau)-s(t)}{2\sqrt{\tau}}$, one obtains from (3.6)

$$\Gamma(s(t), t) = k[\ 1 + o(1)\ ] \int_0^{\sqrt{\alpha}} \frac{2e^{-\eta^2}}{\sqrt{\pi}} \Big(2 - \frac{s(t) - s(t-\tau)}{\dot{s}(t-\tau)\tau}\Big)^{-1} d\eta = k + o(1)$$

since when τ/t is small, $\frac{s(t)-s(t-\tau)}{\dot{s}(t-\tau)\tau} \approx 1$, whereas when τ/t is not small, $\eta \gg 1$ so that $\frac{s(t)-s(t-\tau)}{\dot{s}(t-\tau)\tau}$ can be replaced by 1 as an approximation.

Remark 3.4. A system exactly equivalent to (3.7) was derived in [25] and was used to derive formally (SSC) in section 1. The system was also used to obtain accurate approximations of s for small t, via an iteration scheme: starting with $s \equiv 0$, update s by solving (3.7) with the right-hand side evaluated at a previous s. Nevertheless, this scheme does not seem to converge, although its first several iterations converge rapidly (for small t); for more details, see [8, 25]. This is one of our reasons for deriving (3.9) and using it to analyze $s(t)$ theoretically and also numerically.

Remark 3.5. The nonpositivity assumption on s for the integrodifferential equation in (3.9) is important since if s is a solution, then when $k = 1$, $\hat{s} = -s$ is also a solution.

Proof of Lemma 3.1. Since $\Gamma(\cdot, 0)$ is the Delta function,

$$\int_{-\infty}^{0} (1 - e^y)\Gamma(x - y, t)\, dy = p_0(x) + \int_0^t \int_{-\infty}^{0} p_0(y)\Gamma_\tau(x - y, \tau) dy d\tau$$
$$= p_0(x) + \int_0^t \left\{ \Gamma(x, \tau) - k \int_{-\infty}^{0} \Gamma(x - y, \tau)\, dy \right\} d\tau$$

by using $\Gamma_\tau(x - y, \tau) = \Gamma_{xx} + (k - 1)\Gamma_x - k\Gamma$ and integrating by parts. Substituting this identity into (3.2) we obtain (3.3). The rest of the equations, for p_x, p_t, p_{xx}, and p_{xt}, follow by differentiating (3.3) (a substitution $\Gamma_{xx} + k\Gamma_x = \Gamma_\tau + \Gamma_x + k\Gamma$ is needed for p_{xx}). We remark that all the integrals are convergent due to the regularity assumption we made on s. It remains to show the second part of the lemma.

First we assume that (p, s) solves (1.1). Then $p(x, t) - p_0(x)$, as well as all its derivatives, vanish when $x < s(t)$. Thus, letting $x \nearrow s(t)$ we obtain from the equations for $p, p_x, p_t, p_{xx} - p_x$, and p_{xt} the corresponding equations asserted. Here, in taking the limits for p_{xx} and p_{xt}, we need the following fact: for any continuous function f,

$$\lim_{x \to s(t)\pm} \int_0^t \Gamma_x(x - s(t - \tau), \tau) f(t - \tau)\, d\tau = \mp \frac{f(t)}{2} + \int_0^t \Gamma_x(s(t) - s(t - \tau), \tau) f(t - \tau)\, d\tau.$$

Next we assume that s satisfies one of the equations in the second part of the lemma and show that (p, s) solves (1.1). First we notice that p satisfies $p(\cdot, 0) = p_0(\cdot)$ and for $t > 0$, $p_t - \mathfrak{L}[p] = 0$ for $x > s(t)$, and $= k$ for $x < s(t)$. In addition, $p \in W^{2,1}_{r,loc}(\mathbb{R} \times [0, \infty) \setminus \{(0, 0)\})$ for any $r > 0$. From the equations we derived for p, p_x, p_t, p_{xx}, and p_{xt}, we see that the equations in the second part of the lemma are, respectively, equivalent to the conditions $p = p_0$, $p_x = p_{0x}$, $p_t = 0$, $p_{xx} - p_x = p_{0xx} - p_{0x}$, and $p_{xt} = 0$ at $(s(t)-, t)$ for all $t > 0$. Each of these conditions provides, by the uniqueness of solutions of the initial boundary value problem of the parabolic equation $p_t - \mathfrak{L}[p] = k$ in the set $\{(x, t) \mid x \le s(t), t > 0\}$, that $p \equiv p_0$ in the set, which implies that (p, s) satisfies all the equations (1.1). Also, (p, s) satisfies all the other integral equations we derived.

It remains to show that $p > p_0$ when $x > s(t)$. That $p \in W^{2,1}_r$ for any r implies that $p \in C^{1+\beta,(\beta+1)/2}$ so $p_0 = p$ and $p_{0x} = p_x$ on $x = s(t)$. Now the function $q := (p - p_0)_x$ satisfies $q_t - \mathfrak{L}[q] = 0$. In addition, $q = 0$ and $q_x = k > 0$ on $x = s(t)$. Using lap number theory, one concludes that the number of zeroes of $q(\cdot, t)$ is nonincreasing. One can show that for sufficiently small positive t, $q(\cdot, t) = 0$ has exactly one root in $(s(t), \infty)$, so that $q(\cdot, t) = 0$ has at most one root for all $t > 0$. Consequently, $q(\cdot, t) = 0$ has exactly one root in $(s(t), \infty)$. Thus, starting from $s(t)$, $p(\cdot, t) - p_0(\cdot, t)$ first increases, then decreases to zero as $x \to \infty$; i.e. $p > p_0$ for all $x > s(t)$. This completes the proof. ∎

Proof of Theorem 3.2. Assume that (p, s) solves (1.1). Then $p_{xt} = p_t = 0$ at $(s(t)-, t)$ for all $t > 0$. Equation (3.9) then follows from $p_{xt} + \frac{s(t)+2(k-1)t}{4t} p_t = 0$ at $(s(t)-, t)$.

Now we assume that s satisfies (3.9) and show that (p, s) solves (1.1). We need only show that $p_t = 0$ for all $x < s(t)$. Note that p defined in (3.2) is smooth (enough in our subsequent analysis) in the set $\{(x,t) \mid x \le s(t)-, t > 0\}$, and (3.9) implies $p_{xt} + \frac{s(t)+2(k-1)t}{4t} p_t = 0$ at $(s(t)-, t)$ for all $t > 0$. Since $\frac{s(t)+2(k-1)t}{4t} \sim -\frac{\sqrt{\alpha(t)}}{\sqrt{4t}}$ is negative and p_t is singular near the origin, we cannot directly apply a standard parabolic PDE theory to conclude that $p_t = 0$ for $x < s(t)$ and all $t \ge 0$.

Differentiating the equation $p_t - \mathfrak{L}[p] = k$ with respect to t, multiplying the resulting equation by p_t, and integrating over $(-\infty, s(t))$ we obtain, after integration by parts and the substitution $p_{xt} = -\frac{s(t)+2(k-1)t}{4t} p_t$ at the boundary $x = s(t)-$,

$$\frac{1}{2}\frac{d}{dt}\int_{-\infty}^{s(t)} p_t^2 dx + \int_{-\infty}^{s(t)} (p_{xt}^2 + k\, p_t^2) dx$$
$$= \Big\{ \frac{\dot s(t)}{2} - \frac{s(t)}{4t} \Big\} p_t^2(s(t)-, t) = -\frac{\sqrt{t}\,\dot\alpha(t)}{2\sqrt{\alpha(t)}} p_t^2(s(t)-, t)$$

by the definition $\alpha(t) = s^2/(4t)$. Using $p_t^2(s(t)-, t) = \int_{-\infty}^{s(t)} 2 p_t p_{xt} dx \le \int_{-\infty}^{s(t)} (\delta p_t^2 + \delta^{-1} p_{xt}^2) dx$ with $\delta = |\frac{t\dot\alpha(t)}{2\sqrt{\alpha(\tau)}}|$ and canceling the integral involving p_{xt}^2 on both sides we then obtain

$$\frac{d}{dt}\int_{-\infty}^{s(t)} p_t^2(x,t) dx \le \frac{t(\dot\alpha(t))^2}{2\alpha(t)} \int_{-\infty}^{s(t)} p_t^2(x,t) dx \quad \text{for all } t > 0.$$

Solving the differential inequality over (ε, t) $(0 < \varepsilon < t)$ then gives

$$\int_{-\infty}^{s(t)} p_t^2(x,t)\, dx \le \exp\left\{ \int_\varepsilon^t \frac{\tau \dot\alpha^2(\tau)}{2\alpha(\tau)}\, d\tau \right\} \int_{-\infty}^{s(\varepsilon)} p_t^2(x,\varepsilon)\, dx. \tag{3.11}$$

We now show that the right-hand side approaches zero as $\varepsilon \searrow 0$.

First, using the assumptions on α we can calculate

$$\exp\left\{ \int_\varepsilon^t \frac{\tau \dot\alpha^2(\tau)}{2\alpha(\tau)}\, d\tau \right\} \le |\log \varepsilon|^{O(1)}. \tag{3.12}$$

Next we estimate $\int_{-\infty}^{s(\varepsilon)} p_t^2(x,\varepsilon) dx$ by using the representation of p_t in Lemma 3.1. First we consider the integral in the representation of p_t. Observe that our assumption on α implies $\dot s(\varepsilon-\tau) < 0$ and $[s(\varepsilon) - s(\varepsilon-\tau) + (k-1)\tau] < 0$ for all small ε and $\tau \in (0, \varepsilon)$. Therefore, for all $x < s(\varepsilon)$, $\Gamma(x - s(\varepsilon - \tau), \tau) \le C \exp\{-\frac{[x-s(\varepsilon)]^2}{4\varepsilon}\} \tau^{-1/2} \exp\{-\frac{[s(\varepsilon)-s(\varepsilon-\tau)]^2}{4\tau}\}$, where C is independent of ε. Hence, with a change of variable $\tau \to \eta$ via $\eta = [s(\varepsilon-\tau) - s(\varepsilon)]/(2\sqrt{\tau})$ $(d\eta = -\frac{\dot s(\varepsilon-\tau) d\tau}{2\sqrt{\tau}} (1 - \frac{s(\varepsilon)-s(\varepsilon-\tau)}{2\tau \dot s(\varepsilon-\tau)})$, we can estimate

$$0 < -\int_0^\varepsilon \Gamma(x - s(\varepsilon-\tau), \tau) \dot s(\varepsilon-\tau)\, d\tau \le C \exp\{-\tfrac{[x-s(\varepsilon)]^2}{4\varepsilon}\} \int_0^{\sqrt{\alpha(\varepsilon)}} \frac{e^{-\eta^2}}{1 - \frac{s(\varepsilon)-s(\varepsilon-\tau)}{2\tau\dot s(\varepsilon-\tau)}}\, d\eta$$
$$\le C \exp\{-\tfrac{[x-s(\varepsilon)]^2}{4\varepsilon}\}$$

since the assumption on α implies that $0 < \frac{s(\varepsilon)-s(\varepsilon-\tau)}{2\tau\dot{s}(\varepsilon-\tau)} < \frac{3}{4}$ for all small ε and $\tau \in (0,\varepsilon)$. It then follows from the representation for p_t in Lemma 3.1 that

$$\int_{-\infty}^{s(\varepsilon)} p_t(x,\varepsilon)^2 dx \le \int_{-\infty}^{s(\varepsilon)} \Big\{ 2\Gamma^2(x,\varepsilon) + C\exp\{-\tfrac{[x-s(\varepsilon)]^2}{2\varepsilon}\}\Big\}$$
$$\le C\Big\{ \varepsilon^{-1/2} \int_{\sqrt{\alpha(\varepsilon)}}^{\infty} e^{-2\eta^2} d\eta + \sqrt{\varepsilon}\Big\}$$

by using $\Gamma(x,\varepsilon) \le C\varepsilon^{-1/2}\exp\{-\frac{x^2}{4\varepsilon}\}$ and a change of variable $x = -2\sqrt{\varepsilon}\eta$. Since $\alpha(\varepsilon) = -\frac{1}{2}\log\varepsilon + O(1)$ for small ε, $\int_{\sqrt{\alpha(\varepsilon)}}^{\infty} e^{-2\eta^2}\, d\eta = [1+O(\alpha(\varepsilon)^{-1})]e^{-2\alpha(\varepsilon)}/(4\sqrt{\alpha(\varepsilon)}) = O(\varepsilon)$. Hence, $\int_{-\infty}^{s(\varepsilon)} p_t^2(x,\varepsilon) = O(\sqrt{\varepsilon})$. Substituting this last estimate and (3.12) into (3.11) and sending $\varepsilon \searrow 0$ we then conclude that $\int_{-\infty}^{s(t)} p_t^2(x,t) = 0$ for any $t > 0$. This implies that $p_t = 0$ for all $x < s(t)$, thereby completing the proof of the theorem. □

4. Asymptotic behavior of $s(t)$. In this section, we study the integrodifferential equation (3.9) for small t.

4.1. Reformulation of the problem. To study (3.9), it is convenient to study the function $s^2(t)/(4t)$ in the $\log(t)$ scale. For this purpose, we change variables from (s,t) to (u,ξ) by

$$\begin{cases} t = \frac{1}{4\pi k^2} e^{2\xi}, \\ s(t) = -2\sqrt{t\, u(\xi)}, \end{cases} \quad \Leftrightarrow \quad \begin{cases} \xi = \log\sqrt{t} + \log\sqrt{4\pi k^2}, \\ u(\xi) = s^2(t)/(4t)\ (=\alpha(t)). \end{cases}$$

To transform the integrodifferential equation (3.9) into the new unknown $u(\xi)$, we bear in mind that we are interested in small t, i.e., large negative ξ. Also, one expects $u(\xi) = -\xi - O(\xi^{-1})$ and $u'(\xi) = -1 + O(\xi^{-2})$. The absence of a constant term in the expansion of $u(\xi)$ is due to the presence of the particularly chosen constant $4\pi k^2$ in the definition of ξ.

In the sequel, $' = \frac{d}{d\xi}$ and $F[u](\xi)$ denotes the value at ξ of the function $F[u]$ when F is an operator.

Now we transform (3.9) into the new variables (u,ξ). Simple substitution yields

$$s(t)\Big\{\dot{s}(t) - \frac{s(t)}{2kt}\,\Gamma(s(t),t)\ \Big\} = u'(\xi) + G(u(\xi),\xi),$$
$$G(u,\xi) := 2u\Big(1 - \exp\{-u - \xi + \tfrac{(k-1)\sqrt{u}e^{\xi}}{2k\sqrt{\pi}} - \tfrac{(k+1)^2 e^{2\xi}}{16k^2\pi}\}\Big). \tag{4.1}$$

To convert the integral in (3.9), we notice that the singularity on the exponent looks like $\xi\ z$, where $z = \frac{(\sqrt{t}-\sqrt{t-\tau})^2}{\tau}$ if we were to suppose $s(t) \sim \sqrt{4t}\xi$. This suggests the use of the change of variable from τ to z via $\tau = \frac{4zt}{(1+z)^2}$. Then

$$z = \tfrac{\tau}{(\sqrt{t}+\sqrt{t-\tau})^2} = \tfrac{(\sqrt{t}-\sqrt{t-\tau})^2}{\tau}, \qquad \eta := \log\sqrt{4\pi k^2(t-\tau)} = \xi + \log\tfrac{1-z}{1+z}.$$

For notational simplicity, we write $u = u(\xi)$, $\hat{u} = u(\eta)$, and $\hat{u}' = u'(\eta)$. Then $s(t) = -2\sqrt{tu}$, $s(t-\tau) = -2\sqrt{(t-\tau)\hat{u}}$, $\dot{s}(t-\tau) = -\frac{\sqrt{\hat{u}}}{\sqrt{t-\tau}}[1 + \hat{u}'/(2\hat{u})]$, and

$$s(t)\int_0^t \Big(\frac{s(t)}{2t} - \frac{s(t)-s(t-\tau)}{\tau}\Big)\,\dot{s}(t-\tau)\,\Gamma(s(t)-s(t-\tau),\tau)\,d\tau$$
$$= \int_0^1 \Big\{\frac{1+z^2}{z}\sqrt{u} - \frac{1-z^2}{z}\sqrt{\hat{u}}\Big\}\Big\{1 + \frac{\hat{u}'}{2\hat{u}}\Big\}\frac{\sqrt{\hat{u}}\sqrt{-\xi}\,e^{\xi z - b}}{\sqrt{\pi z}}dz,$$

where

$$b = \log\sqrt{-\xi/u} + (u+\xi)z + \log[1+z]$$
$$+ \tfrac{1-z}{2}\{u-\hat{u}\} + \tfrac{1-z^2}{4z}\{\sqrt{u}-\sqrt{\hat{u}}\}^2 - \tfrac{(k-1)e^{\xi}}{2k\sqrt{\pi}}\{\sqrt{u} - \tfrac{1-z}{1+z}\sqrt{\hat{u}}\} + \tfrac{(k+1)^2 z e^{2\xi}}{4k^2\pi(1+z)^2}. \tag{4.2}$$

One observes that $\eta - \xi = \log\frac{1-z}{1+z} \sim 2z$ for small z. This suggests that we can approximate $\hat{u} = u(\eta)$ by $u(\xi - 2z)$. Hence, writing $(\frac{1+z^2}{z}\sqrt{u} - \frac{1-z^2}{z}\sqrt{\hat{u}})\sqrt{\hat{u}}$ as $\frac{u-u(\xi-2z)}{2z} + \frac{u(\xi-2z)-\hat{u}}{2z} - \frac{(\sqrt{u}-\sqrt{\hat{u}})^2}{2z} + z(\sqrt{u\hat{u}} + \hat{u})$, we can transform (3.9) (multiplied by $s(t)$) to

$$u'(\xi) + \int_0^1 \frac{u(\xi) - u(\xi-2z)}{2z}\frac{\sqrt{-\xi}e^{\xi z}}{\sqrt{\pi z}}\,dz + G(u(\xi),\xi) = F[u](\xi), \tag{4.3}$$

where $G(u,\xi)$ is as in (4.1) and the operator F is defined by

$$F[u](\xi) = -\int_0^1 \Big\{f_1 + f_2 + \hat{u}' f_3\Big\}\,dz\,, \tag{4.4}$$
$$f_1 = \Big\{\tfrac{u-u(\xi-2z)}{2z}(e^{-b}-1) + \tfrac{u(\xi-2z)-\hat{u}}{2z}e^{-b} - \tfrac{(\sqrt{u}-\sqrt{\hat{u}})^2}{2z}e^{-b}\Big\}\,\tfrac{\sqrt{-\xi}e^{\xi z}}{\sqrt{\pi z}},$$
$$f_2 = z(\sqrt{u\hat{u}} + \hat{u})\,\tfrac{\sqrt{-\xi}e^{\xi z - b}}{\sqrt{\pi z}},$$
$$f_3 = \Big\{\tfrac{1+z^2}{z}\sqrt{\tfrac{u}{\hat{u}}} - \tfrac{1-z^2}{z}\Big\}\tfrac{\sqrt{-\xi}e^{\xi z - b}}{\sqrt{\pi z}}.$$

Now we introduce a linear operator $\mathbf{L} : \phi \to \mathbf{L}[\phi]$ by

$$\mathbf{L}[\phi](\xi) = \int_0^1 \Big(\frac{1}{z}\int_0^z \phi(\xi - 2\zeta)d\zeta\Big)\frac{\sqrt{-\xi}e^{\xi z}}{\sqrt{\pi z}}\,dz\,.$$
$$= \int_0^1 \phi(\xi-2\zeta)\varrho(\xi,\zeta)\,d\zeta, \quad \varrho(\xi,\zeta) := \int_\zeta^1 \frac{\sqrt{-\xi}e^{\xi z}}{z\sqrt{\pi z}}\,dz\,. \tag{4.5}$$

Then equation (3.9) or (4.3) can be written as, for $\xi < 0$ (i.e., $t < 1/(4\pi k^2)$),

$$(\mathbf{I} + \mathbf{L})[u'](\xi) + G(u(\xi),\xi) = F[u](\xi), \tag{4.6}$$

where $\mathbf{I}$ represents the identity operator.

THEOREM 4.1. *Assume that $k > 0$. Then there exists a constant $\xi_0 < 0$ such that* (4.6) *admits a unique positive solution $u(\cdot) \in C^2((-\infty,\xi_0])$ having the asymptotic expansion, as $\xi \to -\infty$,*

$$u(\xi) = u_0(\xi) + O(\xi^{-4}), \quad u_0 := -\xi - \frac{1}{2\xi} + \frac{1}{8\xi^2} + \frac{17}{24\xi^3}. \tag{4.7}$$

4.2. Proof of Theorem 4.1. We use the following Schauder's fixed point theorem:

> *A continuous map $\mathbf{T}$ from a compact and convex subset $\mathbf{D}$ of a Banach space $\mathbf{X}$ to $\mathbf{D}$ possesses at least a fixed point.*

To apply this theorem, we define, for some negative constant ξ_0 to be chosen later,

$$\mathbf{X} = C^1((-\infty,\xi_0])\,, \qquad \|v\|_{\mathbf{X}} := \sup_{\xi\in(-\infty,\xi_0]}\{|v(\xi)| + |v'(\xi)|\}\,,$$
$$\mathbf{D} = \{v \in \mathbf{X} \;|\; |v(\xi)| + |v'(\xi)| + |v''(\xi)| \le |\xi|^{-1} \;\; \forall\,\xi \in (-\infty,\xi_0]\}\,.$$

The uniform decay of derivatives of functions in $\mathbf{D}$ ensures that $\mathbf{D}$ is compact and convex in $\mathbf{X}$. We also need two technical lemmas.

LEMMA 4.2. *Let u_0 be as in* (4.7). *There exists a positive constant C_0, depending only on k, such that for every $\xi_0 \le -2$ and every $u \in \{u_0\} + \mathbf{D}$, the function $F[u](\cdot)$ defined in* (4.4) *is continuous differentiable on $(-\infty, \xi_0]$ and*

$$\left|F[u](\xi) + 1\right| + \left|\frac{d}{d\xi}F[u](\xi)\right| \le C_0|\xi|^{-2} \qquad \forall\, \xi \in (\infty, \xi_0]\ . \tag{4.8}$$

LEMMA 4.3. *Let C_0 be as in the previous lemma. There exist a positive constant $M(C_0)$ and a negative constant $\Xi(C_0) \le -2$ such that if $\xi_0 \le \Xi[C_0]$ and $f \in C^1((-\infty, \xi_0])$ satisfies*

$$|f(\xi) + 1| + |f'(\xi)| \le C_0|\xi|^{-2} \qquad \forall \xi \in (-\infty, \xi_0],$$

then there exists a unique solution $w \in C^2((-\infty, \xi_0])$ to

$$(P1) \quad \begin{cases} (\mathbf{I} + \mathbf{L})[w'](\xi) + G(w(\xi), \xi) = f(\xi) & \forall\, \xi \in (-\infty, \xi_0], \\ \lim_{\xi \to -\infty}(w(\xi) + \xi) = 0\ . \end{cases} \tag{4.9}$$

In addition, the solution satisfies, for u_0 defined as in (4.7),

$$|w(\xi) - u_0(\xi)| + |w'(\xi) - u_0'(\xi)| + |\xi|^{-1}\ |w''(\xi) - u_0''(\xi)| \le M(C_0)|\xi|^{-3} \quad \forall\, \xi \le \xi_0.$$

Proof of Theorem 4.1. We choose $\xi_0 = \min\{\Xi[C_0], -M(C_0)\}$, where $C_0, M(C_0)$ and $\Xi[C_0]$, depending only on k, are as in the previous two lemmas.

For every $v \in \mathbf{D}$, we define $\mathbf{T}[v] = w - u_0$, where w is the solution to (4.9) with $f := F[u_0 + v]$. By Lemma 4.2 and Lemma 4.3, $\mathbf{T}$ is well defined. In addition, from the estimates for w and the definition of ξ_0, $\mathbf{T}[v] \in \mathbf{D}$; i.e., $\mathbf{T}$ maps $\mathbf{D}$ into itself.

Now we show that $\mathbf{T} : \mathbf{D} \subset \mathbf{X} \to \mathbf{X}$ is continuous. For this purpose, let v_j, $j = 1, 2, \ldots, \infty$, be functions in $\mathbf{D}$ such that $v_j \to v_\infty$ in $\mathbf{X} = C^1((-\infty, \xi_0])$ as $j \to \infty$. We want to show that $\mathbf{T}[v_j] \to \mathbf{T}[v_\infty]$ in $\mathbf{X}$. As every member of $\{\mathbf{T}[v_j]\}_{j=1}^{\infty}$ is in $\mathbf{D}$, which is compact in $\mathbf{X}$, any subsequence of $\{\mathbf{T}[v_j]\}$ has a subsubsequence convergent to a limit, say, $\tilde{v}$, in $\mathbf{X}$. Since along that subsubsequence, $F[u_0 + v_j] \to F[u_0 + v_\infty]$, $(\mathbf{T}[v_j])' \to \tilde{v}'$, and $\mathbf{L}[(u_0 + \mathbf{T}[v_j])'\ ] \to \mathbf{L}[(u_0 + \tilde{v})']$ in $C^0((-\infty, \xi_0])$, we conclude by taking the limit of the integrodifferential equation satisfied by $\mathbf{T}[v_j]$ that $u_0 + \tilde{v}$ is a solution to (4.9) with $f = F[u_0 + v_\infty]$. It then follows by uniqueness of (4.9) that $\tilde{v} = \mathbf{T}[v_\infty]$. Consequently, the whole sequence $\{\mathbf{T}[v_j]\}$ converges to $\mathbf{T}[v_\infty]$ in $\mathbf{X}$. Thus, $\mathbf{T}$ is continuous. The Schauder fixed point theorem then shows that $\mathbf{T}$ has at least one fixed point, which, after adding u_0, gives a solution to (4.3).

Finally, by Theorem 3.2, such a solution u is unique. This proves Theorem 4.1. ∎

4.3. Proof of Lemma 4.2. For notational simplicity, in the sequel, $O(f)$ stands for a quantity satisfying $|O(f)| \le C|f|$, where C is a positive constant depending only on k.

Since $u - u_0 \in \mathbf{D}$, $|u'(\xi) + 1| \le 2/|\xi|$, so that, by the mean value theorem and the definition $\eta = \xi + \log\frac{1-z}{1+z}$,

$$u - \hat{u} = u(\xi) - u(\eta) = u'(\xi - \theta)\{\xi - \eta\} = [1 + O(|\xi|^{-1})]\log\tfrac{1-z}{1+z}.$$

It then follows from the Lebesgue dominated convergence theorem that $F \in C^1((-\infty, \xi_0])$. It remains to estimate $F[u]$ and its derivative.

Note that b, defined in (4.2), is uniformly bounded in $z \in (0,1)$ and $\xi \in (-\infty, \xi_0]$. When $z \in [1/2, 1]$, the term $e^{\xi z} < e^{\xi/2}$ is exponential small in terms of $\xi < \xi_0 \le -2$. All other terms are of order $\log[1-z]$, so the total contribution of the integral for $z \in [1/2, 1]$ is exponential small in ξ. Thus, we need only pay attention when $z \in (0, 1/2]$.

First we estimate f_1. That $u - u_0 \in \mathbf{D}$ and $\eta - \xi = \log \frac{1-z}{1+z} = -2z + (z^2)$ implies $b = O(z^2 + z|\xi|^{-1} + \xi^{-2})$ so that $|f_1| = O(z^2 + z|\xi|^{-1} + \xi^{-2}) \frac{\sqrt{-\xi} e^{\xi z}}{\sqrt{\pi z}}$. Note that for every $\xi < 0$ and $i > -1/2$,

$$\int_0^1 z^i \frac{\sqrt{-\xi} e^{\xi z}}{\sqrt{\pi z}}\, dz \le \frac{|\xi|^{-i}}{\sqrt{\pi}} \int_0^\infty z^{i-1/2} e^{-z}\, dz = O(|\xi|^{-i}).$$

It then follows that $\int_0^1 |f_1| dz = O(\xi^{-2})$. A direct differentiation also shows that $\int_0^1 |\frac{d}{d\xi} f_1|\, dz = O(\xi^{-2})$ since $u'' = O(|\xi|^{-1})$.

Similarly, we can show that $\int_0^1 (|\hat{u}' f_3| + |\frac{d}{d\xi} \hat{u}' f_3|) dz = O(\xi^{-2})$.

For the integral involving f_2, we write $\sqrt{u\hat{u}} + \hat{u} = 2u - \frac{1}{2}(\sqrt{u} - \sqrt{\hat{u}})^2 - \frac{3}{2}(u - \hat{u}) = -2\xi + O(|\xi|^{-1} + |\xi|^{-1} \log^2 \frac{1+z}{1-z} + \log \frac{1+z}{1-z})$. Hence, $\int_0^1 f_2\, dz = \int_0^1 \frac{-2z\xi\sqrt{-\xi} e^{\xi z}}{\sqrt{\pi z}}\, dz + O(|\xi|^{-2}) = -1 + O(\xi|^{-2})$. A differentiation also gives $|\int_0^1 \frac{d}{d\xi} f_2\, dz| = O(\xi^{-2})$. This completes the proof. □

Remark 4.1. The integral in (3.10) is of size $|\xi|/s(t)$ if $\lambda \neq 1$. When $\lambda = 1$, this integral is of size $(\mathbf{L}[u'] - F[u])/s(t) = O(\xi^{-2})/s(t)$ since $u' = -1 + O(\xi^{-2})$ and $\mathbf{L}[u'] = -1 + O(\xi^{-2})$. Hence, the ratio of the integral and the first term on the right-hand side of (3.9) is of size $O(\xi^{-3})$; see Remark 3.1.

4.4. Idea for the proof of Lemma 4.3. To complete the proof of Theorem 4.1, it remains to prove Lemma 4.3, which will be done in the next two sections. Here we provide the main idea of the proof.

We first investigate in section 5 the linear operator $\mathbf{L}$. In particular, we show that the inverse operator $(\mathbf{I} + \mathbf{L})^{-1}$ is a bounded operator from C^0 to C^0. Also, we show that $\mathbf{L}[\phi]$ is always $1/2$ more differentiable than ϕ.

Then in section 6, we study, for any large integer j, an initial value problem $(\mathrm{P1})_j$ of the integrodifferential equation in (P1) in the interval $[-j, \xi_0]$ with "initial value" $w = u_0$ in $(-\infty, -j]$. The existence of a solution follows from a standard Picard iteration technique.

To obtain certain desired behavior of the solution of $(\mathrm{P1})_j$, we find that $(\mathrm{P1})_j$, as well as (P1), satisfy a comparison principle: larger initial data and larger source term produce larger solutions. Because of the large positive derivative $\frac{\partial}{\partial u} G(u_0, \xi) \sim 2|\xi|$, this comparison principle allows us to construct sub and super solutions of the form $u_0 \pm M|\xi|^{-3}$ to sandwich the solution to $(\mathrm{P1})_j$. Thus, we can take the limit $j \to \infty$ to obtain a solution to (P1) with the desired asymptotic behavior. Uniqueness of solutions to (P1) also follows from the comparison principle.

5. The operator L. In this section we study the operator $\mathbf{L}$ defined in (4.5).

LEMMA 5.1. *There exists a universal positive constant c_0 such that for every $\xi_0 \le -2$,*

$$\begin{aligned} c_0 \|\phi\|_{C^0(-\infty,\xi_0])} &\le \|(\mathbf{I} + \mathbf{L})[\phi]\|_{C^0(-\infty,\xi_0])} \\ &\le 2\|\phi\|_{C^0((-\infty,\xi_0])} \quad \forall\, \phi \in C^0((-\infty, \xi_0])\,. \end{aligned} \tag{5.1}$$

Consequently, $\mathbf{I}+\mathbf{L}$ *admits a bounded inverse* $(\mathbf{I}+\mathbf{L})^{-1}$ *from* $C^0((-\infty,\xi_0])$ *to itself and*

$$\tfrac{1}{2} \le \| \, (\mathbf{I}+\mathbf{L})^{-1} \, \|_{C^0((-\infty,\xi_0]\to C^0((-\infty,\xi_0])} \le \tfrac{1}{c_0}. \tag{5.2}$$

Proof. The second inequality in (5.1) follows easily from the first definition of $\mathbf{L}$ (one line above (4.5)).

To prove the first inequality, we make a scaling change a change of variables $z \to Z/|\xi|$ and $\zeta \to \theta/|\xi|$. Then $\mathbf{L}[\phi]$ in (4.5) can be written as

$$\mathbf{L}[\phi](\xi) = \int_0^{-\xi} \phi(\xi + 2\theta/\xi)\varrho_1(\xi,\theta)\,d\theta, \quad \varrho_1(\xi,\theta) = \int_\theta^{-\xi} \frac{e^{-Z}}{Z\sqrt{\pi Z}}\,dZ.$$

To prove the first inequality in (5.1), we notice that $\mathbf{L}$ is linear, so that we can without loss of generality assume $\|\phi\|_{C^0} = 1 = \sup_{\xi<\xi_0}\phi(\xi)$.

Let $j \ge 2$ be any integer, and let $\xi_j \in (-\infty,\xi_0]$ be a point such that $\phi(\xi_j) \ge 1 - 1/j$. Since $\int_0^{-\xi_j} \varrho_1(\xi_j,\theta) < 1$ for any $m \in (0,1/2)$,

$$\begin{aligned}
&(\mathbf{I}+\mathbf{L})[\phi](\xi_j) \\
&\quad \ge \int_0^{-\xi_j} \Big\{1 + \phi(\xi_j + 2\theta/\xi_j)\Big\}\varrho_1(\xi_j,\theta)\,d\theta - \frac{1}{j} \ge m\varrho_1(\xi_j,1)\{1 - A(\xi_j,m)\} - \frac{1}{j},
\end{aligned}$$

where $A(\xi,m) = \text{measure}\{\theta \in [0,1] \;\big|\; \phi(\xi + 2\theta/\xi) + 1 < m\}$. Suppose $A(\xi_j,m) > 0$. Then there is a unique $\hat{\xi}_j \in (\xi_j + 2/\xi_j, \xi_j)$ such that $\phi(\xi) + 1 > m$ in $(\hat{\xi}_j,\xi_j]$ and $\phi(\hat{\xi}_j) + 1 = m$. Since $\hat{\xi}_j + 2/\hat{\xi}_j < \xi_j + 2/\xi_j$,

$$\begin{aligned}
A(\xi_j,m) &= \frac{\xi}{2}\text{measure}\{z \in [\xi_j + 2/\xi_j, \xi_j] \mid \phi(z) < m-1\} \\
&\le \frac{\xi}{2}\text{measure}\{z \in [\hat{\xi}_j + 2/\xi_j, \hat{\xi}_j] \mid \phi(z) < m-1\} = \frac{\xi}{\hat{\xi}} A(\hat{\xi}_j,m) \leqslant A(\hat{\xi}_j,m).
\end{aligned}$$

Hence

$$\begin{aligned}
-(\mathbf{I}+\mathbf{L})[\phi](\hat{\xi}_j) &= 1 - m - \mathbf{L}[\phi](\hat{\xi}_j) \ge -m + \int_0^{-\hat{\xi}_j} \{1 - \phi(\hat{\xi}_j + 2\theta/\hat{\xi}_j)\}\varrho_1(\hat{\xi}_j,\theta)\,d\theta \\
&\ge -m + (2-m)A(\hat{\xi}_j,m)\varrho_1(\hat{\xi}_j,1) \ge -m + (2-m)A(\xi_j,m)\varrho_1(\xi_j,1) \\
&\ge \varrho(\xi_j,1)[-m + (1-m)A(\xi_j,m)].
\end{aligned}$$

It then follows that, regardless of the size of $A(\xi_j,m)$,

$$\begin{aligned}
\|(\mathbf{I}+\mathbf{L})[\phi]\|_{C^0} + \tfrac{1}{j} &\ge \varrho_1(\xi_j,1)\max\Big\{m[1 - A(\xi_j,m)], -m + (2-m)A(\xi_j,m)\Big\} \\
&\ge \varrho_1(-2,1)m(1-m).
\end{aligned}$$

Sending $j \to \infty$ and taking $m = 1/2$ we then conclude that (5.1) holds with $c_0 = \varrho_1(-2,1)/4$.

The invertibility of $\mathbf{I}+\mathbf{L}$ and the estimate (5.2) follow from (5.1) and the Hahn–Banach theorem. ∎

Next, we show that $\mathbf{L}[\phi]$ is 1/2 more differentiable than ϕ.

LEMMA 5.2. *For every $\beta \in [0,1]$, there exists a positive constant $C(\beta)$ such that*

$$\|\phi\|_{C^0([a-2,b])} + C(\beta)\sqrt{|a|}\ [\phi]_{\beta,[a-2,b]} \ \geq \ \begin{cases} [\ \mathbf{L}[\phi]\ ]_{\beta+1/2,[a,b]} & \text{if } \beta \in [0,1/2), \\ [\mathbf{L}[\phi]\ ]^*_{1,[a,b]} & \text{if } \beta = 1/2, \\ [(\mathbf{L}[\phi])'\ ]_{\beta-1/2,[a,b]} & \text{if } \beta \in (1/2,1], \end{cases} \tag{5.3}$$

where $[\psi]^*_{1,[a,b]} := \sup_{a\leq\xi_2<\xi_1\leq b} \frac{|\psi(\xi_2)-\psi(\xi_1)|}{|\xi_2-\xi_1|\ \max\{1,|\log(\xi_1-\xi_2)|\}}$ *and*

$$[\psi]_{\beta,[c,d]} = \sup_{c\leq\xi_2<\xi_1\leq d} \frac{|\psi(\xi_2)-\psi(\xi_1)|}{|\xi_2-\xi_1|^\beta} \quad \forall \beta \in [0,1]. \tag{5.4}$$

We remark that $\mathbf{L}[\phi](\xi)$ depends only on values of ϕ in $[\xi-2,\xi]$. Also the factor $\sqrt{|a|}$ on the left-hand side of (5.3) is necessary since $\lim_{\xi\to\infty}\{\mathbf{L}[\phi](\xi)-\phi(\xi)\}=0$ for any bounded and uniformly continuous function ϕ.

Proof. Let ξ_1 and ξ_2 be any numbers such that $a \leq \xi_2 < \xi_1 \leq b$. Set $h = \xi_1 - \xi_2$. Since $\|\mathbf{L}\|_{C^0\to C^0} \leq 1$, we need only consider the case $h < 1/4$. Also, $\mathbf{L}[\mathbf{1}](\xi) = \int_0^{-\xi} \varrho(\xi,\theta)d\theta = \int_0^{-\xi} \frac{e^{-z}}{\sqrt{\pi z}}\,dz$, so that $\|\mathbf{L}[\mathbf{1}]\|_{C^1} \leq 1$. Hence, by considering the function $\phi(\xi) - \phi(\xi_2)\mathbf{1}$ if necessary, we can assume that $\phi(\xi_2) = 0$.

First, we consider the case $\beta \in [0,1/2]$. Using (4.5) and changing of variable $\zeta \to \zeta + h/2$ for the integral for $\mathbf{L}[\phi](\xi_1)$ we have

$$\begin{aligned} \mathbf{L}[\phi](\xi_2) - \mathbf{L}[\phi](\xi_1) = &\int_0^{1-h/2} \phi(\xi_2-2\zeta)\Big\{\varrho(\xi_2,\zeta) - \varrho(\xi_1,\zeta+h/2)\Big\}\,d\zeta \\ &+ \int_{1-h/2}^1 \phi(\xi_2-2\zeta)\varrho(\xi_2,\zeta)\,d\zeta - \int_0^{h/2} \phi(\xi_1-2\zeta)\varrho(\xi_1,\zeta)\,d\zeta\ . \end{aligned}$$

Now $\phi(\xi_2)=0$ implies $|\phi(\xi_2-2\zeta)| \leq [\phi]_\beta |2\zeta|^\beta$ for all $\zeta \in [0,1]$ and $|\phi(\xi_1-2\zeta)| \leq h^\beta [\phi]_\beta$ for all $\zeta \in [0,h/2]$. Hence, $|\mathbf{L}[\phi](\xi_2) - \mathbf{L}[\phi](\xi_1)|$ is bounded by

$$\begin{aligned} [\phi]_\beta\ \Bigg\{ &\int_0^{1-h/2} (2\zeta)^\beta \Big|\varrho(\xi_1-h,\zeta) - \varrho(\xi_1,\zeta+h/2)\Big|\,d\zeta \\ &+ \int_{1-h/2}^1 \varrho(\xi_2,\zeta)\,d\zeta + \int_0^{h/2} h^\beta \varrho(\xi_1,\zeta)\,d\zeta \Bigg\}. \end{aligned} \tag{5.5}$$

Note that $\varrho(\xi,\zeta) < \sqrt{|\xi|/\pi}\int_\zeta^1 z^{-3/2}dz \leq 2\sqrt{|\xi|}\zeta^{-1/2}$. The last two integrals are bounded by $O(1)\sqrt{|\xi_2|}h^{\beta+1/2}$ when $\beta \in [0,1/2]$. Also

$$\begin{aligned} |\varrho(\xi_1-h,\zeta) - \varrho(\xi_1,\zeta+h/2)| \leq &\int_{\zeta+h/2}^1 \frac{|\sqrt{|\xi_1-h|}e^{(\xi_1-h)z} - \sqrt{\xi_1}e^{\xi_1 z}|}{\sqrt{\pi z^3}}dz \\ &+ \int_\zeta^{\zeta+h/2} \frac{\sqrt{|\xi_1-h|}e^{(\xi_1-h)z}}{\sqrt{\pi z^3}}dz \end{aligned}$$

The first integral is bounded by $O(1)\sqrt{|\xi_2|}h\zeta^{-1/2}$ whereas the second integral is bounded by

$$O(1)\sqrt{|\xi_2|}\Big(\frac{1}{\sqrt{\zeta+h/2}} - \frac{1}{\sqrt{\zeta}}\Big) = \frac{O(1)\sqrt{|\xi_2|}\ \ h/2}{\sqrt{\zeta}\sqrt{\zeta+h/2}(\sqrt{\zeta+h/2}+\sqrt{\zeta})} = \frac{O(1)\sqrt{|\xi_2|}\,h}{\zeta\sqrt{\zeta+h/2}}.$$

Finally, note that

$$\int_0^1 \frac{\zeta^\beta h}{\zeta\sqrt{\zeta+h/2}}d\zeta = h^{\beta+1/2}\int_0^{1/h}\frac{z^{\beta-1}}{\sqrt{z+1/2}}dz = \begin{cases} O(\beta)h^{\beta+1/2} & \text{if } \beta\in(0,1/2),\\ O(1)h|\log h| & \text{if } \beta = 1/2.\end{cases}$$

Therefore, the quantity in (5.5) is bounded by $C(\beta)\sqrt{|\xi_2|}[\phi]_\beta h^{\beta+1/2}$ if $\beta\in[0,1/2)$ and by $C(\beta)\sqrt{|\xi_2|}[\phi]_\beta h\,|\log h|$ if $\beta = 1/2$. This proves (5.3) for the case $\beta\in[0,1/2]$.

Next we consider $\beta\in(1/2,1]$. Set $\Phi(\xi) = \int_{\xi_2}^{\xi}\phi(\eta)\,d\eta$. Then

$$\mathbf{L}[\phi] = \int_0^1 \frac{\Phi(\xi)-\Phi(\xi-2z)}{2z}\frac{\sqrt{-\xi}e^{\xi z}}{\sqrt{\pi z}},$$

and

$$\frac{d}{d\xi}\mathbf{L}[\phi](\xi) = \int_0^1 \frac{\phi(\xi)-\phi(\xi-2z)}{2z}\frac{\sqrt{-\xi}e^{\xi z}}{\sqrt{\pi z}}\,dz + \frac{1}{2\xi}\int_0^1 \frac{\Phi(\xi)-\Phi(\xi-2z)}{2z}\frac{(1+2\xi z)\sqrt{-\xi}e^{\xi z}}{\sqrt{\pi z}}\,dz =: \mathrm{I}(\xi)+\mathrm{II}(\xi)\,.$$

Note that $\frac{d}{d\xi}\mathrm{II}(\xi)$ is bounded by $[\phi]_\beta$ since $\Phi' = \phi$ and $|\phi(\xi)-\phi(\xi-2z)|\le[\phi]_\beta(2z)^\beta$ and $\beta>1/2$. It remains to consider $I(\cdot)$. We write

$$I(\xi_1)-I(\xi_2) = \int_0^1 \frac{\phi(\xi_1)-\phi(\xi_1-2z)-\phi(\xi_2)+\phi(\xi_2-2z)}{2z}\frac{\sqrt{-\xi_1}e^{\xi_1 z}}{\sqrt{\pi z}}\,dz + \int_0^1 \frac{\phi(\xi_2)-\phi(\xi_2-2z))}{2z}\,\frac{\sqrt{-\xi_1}e^{\xi_1 z}-\sqrt{-\xi_2}e^{\xi_2 z}}{\sqrt{\pi z}}\,dz.$$

Since $|\phi(\xi_2)-\phi(\xi_2-2z)|\le[\phi]_\beta(2z)^\beta$ with $\beta>1/2$, the second integral is bounded by $h[\phi]_\beta|\xi_1|^{-\beta}\int_0^\infty(1+Z)Z^{\beta-3/2}e^{-Z}\,dZ \le Ch[\phi]_\beta|\xi_1|^{-\beta}$. To estimate the first integral, we use

$$|\phi(\xi_1)-\phi(\xi_1-2z)-\phi(\xi_2)+\phi(\xi_2-2z)|\le 2[\phi]_\beta\min\{h^\beta,(2z)^\beta\}$$

so that the first integral is bounded by

$$2[\phi]_\beta\int_0^\infty \frac{\min\{(2z)^\beta,h^\beta\}}{2z}\frac{\sqrt{-\xi_2}e^{\xi_2 z}}{\sqrt{\pi z}}\,dz \le C\sqrt{-\xi_2}[\phi]_\beta h^{\beta-1/2}.$$

In summary, we have $|\frac{d}{d\xi}\mathbf{L}[\phi](\xi_1)-\frac{d}{d\xi}\mathbf{L}[\phi](\xi_2)|\le C[\phi]_\beta\sqrt{-\xi_2}h^{\beta-1/2}$. This completes the proof. ∎

Remark 5.1. With the same technique, one can show that for any positive non-integer β, $\mathbf{L}[\phi]\in C^\beta$ if $\phi\in C^{\beta-1/2}$. Also we can show that $F[u]$ defined in (4.4) is always 1/2 more differentiable that u', assuming that $u\sim-\xi+o(1/\xi)$. We omit the details.

6. Proof of Lemma 4.3.

6.1. The truncated problem. We first study problem (P1) in a finite interval $[-j,\xi_0]$:

$$(\mathrm{P1})_j \qquad \begin{cases} (\mathbf{I}+\mathbf{L})[w'](\xi)+G(w(\xi),\xi) = f(\xi) & \forall\,\xi\in(-j,\xi_0],\\ w(\xi) = u_0(\xi) & \forall\xi\in(-\infty,-j]\,.\end{cases} \tag{6.1}$$

Since we aim for positive solutions, we extend $G(w,\xi)$ for negative w by 0.

LEMMA 6.1. *Let $-j < \xi_0 \le -2$, $f(\cdot)$ be any continuous function on $[-j, \xi_0]$, and u_0 be any differentiable function on $(-\infty, \xi_0]$. Then $(\mathrm{P1})_j$ admits a unique solution $w \in C^1([-j+, \xi_0])$.*

Proof. We first note that $\mathbf{L}[w'] = \int_0^1 \frac{w(\xi)-w(\xi-2z)}{2z} \frac{\sqrt{-\xi}e^{\xi z}}{\sqrt{\pi z}}\, dz$, so that $\|\mathbf{L}[w']\|_{C^0([-j,\xi])}$ $\le C_j [w]_{3/4,[-j-2,\xi]}$ for all $\xi \in [-j, \xi_0]$, where $C_j = j^{1/4} \int_0^\infty Z^{-3/4} e^{-Z} dZ$ and $[\phi]_{\beta,[a,b]}$ is as in (5.4). We remark that if $\phi(a) = 0$, then $[\phi]_{0,[a,b]} \ge \|\phi\|_{C^0([a,b])} := \sup_{\xi\in[a,b]} |\phi(\xi)|$.

Next we note that the function $G(w, \xi)$, after extension by 0 for negative w, is bounded, uniformly in $\xi \in [-j, \xi_0]$ and $u \in [-M, M]$ for every $M > 0$. Also $L_j = \sup_{\xi\in[-j,0], w\ge 0} |G_w(w, \xi)| < \infty$.

We now use the Picard iteration to establish the existence and uniqueness. Starting with $w_0 \equiv u_0$, we successively define w_i, $i = 1, 2, \ldots$, by $w_i = u_0$ in $(-\infty, -j]$ and

$$w_i(\xi) = u_0(-j) + \int_{-j}^{\xi} \Big\{ f(\hat\xi) - G(w_{i-1}(\hat\xi), \hat\xi) - \mathbf{L}[w'_{i-1}](\hat\xi) \Big\} d\hat\xi, \quad \xi \in [-j, \xi_0].$$

Taking the difference of the equations (and also their derivative) for w_{i+1} and w_i we obtain for all $i \ge 1$ and all $\xi \in (-j, \xi_0]$

$$[w_{i+1} - w_i]_{0,[-j,\xi]} \le L \int_{-j}^{\xi} \Big\{ [w_i - w_{i-1}]_{0,[-j,\hat\xi]} + [w_i - w_{i-1}]_{3/4,[-j,\hat\xi]} \Big\} d\hat\xi,$$
$$[w_{i+1} - w_i]_{1,[-j,\xi]} \le L \Big\{ [w_i - w_{i-1}]_{0,[-j,\xi]} + [w_i - w_{i-1}]_{3/4,[-j,\xi]} \Big\},$$

where $L = \max\{C_j, L_j\}$. Since $[\phi]_{3/4,[a,b]} \le ([\phi]_{1,[a,b]})^{3/4} ([\phi]_{0,[a,b]})^{1/4}$ for any ϕ and any interval $[a, b]$, mathematical induction then gives for $\beta = 0, 3/4, 1$ and all $i \ge 2$

$$[w_{i+1} - w_i]_{\beta,[-j,\xi]} \le \frac{M^i (\xi + j)^{i/4}}{(i\,!)^{1/4}} \Big(\frac{i}{4(\xi + j)} \Big)^{\beta} \qquad \forall \xi \in (-j, \xi_0]$$

for some sufficiently large constant M depending only on j and u_0. Following the rest steps of the Picard iteration method (see, for example, [9]), we then complete the proof. □

To take the limit $j \to \infty$ for solutions of $(P1)_j$, we need certain estimates. This will be done via a comparison principle and construction of sub and super solutions.

6.2. The comparison principle. For convenience, we introduce a nonlinear operator $\mathbf{N}$ defined by

$$\mathbf{N}[w](\xi) = (\mathbf{I} + \mathbf{L})[w'](\xi) + G(w(\xi), \xi)\,. \tag{6.2}$$

LEMMA 6.2 (comparison principle). *Let $\xi_0 \le -2$ be any number and w_1 and w_2 be two (piecewise) continuous differentiable functions on $(-\infty, \xi_0]$ satisfying the following:*

(i) $\min\{w_1, w_2\} \ge 3/2$ *in* $(-\infty, \xi_0]$;

(ii) *there exists* $j \in \{\infty\} \cup (-\xi_0, \infty)$ *such that*

$$\mathbf{N}[w_1](\xi) \ge \mathbf{N}[w_2](\xi) \quad \forall\, \xi \in (-j, \xi_0]$$

and $\liminf_{\xi\to-\infty}\{w_1(\xi) - w_2(\xi)\} \ge 0$ *if* $j = \infty$, *and* $w_1(\xi) \ge w_2(\xi)$ *on* $(-\infty, -j]$ *if* $j < \infty$.

Then $w_1(\xi) \ge w_2(\xi)$ *for all* $\xi \in (-\infty, \xi_0]$.

Proof. Let $\varepsilon \in (0, 1/4)$ be any constant. We define

$$\xi_\varepsilon = \sup\{\xi \le \xi_0 \mid w_1 + \varepsilon > w_2 \ \text{ in } \ (-\infty, \xi)\}.$$

The "initial condition" in assumption (ii) implies that ξ_ε is well defined and $\xi_\varepsilon > -j$. We claim that $\xi_\varepsilon = \xi_0$. In fact, if this is not true, then $w_2 < w_1 + \varepsilon$ in $(-\infty, \xi_\varepsilon)$, and at $\xi = \xi_\varepsilon$, $w_2 = w_1 + \varepsilon$ and $w_2' \ge w_1'$. In addition, $G(w_2(\xi_\varepsilon), \xi_\varepsilon) = G(w_1(\xi_\varepsilon) + \varepsilon, \xi_\varepsilon) > G(w_1(\xi_\varepsilon), \xi_\varepsilon)$ since $w_1(\xi_\varepsilon) = -\varepsilon + \max\{w_1(\xi_\varepsilon), w_2(\xi_\varepsilon)\} > 5/4$ and $G_w(w, \xi) > 0$ when $w > 5/4$. Hence

$$\begin{aligned}
&\mathbf{N}[w_1](\xi_\varepsilon) \\
&\quad = w_1'(\xi_\varepsilon) + \int_0^1 \frac{\{w_1(\xi_\varepsilon) + \varepsilon\} - \{w_1(\xi_\varepsilon - 2z) + \varepsilon\}}{2z} \frac{\sqrt{-\xi_\varepsilon} e^{\xi_\varepsilon z}}{\sqrt{\pi z}} dz + G(w_1(\xi_\varepsilon), \xi_\varepsilon) \\
&\quad < w_2'(\xi_\varepsilon) + \int_0^1 \frac{w_2(\xi_\varepsilon) - w_2(\xi_\varepsilon - 2z)}{2z} \frac{\sqrt{-\xi_\varepsilon} e^{\xi_\varepsilon z}}{\sqrt{\pi z}}\, dz + G(w_2(\xi_\varepsilon), \xi_\varepsilon) = \mathbf{N}[w_2](\xi_\varepsilon),
\end{aligned}$$

which contradicts the assumption that $\mathbf{N}[w_1] \ge \mathbf{N}[w_2]$ in $(-j, \xi_0]$. This contradiction shows that $\xi_\varepsilon = \xi_0$; namely, $w_1(\xi) + \varepsilon \ge w_2(\xi)$ in $(-\infty, \xi_0]$. Sending ε to 0 we then obtain the assertion of the lemma. □

One notices that the condition (i) is used only to ensure that $G_w(w, \xi) > 0$ for any $w \ge \max\{w_1, w_2\}$.

For later applications, we also provide the following maximum principle.

LEMMA 6.3 (maximum principle). *Let $L(\cdot)$ be a continuous and uniformly positive function on $(-\infty, \xi_0]$ and W be a Lipschitz continuous functions on $(-\infty, \xi_0]$ satisfying*

$$(\mathbf{I} + \mathbf{L})[W'](\xi) + L(\xi) W(\xi) \ge 0 \quad \forall \xi \in (-\infty, \xi_0], \qquad \inf_{\xi \le \xi_0} W(\xi) > -\infty.$$

Then $W \ge 0$ on $(-\infty, \xi_0]$.

The proof follows closely the proof for the previous lemma and is omitted.

6.3. Estimates for solutions of $(\mathbf{P1})_j$. Let C_0 be the constant in Lemma 4.2.

LEMMA 6.4. *There exists a large negative constant $\Xi_1(C_0)$ such that if $\xi_0 \le \Xi_1(C_0)$ and $f(\cdot) \in C^0((-\infty, \xi_0])$ satisfying*

$$|f(\xi) + 1| \le C_0 \xi^{-2}, \tag{6.3}$$

then the unique solution w to $(\mathrm{P1})_j$ with $u_0 = -\xi - \frac{1}{2}\xi^{-1} + \frac{3}{8}\xi^{-2} + \frac{17}{24}\xi^{-3}$ satisfies

$$|w(\xi) - u_0| \le (1 + \tfrac{1}{2} C_0) |\xi|^{-3} \qquad \forall \xi \in (-\infty, \xi_0], \tag{6.4}$$

$$|w'(\xi) - u_0'| \le \frac{2C_0 + 2}{c_0} \xi^{-2} \qquad \forall \xi \in (-\infty, \xi_0]. \tag{6.5}$$

Proof. Let $w_\pm = u_0 \mp M\xi^{-3}$, where $M > 0$ is to be determined. Then $\frac{w_\pm(\xi) - w_\pm(\xi - 2z)}{2z}$ $= w_\pm'(\xi) + z\, O(|\xi|^{-3})$. It then follows that $\mathbf{L}[w_\pm'] = \int_0^1 \frac{w_\pm(\xi) - w_\pm(\xi - 2z)}{2z} \frac{\sqrt{-\xi} e^{\xi z}}{\sqrt{\pi z}}\, dz =$ $w_\pm'(\xi) + O(|\xi|^{-4})$. Consequently,

$$\begin{aligned}
&\mathbf{N}[w_\pm] - f(\xi) = (\mathbf{I} + \mathbf{L})[w_\pm'] + G(w_\pm, \xi) - f(\xi) \\
&= 2w_\pm'(\xi) + 2w_\pm(1 - e^{-w_\pm - \xi}) - f(\xi) + O(|\xi|^{-4}) \\
&= -1 - f(\xi) \pm 2M\xi^{-2} + O(|\xi|^{-3}).
\end{aligned}$$

Hence, taking $M = C_0/2 + 1$ and $\Xi_1[C_0]$ large enough, we have $\pm\mathbf{N}[w_\pm] > 0$ in $(-\infty, \xi_0]$. The comparison principle then gives $w_- \le w \le w_+$ in $(-\infty, \xi_0]$ and therefore also (6.4). In addition,

$$(\mathbf{I}+\mathbf{L})[w' - u_0'](\xi) = f(\xi) - G(w(\xi), \xi) - 2u_0' + O(|\xi|^{-4}) = O(\xi^{-2}).$$

The estimate (6.5) then follows from the boundedness of $\|(\mathbf{I}+\mathbf{L})^{-1}\|_{C^0 \to C^0}$. □

Now we are ready to prove Lemma 4.3.

6.4. Proof of Lemma 4.3. Let C_0 and $\Xi_1(C_0)$ be as in Lemmas 4.2 and 6.4, respectively. Let $\xi_0 \le \Xi_1(C_0]$.

For each integer $j > |\xi_0|$, let w_j be the solution to $(P1)_j$. From Lemma 6.4, we know that we can extract a subsequence from $\{w_j\}_{j>|\xi_0|}$, which converges to w in $C^\beta((-\infty, \xi_0])$ for any $\beta \in (0,1)$ and some Lipschitz continuous function w satisfying the estimates (6.4) and (6.5). Consequently, $G(w_j(\cdot), \cdot) \to G(w(\cdot), \cdot)$ in $C^0((-\infty, \xi_0])$. Also, from the expression $\mathbf{L}[w_j'](\xi) = \int_0^1 \frac{w_j(\xi) - w_j(\xi - 2z)}{2z} \frac{\sqrt{-\xi} e^{\xi z}}{\sqrt{\pi z}}\, dz$ we see that $\mathbf{L}[w_j'] \to \mathbf{L}[w']$ uniformly in $(-\infty, \xi_0]$. Hence, from $w_j' = f - \mathbf{L}[w_j'] - G(w_j, \cdot)$, we conclude that $w_j(\cdot)' \to w'(\cdot)$ in $C^0((-\infty, \xi_0])$ and w is a $C^1((-\infty, \xi_0])$ solution to (P1). Uniqueness of the solution to (P1) follows from the comparison principle, i.e., Lemma 6.2 with $j = \infty$.

It remains to show that $w \in C^2$ and to estimate w'' and w' (better than (6.5)).

First, $f(\cdot) + G(w(\cdot), \cdot)$ is differentiable. Also, that $w' \in C^0$ and Lemma 5.2 implies that $\mathbf{L}[w'] \in C^{1/2}$ so that $w' = f - G(w, \cdot) - \mathbf{L}[w'] \in C^{1/2}$. Repeating this process, we then conclude that $w \in C^2$.

Once we know that w is C^2, we can differentiate the equation for w to obtain

$$(\mathbf{I}+\mathbf{L})[w''] + L(\xi)w' = f_\xi - G_\xi(w, \xi) - \psi,$$

where $L(\xi) = G_w(w(\xi), \xi))$ and $\psi = \int_0^1 \frac{w(\xi) - w(\xi - 2z)}{2z} \frac{\sqrt{-\xi} e^{\xi z}}{\sqrt{\pi z}} \left\{ \frac{1}{\xi} + z \right\} dz = O(\xi^{-2})$ since $w' = -1 + O(|\xi|^{-1})$.

Using the estimate (6.4) and the definition of G in (4.1) we see that $L(\xi) = G_w(w, \xi) = -2\xi - 1 - 2\xi^{-1} + O(|\xi|^{-2})$ and $G_\xi = -2\xi - 1 - \xi^{-1} + O(\xi^{-2})$. Since $(\mathbf{I}+\mathbf{L})[u_0''] = O(|\xi|^{-3})$ we have $(\mathbf{I}+\mathbf{L})[(w - u_0)''] + L(\xi)(w - u_0)' = f_\xi - G_\xi - \psi - L(\xi)u_0' + O(|\xi|^{-3}) = O(\xi^{-2})$ by the assumption on f_ξ.

Now we can use the maximum principle (Lemma 6.3) to estimate $(u - u_0)'$ and $(u - u_0)''$. For large constant M to be determined, set $W_\pm = -M\xi^{-3} \pm (u - u_0)'$. Then $(\mathbf{I}+\mathbf{L})[W'] + L(\xi)W = 2M\xi^{-2} + O(1)|\xi|^{-2} + O(|\xi|^{-3})$, where $O(1)$ depends only on C_0. Hence, taking M large (depending only on C_0) such that $M + O(1) = 1$ and then taking $\Xi(C_0)$ large enough negative, we have, when $\xi_0 \le \Xi(C_0)$, that $W_\pm > 0$ in $(-\infty, \xi_0]$, i.e., $|(w - u_0)'| \le M|\xi|^{-3}$. (Note the improvement over (6.5).) In addition, from $(\mathbf{I}-\mathbf{L})[(w - u_0)''] = O(|\xi|^{-2}) - L(w - u_0)' = O(|\xi|^{-2})$ and the boundedness of $(\mathbf{I}+\mathbf{L})^{-1}$ we conclude that $(w - u_0)'' = O(\xi^{-2})$. This completes the proof of Lemma 4.3 and also of Theorem 4.1. □

Remark 6.1. Using the same argument as in problem $(P1)_j$ (the Picard iteration), one can extend the solution u of (4.3) to $\xi \in (-\infty, \infty)$.

Remark 6.2. With the preceding argument for the C^2 differentiability of w, one can actually show that solution u to (4.3) is C^∞. To do this, one writes the equation as

$$u' + \mathbf{L}[u'] + \mathbf{L}_1^u[u'] = (F[u] + \mathbf{L}_1^u[u']) - G(u, \cdot),$$

where $\mathbf{L}_1^u[\phi'] = \int_0^1 \phi'(\eta) f_3\, dz$ is the part of $F[u]$ involving the integral of $u'(\eta)f_3$. Then the right-hand side of the equation is always 1/2 more differentiable than that of u'. As the operator norm from $C^0((\infty,\xi])$ to $C^0((-\infty,\xi])$ of $\mathbf{L}_1^u$ is of order $|\xi|^{-1}$, one sees that $(\mathbf{I}+\mathbf{L}+\mathbf{L}_1^u)^{-1}$ is bounded from C^0 to C^0. It then follows from a boot strap argument that $u \in C^\infty$. See also Remark 2.2.

6.5. Higher order expansions.

THEOREM 6.5. *There exist constants $c_1, c_2, c_3, \ldots$ such that as $\xi \to -\infty$, the unique solution to* (P1) *has the asymptotic expansion $u \sim -\xi + \sum_{i=1}^\infty c_i \xi^{-i}$; in particular,* (1.2) *holds (with $\alpha(t)$ replaced by $u(\xi)$).*

Proof.

Construction of the asymptotic expansion. First, we can replace $G(u,\xi)$ defined in (4.1) by $2u(1-e^{-u-\xi})$ since the terms dropped are of order $O(|\xi|e^\xi)$. Similarly, we can drop the exponentially small terms in b defined in (4.2). (These are equivalent to replacing the fundamental solution $\Gamma(x,t)$ in (3.2) by $\frac{1}{2\sqrt{\pi t}}e^{-x^2/(4t)}$.) Writing (4.6) as $G(u,\xi) = F[u] - (\mathbf{I}+\mathbf{L})[u]$ and solving for u we obtain

$$u(\xi) = -\xi - \log\left\{1 + \frac{(\mathbf{I}+\mathbf{L})[u'] - F[u]}{2u(\xi)}\right\}.$$

Starting with $u = \xi + O(1)$ and successively replacing u on the right-hand side by its previous expansion, we then obtain expansions of all order. The key here is that the right-hand side produces a unique $n+1$th order expansion, if an nth order expansion of u is given, because of the denominator $2u(\xi)$.

With the help of Mathematica's symbolic package, we obtain, in particular, the expansion (1.2); see www.math.pitt.edu/~xfc.

Rigorous verification of the expansion.

For every $n \ge 2$, set $u_n = -\xi + \sum_{i=1}^n c_i \xi^{-i}$ and define

$$\mathbf{X}_n = \left\{ w \in C^1((-\infty,\xi_n]) \,\middle|\, |(w-u_n)(\xi)| + \tfrac{1}{|\xi|}|(w-u_n)'(\xi)| \le M_n|\xi|^{-n-1} \ \forall\ \xi \le \xi_n \right\}.$$

We shall use mathematical induction to show that for every integer $n \ge 2$, $u \in \mathbf{X}_n$ provided that we take ξ_n and M_n large enough.

Suppose $u \in \mathbf{X}_n$. Then one can verify that $F[u]-F[u_n] = O(|\xi|^{-n-1})$. In deriving this, we need

$$u(\xi)-u(\eta) = \int_\xi^\eta u'(\hat\xi)\, d\hat\xi = \int_\xi^\eta (u_n' + O(|\xi|^{-n})\, d\hat\xi = u_n(\xi) - u_n(\eta) + |\xi-\eta|O(|\xi|^{-n}).$$

Now define $w_\pm = u_{n+1} \pm M|\xi|^{-n-2}$, where M is to be determined. We can calculate $\mathbf{N}[w_\pm] := (\mathbf{I}+\mathbf{L})[w_\pm'] + G(w_\pm,\xi) = \mathbf{N}[u_{n+1}] \pm 2M|\xi|^{-n-1} + O(|\xi|^{-n-2})$ since $L(\xi) := G_w(u,\xi) = 2|\xi| + O(1)$.

From the construction of u_{n+1}, we have $\mathbf{N}[u_{n+1}] = F[u_n] + O(|\xi|^{-n-1})$, and we then conclude that $\mathbf{N}[w_\pm] - \mathbf{N}[u] = F[u_n] - F[u] \pm 2M|\xi|^{-n-1} + O(1)|\xi|^{-n-1} = 2 \pm M|\xi|^{-n-1} + O(1)|\xi|^{-n-1}$, where $O(1)$ is independent of M if ξ is large enough. Hence, there exist a large constant M_{n+1} and large negative constant ξ_{n+1} such that for $M = M_{n+1}$, $\pm(\mathbf{N}[u] - \mathbf{N}[w_\pm]) > 0$ in $(-\infty,\xi_{n+1})]$. Therefore, by comparison, $w_- < u < w_+$, i.e., $|u-u_{n+1}| \le M_{n+1}|\xi|^{-n-2}$ for all $\xi \le \xi_{n+1}$. We take M large enough. With this estimate, we also obtain $(\mathbf{I}+\mathbf{L})[(u-u_{n+1})'] = \{\mathbf{N}[u]-\mathbf{N}[u_{n+1}]\} - \{G(u,\xi)-G(u_{n+1},\xi)\} = O(|\xi|^{-n-1})$. Consequently, by the boundedness of $(\mathbf{I}+\mathbf{L})^{-1}$, $|u'-u_{n+1}| = O(|\xi|^{-n-1})$. Thus, $u \in \mathbf{X}_{n+1}$. This completes the proof. ∎

Remark 6.3. We did not include the second order derivative of u in the definition of $\mathbf{X}_n$ since we do not intend to establish the estimate for u''. On the other hand, we do need to include the second order derivative of u in $\mathbf{X}$ in the proof of Theorem 4.1 to make the set $\mathbf{D}$ compact in $\mathbf{X}$.

7. Approximations of the early exercise boundary. In applications, one needs to find quickly the early exercise boundary

$$S_f(T) = Ee^{s(t)}, \quad s(t) = -2\sqrt{t\,\alpha(t)}, \quad \alpha(t) = u(\xi),$$
$$\xi = \log\sqrt{4\pi k^2 t},\ k = 2r\sigma^{-2}, \quad t = \tfrac{1}{2}\sigma^2(T_F - T)$$

of the American put option. As mentioned in the introduction, there have been a number of theoretical approximations; see, for example, [25], [18], [5], and [2, 3, 14]. In this section we derive our new approximations mentioned in section 1 as well as provide numerical comparisons. In what follows, σ^2 and r are measured in annualized units, i.e., have units 1/year.

7.1. An explicit approximation near expiry. One notices that expansions such as (1.2) cannot be used for $\xi \geq 0$ (equivalent to $t \geq 1/(4\pi k^2)$ or $T_F - T > \sigma^2/(8\pi r^2)$). Indeed, our numerical evidence [8] shows that approximations based on the truncations of (1.2) break down much earlier, and higher order expansions approximate $s(t)$ better than the second order only if $T_F - T$ is shorter than a few minutes and, therefore, are of no practical use. In this aspect, the best choice for practical estimation of $S_f(T)$ near expiry is the second order approximation $u(\xi) \approx -\xi - \frac{1}{2\xi}$. It is good for $T_F - T$ less than a week when $\sigma = 0.25/\sqrt{\text{year}}, r = 0.1/\text{year}$. Nevertheless, we still want to use (1.2) to obtain better approximations.

We recall that the particular choice of the constant $4\pi k^2$ in the definition of $\xi = \log\sqrt{t} + \log\sqrt{4\pi k^2}$ is to eliminate the constant term in the expansion of $u(\xi)$. If we use another variable such as $\hat{\xi} = \log\sqrt{Bt}$ and expand u in terms of $1/\hat{\xi}$, then the corresponding expansions make sense for all $t < 1/B$. Based on this idea, for any $a > 0$, we expand $u(\xi)$ as $u = -\xi - \frac{1}{2(\xi-a)} + \frac{1/8+a/2}{(\xi-a)^2} + \frac{17/24-a/4-a^2/2}{(\xi-a)^3} + \cdots$. Being equivalent to (1.2) as $\xi \to -\infty$, this new expansion, however, can be evaluated for all $t < e^a/(4\pi k^2)$. In particular, taking $a = 0.96621$ to be the positive root to $17/24 - a/4 - a^2/2 = 0$ and truncating the expansion at the fourth order, we obtain (expl) in section 1. Numerical evidence shows that this new approximation (expl) is better than any of the straightforward truncations of (1.2), both in accuracy and in the length of the interval of validity. For $\sigma = 0.25/\sqrt{\text{year}}$ and $r = 0.1/\text{year}$, the approximation is very accurate for $T_F - T$ less than 1 month.

7.2. An implicit/series approximation. We can extend further the above idea. We seek approximations which meet two requirements: (i) they are valid asymptotic expansions as $\xi \to -\infty$, and (ii) they are analytic for all $\xi \in \mathbb{R}$. We find that such approximations can be easily obtained if we regard ξ as function of u, i.e., the inverse function of $\xi = \xi(u)$.

For every $a > 0$, we convert (1.2) into its equivalent form $\xi = -u - \log\{1 - \frac{1}{2(u+a)} - \frac{a}{2(u+a)^2} + \frac{1-a^2}{2(u+a)^2} + \cdots\}$. Hence, taking $a = 1$ and truncating the expansion at the fourth order, we obtain the implicit formula (imp1) in section 1. As a special advantage, this expansion is meaningful for all time since for every $\xi \in \mathbb{R}$, there is a unique u solving (imp1) and $\xi \to \infty$ as $u \to 0$, which is compatible with the fact that $u = \frac{s(t)^2}{4t} \to 0$ as $\xi = \log\sqrt{4\pi k^2 t} \to \infty$. Our numerical experiments in [8] show that

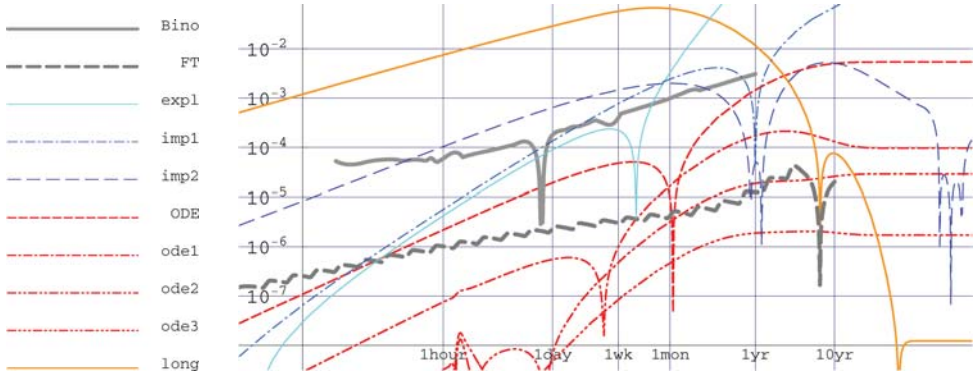


FIG. 7.1. *Error (in* $\log_{10}$*) of approximations for* $r = 0.1$ *and* $\sigma = 0.25$.

(impl1) is much better than (expl). It is reasonably good for $T_F - T$ as long as one year when $\sigma = 0.25$ and $r = 0.1$. See Figure 7.1.

7.3. Implicit/interpolation approximation. The approximation (impl1) is based on the asymptotic expansion (1.2) which concerns *only* the behavior of $s(t)$ near expiry. We now derive an approximation which incorporates as well the asymptotic behavior of u for large $t = e^{2\xi}/(4\pi k^2)$.

Using (3.6) and the change of variable $\tau \to z$ via $z = (s(t) - s(t-\tau))/(2\sqrt{\tau})$ we obtain

$$e^{-u-\xi} = \frac{2}{\sqrt{\pi}} \int_0^{\sqrt{u}} e^{-z^2} \theta_1(t,z)\theta_2(t,z)\, dz =: \frac{2\theta(t)}{\sqrt{\pi}} \int_0^{\sqrt{u}} e^{-z^2}\, dz, \tag{7.1}$$

where

$$\theta_1(t,z) = \Big\{2 - \frac{s(t) - s(t-\tau)}{\tau\, s'(t-\tau)}\Big\}^{-1}, \qquad \theta_2(t,z) = \exp\{\tfrac{(k-1)s(t-\tau)}{2} + \tfrac{(k+1)^2(t-\tau)}{4}\}.$$

Note that $\theta_1(t,0) = 1$ and $\theta(t,\sqrt{u}) = 1/2$. Also, $\lim_{t\to 0} \theta_2(t,z) = 1$ uniformly in z and $\lim_{t\to 0} \theta_1(t,z) = 1$ for any fixed finite z. Hence $\lim_{t\to 0} \theta(t) = 1$; cf. Remark 3.3.

Now we consider $\theta(t)$ for large t. From Theorem 2.3, we obtain $u = s^2(t)/(4t) \approx \log^2[1 + 1/k]/(4t)$, and $1/\theta(t) = \frac{2}{\sqrt{\pi}} e^{u+\xi} \int_0^{\sqrt{u}} e^{-z^2} dz \approx 2k \log[1 + 1/k]$ for large t (or ξ). Once we know the behavior of $\theta(t)$ for small and large t, we can approximate $\theta(t)$ for any t by interpolation. Without considering any more detailed behavior of θ for intermediate sizes of t, we choose, for simplicity, the approximation

$$\frac{1}{\theta(t)} \approx \frac{\theta(0)e^u + \theta(\infty)e^{1/u}}{e^u + e^{1/u}} = \frac{e^u + 2k\log(1 + 1/k)e^{1/u}}{e^u + e^{1/u}}.$$

Substituting this approximation into (7.1) and taking the log of both sides, we then obtain (imp2) in section 1. Figure 7.1 shows that (imp2) is better than (impl1) when $T_F - T$ is larger than 1 month (for $r = 0.1$ and $\sigma = 0.25$). When $T_F - T$ is less than a month, (impl1) is better than (imp2) since for small t, (impl1) is a fourth order approximation whereas (imp2) is only first order.

We remark that (impl1) can be revised to provide approximations which have higher order (as $t \to 0$) than (imp2), yet still capture the asymptotic behavior $s(t) \sim s_\infty$ for large t.

7.4. An ODE approximation. The ODE approximation in section 1 is obtained by neglecting the integral in (3.10). In the (u,ξ) variable, it can be written as

$$(7.2)\quad \begin{cases} \dfrac{1}{2u}\dfrac{du}{d\xi} = \exp\Big\{ -u-\xi+\frac{(k-1)\sqrt{u}e^{\xi}}{2k\sqrt{\pi}} - \frac{(k+1)^2e^{2\xi}}{16k^2\pi}e^{2\xi}\Big\} - 1 \quad \text{for} \quad \xi\in\mathbb{R}^1, \\ \lim_{\xi\to-\infty}\{u(\xi)+\xi\} = 0. \end{cases}$$

In numerically solving this ODE problem, the initial condition can be taken as $u|_{\xi=\xi_0} = -\xi_0 - \frac{1}{2\xi_0}$ for large negative ξ_0, say, $\xi_0 = -7$. This initial value problem is extremely stable with respect to the initial condition. Indeed, solutions with initial conditions $u|_{\xi=-7} = 1, 7, 10$ are indistinguishable when $\xi = -6$, since the initial difference decays with a speed $e^{-(\xi-\xi_0)G_w(u_0,\xi)} \approx e^{-2|\xi_0|(\xi-\xi_0)}$. Also, the computing time is almost instantaneous. Our numerical simulations indicate that this ODE approximation is better than any of the previous three algebraic approximations and is very accurate for a variety of parameter ranges of σ and r and almost all time $t > 0$.

7.5. An ODE iterative scheme. For the purpose of numerical comparison of the accuracy of theoretical approximations, highly accurate solutions for $s(t)$ are needed. Such solutions can be obtain by an iteration based on (3.10). We write (3.10) as (1.3). Asymptotic expansion gives for small t, $m(t) = 0 + 0\xi^{-1} + 0\,\xi^{-2} + \frac{1}{4}\,\xi^{-3} + O(\xi^{-4}) \approx \frac{1}{4}\xi^{-3}$, and for large t,

$$m(t) \approx m(\infty) = k\int_0^\infty \Big\{1 - \frac{2s(\tau)}{s(\infty)}\Big\}\exp\left\{\frac{(k-1)}{2}s(\tau) + \frac{(k+1)^2}{4}\tau\right\}\dot{s}(\tau)\,d\tau.$$

We remark that the integral is finite since letting t in (3.6) approach ∞ gives the identity

$$1 = -k\int_0^\infty \exp\left\{\frac{(k-1)}{2}s(\tau) + \frac{(k+1)^2}{4}\tau\right\}\dot{s}(\tau)\,d\tau.$$

Hence, $|m(\infty)| < 1$. Our numerical simulation shows that $m(t)$ changes sign exactly once, and this occurs near $\xi = 0$; for $r = 0.1$ and $\sigma = 0.25$, the minimum of $m(t)$ is $-0.003\ldots$, which occurs near $\xi = -5$, and the maximum is $0.17\ldots$, which is attained at $\xi = \infty$. From here, we can see why the ODE approximation (approximating m by 0) is very accurate for all $t > 0$.

The ODE iterative scheme that we propose is as follows: update s by solving (1.3) for s with $m(t)$ evaluated at a previous s; more precisely, $m^{(0)}(t) \equiv 0$, and for $n = 0, 1, \ldots$, $s^{(n)}(t) = -2\sqrt{t\,u^{(n)}(t)}$, where

$$(7.3)\quad \begin{cases} \dfrac{du^{(n)}}{d\log\sqrt{4\pi k^2 t}} = 2u^{(n)}\Big\{\frac{1+m^{(n)}}{\sqrt{4\pi k^2 t}}\exp\Big[\frac{k-1}{2}\sqrt{t\,u^{(n)}} - u^{(n)} - \frac{(k+1)^2}{4}t\Big] - 1\Big\}, \quad t \ge \delta, \\ u^{(n)}(\delta) = -\frac{1}{2}\log[4\pi k^2\delta] - \frac{1}{2}\log^{-1}[4\pi k^2\delta]), \end{cases}$$

$$m^{(n+1)}(t) = \begin{cases} ks^{(n)}(\delta) + \int_\delta^t \Big\{\frac{s^{(n)}(t)-s^{(n)}(\tau)}{t-\tau}\frac{2t}{s^{(n)}(t)} - 1\Big\}\frac{k\Gamma(s^{(n)}(t)-s^{(n)}(\tau),t-\tau)}{\Gamma(s^{(n)}(t),t)}\dot{s}^{(n)}(\tau)\,d\tau, & t \ge \delta_1, \\ \frac{1}{4}\log^{-3}\sqrt{4\pi k^2 t}, & t\in[\delta,\delta_1], \end{cases}$$

where δ and δ_1 are small numbers, say, $\delta = e^{-27}$ and $\delta_1 = e^{-18}$. The ODE for u is solved in the $\xi = \log\sqrt{4\pi k^2 t}$ variable, and a change of variable $\tau = t\eta^2$ is used in evaluating the integral for m. For $t \in [\delta_1, t_{\max}]$ ($t_{\max} = e^5$ will make $s(t_{\max})$ within 0.1% of s_∞), an increasing number, say, $32, 48, 72, 108, \ldots$, of points evenly distributed on the $\log t$ scale can be used to interpolate the function $m^{(1)}(t), m^{(2)}(t), m^{(3)}(t), \ldots$. In general, three iterations will provide a solution of relative error less than 10^{-5}, for all t, and five iterations will produce a numerical fixed point, which costs a total of less than 10 minutes of computing time on a Sparc server.

7.6. An approximation for large time. Although options with very long expiry rarely exist in practice, it is still useful to find the long time behavior (more precise than $S_f \approx Ek/(1+k)$) of the optimal exercise boundary since the scaling $t = \frac{\sigma^2}{2}(T_F - T) = \frac{r}{k}(T_F - T)$ tells us that short time expiration can be considered as long if r or σ is large. For this reason, we provide an approximation for large time, so that when incorporated with our ODE and/or implicit approximations, it will provide instantaneously a reliable approximation valid for all t and consequently for all ranges of σ and r.

Approximating $\Gamma(s(t), t)\{1 + m(t)\}$ on the right-hand side of (1.3) by $\frac{1+m(\infty)}{2\sqrt{\pi t}}$ $e^{-\frac{(1-k)}{2}s(\infty) - \frac{(k+1)^2}{4}t}$ and integrating the resulting approximation from t to ∞ we obtain (long) in section 1. As the ODE solution is a good approximation, $\hat{m}$ in (long) can be approximately calculated by using the ODE approximation for $s^{(0)}(\tau)$ in the integral; the computing time for this approximation is almost instantaneous since we can do so for $s^{(0)}$.

7.7. A numerical example. To give the reader an idea of the accuracy of our approximations, we provide in Figure 7.1 the results of a numerical simulation with typical parameters $E = 1$ (dollar), $r = 0.1$ (1/year), $\sigma = 0.25$ ($1/\sqrt{\text{year}}$), and $k = 2r/\sigma^2 = 3.2$.

In Figure 7.1, the vertical axis is $\log_{10}$ (errors) (with labeling being the actual size of the errors) of the various approximations for the optimal exercise boundary $S_f(T)$, whereas the horizontal axis is the time to expiry ($(T_F - T)$) in the log scale. In calculating the errors, the "exact" solution to which all the approximations are compared is actually the fifth iteration of (7.3), which is a numerical fixed point to (3.9). The labels stand for the binomial tree method (Bino), the front tracking (FT) and extrapolation methods (www.math.pitt.edu/˜xfc), the explicit approximation (expl), the implicit/series approximation (imp1), the implicit/interpolation approximation (imp2), the ODE approximation (ODE), the iterative ODE approximation (7.3) of the first, second, and third iterations, and the large time approximation (long). All the cusps (except those near the right and lower edges of the figure) are the points where errors change sign (since $\log_{10}(\text{error}) = -\infty$ at these points). The nonsmoothness of the curve marked imp2 near the right edge of the figure is due to the inefficiency of our Newton method in finding the roots u of (imp2) for large t. The bumps of ode2 and ode3 at the lower edge of the figure are numerical round-off errors.

The classical binomial and/or trinomial tree methods are typically used in the literature to find solutions to serve as the exact solutions with which approximations are to be compared. In calculating the optimal boundary $S_f(T)$, the point where the functions $P(S,T)$ and $E - S$ depart (tangentially), these tree methods are computing-time extensive. Depending on the initial guess of S_f, for each given T, it takes, with the number of division points $n = 1000$, about 5 to 20 minutes to find $S_f(T)$. The complexity of the method is $O(n^2)$ and the error is of order $O(\frac{\log n}{n})$. The solution

used in the figure contains 50 different sample Ts so that it takes about 10 hours of computing time.

The front tracking method that one of the authors designed has the same complexity $O(n^2)$ and error $O(\frac{\ln n}{n})$ as that of the binomial tree method. However, one can use solutions obtained with divisions $n, n/2$, and $n/4$, respectively, to extrapolate a much more accurate solution, as one can see from the significant difference between the curves marked Bino and FT. Given a fixed time $T_{\max}$, it takes, with $n = 2000$, about 15 minutes to find $S_f(T)$ for all $(T_F - T) \le T_{\max}$. The solution used in the figure is actually the union of solutions for $T_F - T$ in the interval $(\frac{1}{2}T_{\max}, T_{\max}]$ with $T_{\max} = 10/2^i$ (year) for $i = 0, 1, \ldots, 25$, and therefore, it takes a total of 10 computing hours.

As mentioned earlier, the ODE approximation is almost instantaneous. From the figure, one can see that the ODE approximation has already surpassed that obtained from the binomial method (with 1000 division). For the ODE iterative scheme, the computing time for the first iteration takes about 1 minute. To finish the fifth iteration takes a total of about 10 minutes.

Acknowledgment. The authors would like to express their appreciation to the referees for their valuable comments and suggestions.

REFERENCES

[1] F. Aitsahlia and T. Lai, *Exercise boundaries and efficient approximations to American option prices*, J. Computational Finance, 4 (2001), pp. 85–103.
[2] G. Barone-Adesi and R. E. Whaley, *Efficient analytic approximation of american option values*, J. Finance, 42 (1987), pp. 301–320.
[3] G. Barone-Adesi and R. Elliott, *Approximations for the values of American options*, Stochastic Anal. Appl., 9 (1991), pp. 115–131.
[4] G. Barles, J. Burdeau, M. Romano, and N. Samsone, *Critical stock price near expiration*, Math. Finance, 5 (1995), pp. 77–9,5.
[5] D. A. Bunch and H. Johnson, *The American put option and its critical stock price*, J. Finance, 55 (2000).
[6] P. Carr, R. Jarrow, and R. Myneni, *Alternative characterization of American put option*, Math. Finance, 2 (1992), pp. 87–105.
[7] Xinfu Chen, J. Chadam, L. Jiang, and W. Zhang, *Convexity of the Exercise Boundary of the American Put Option on a Zero Dividend Asset*, Math. Finance, to appear.
[8] Xinfu Chen and J. Chadam, *Analytic and numerical approximations for the early exercise boundary for American put options*, Dyn. Contin. Discrete Impuls. Syst., 10 (2003), pp. 649–657.
[9] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[10] E. Ekstrom, *Convexity of the optimal stopping boundary for the American put option*, J. Math. Anal. Appl., 299 (2004), pp. 147–156.
[11] J. D. Evans, R. R. Kuske, and J. B. Keller, *American options of assets with dividends near expiry*, Math. Finance, 12 (2002), pp. 219–237.
[12] A. Friedman, *Variational Principles and Free Boundary Problems*, John Wiley & Sons, New York, 1982.
[13] A. Friedman, *Partial Differential Equations of Parabolic Type,* Prentice–Hall, Englewood Cliffs, NJ, 1964.
[14] J. Hull, *Options, Futures and Other Derivative Securities*, 3rd ed., Prentice–Hall, New York, 1997.
[15] P. Jaillet, D. Lamberton, and B. Lapeyre, *Variational inequalities and the pricing of American options*, Acta Appl. Math., 21 (1990), pp. 263–289
[16] I. J. Kim, *The analytic valuation of American options*, Rev. Financial Stud., 3, pp. 547–572.
[17] C. Knessl, *Asymptotic analysis of the American call option with dividends*, European J. Appl. Math., 13 (2002), pp. 587–616.

[18] R. A. Kuske and J. B. Keller, *Optimal exercise boundary for an american put option*, Appl. Math. Finance, 5 (1998), pp. 107–116.
[19] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi–linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[20] L. W. MacMillan, *Analytic approximation for the american put option*, Adv. Future Option Res., 1 A (1997), pp. 119–139.
[21] H. P. McKeanm, Jr., *Appendix A: A free boundary problem for the heat equation arising from a problem in mathematical economics*, Industrial Management Rev., 6 (1965), pp. 32–39.
[22] R. Merton, *The theory of rational option pricing*, Bell J. Economics 4 (1973), pp. 141–183.
[23] G. Peskir, *On the American option problem*, Math. Finance, 15 (2005), pp. 169–181.
[24] D. M. Salopek, *American Put Option*, Pitman Monogr. Surveys Pure Appl. Math. 84, Addison–Wesley Longman, Reading, MA, 1997.
[25] R. Stamicar, D. Ševčovič, and J. Chadam, *The early exercise boundary for the American put near expiry: Numerical approximation*, Canad. Appl. Math. Quart., 7 (1999), pp. 427–444.
[26] P. Van Moerbeke, *On optimal stopping and free boundary problems*, Arch. Ration. Mech. Anal., 60 (1976), pp. 101–148.
[27] P. Wilmott, J. Dewynne, and S. Howison, *Option Princing: Mathematical Models and Computation*, Cambridge University Press, New York, 1995.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1642–1659

YOUNG MEASURES AND ORDER-DISORDER TRANSITION IN STATIONARY FLOW OF LIQUID CRYSTALS*

MARIA-CARME CALDERER† AND ALEXANDER PANCHENKO‡

Abstract. We study a system of nonlinear second order ordinary differential equations modeling Poiseuille flow of liquid crystals with variable degree of orientation, at the limit of large Ericksen number. The system is singularly perturbed and degenerate, and as a result the solutions are highly oscillatory. We obtain the relations satisfied by the Young measures generated by sequences of weak solutions, and show that the persistent oscillations are encoded in the Young measure generated by the molecular alignment variable. The effective equations correspond to the macroscopic isotropic Newtonian flow with a liquid crystalline microstructure indicating a remnant alignment.

Key words. Nematic liquid crystals, Young measures, non-Newtonian flows, singular perturbations, effective viscosity

AMS subject classifications. 34D15, 34E10, 76A05, 76A15

DOI. 10.1137/050623735

1. Introduction.

In this article we study stationary flow of nematic liquid crystals with large Ericksen number, $\mathcal{E}$, in terms of the Young measures generated by sequences of weak solutions of the governing equations. It is experimentally well known that liquid crystal flows with large Ericksen number present a high density of defects and texture which increases with increasing values of $\mathcal{E}$ (cf. [19], [11], [16], [13], and [18]).

The system that we analyze consists of ordinary differential equations for the variable fields $s(x)$, $\phi(x)$, and $v(x)$, with $x \in [-1,1]$ with $\mathbf{R}^2 \times (-1,1)$, representing the domain of the flow. The governing system is highly nonlinear, nonautonomous and singularly perturbed with respect to the small parameter $\mu = \mathcal{E}^{-1}$. Its principal part as well as the boundary conditions become degenerate at $s = 0$. These combined features result in a highly oscillatory behavior of weak solutions. The goal of the present analysis is to encode oscillations persistent at the limit $\mu \to 0$ into Young measures.

We study a plane Poiseuille flow, which is driven by a prescribed pressure gradient, with vanishing velocity field at the boundary. The variable ϕ corresponds to the angle between the unit molecular director, $\mathbf{n} = (\sin\phi, 0, \cos\phi)$, and the velocity, $\mathbf{v} = (0, 0, v(x))$, of the flow. The variable degree of orientation, $s \in (-\frac{1}{2}, 1)$, gives the quality of alignment of the molecules with the director field, with $s = 1$ corresponding to perfect alignment, and $s = -\frac{1}{2}$ describing the case with molecules placed on a plane perpendicular to $\mathbf{n}$. Especially relevant to the present study is the isotropic case, $s = 0$, with randomly oriented molecules. Points, lines or planes in the flow region with $s = 0$ correspond to nematic liquid crystal defects, with undefined ϕ. Moreover, ϕ becomes discontinuous across defect lines and planes. The variables s and $\mathbf{n}$ correspond to an

*Received by the editors February 1, 2005; accepted for publication (in revised form) June 12, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/sima/38-5/62373.html
†Department of Mathematics, University of Minnesota, Minneapolis, MN 55455 (mcc@math.umn.edu). This author's work has been supported in part by the NSF Contract DMS-0128832.
‡Department of Mathematics, Washington State University, Pullman, WA 99163 (panchenko@math.wsu.edu). This author's work has been supported in part by the DOE grant DE-FG02-05ER25709.

eigenvalue and eigenvector, respectively, of the optically uniaxial and traceless order tensor Q. The latter represents a second order moment of the molecular orientation field of a rigid polymer.

The equations that we analyze follow from those derived by Ericksen to model flow of liquid crystals with variable degree of orientation [8]. They yield the Leslie–Ericksen equations when the order parameter s is taken to be a nonzero constant. Relevant features of the model include the Helmholtz free energy and the viscous, anisotropic, stress tensor. The latter is characterized by a set of anisotropic viscosity functions, $\alpha_i(s)$, $1 \leq i \leq 6$, known as Leslie coefficients (in particular, $\frac{1}{2}\alpha_4(0)$ represents the Newtonian viscosity). The free energy density is of the form $a_1|\nabla s|^2 + a_2 s^2|\nabla \mathbf{n}|^2 + \mathcal{J}^{-1} f(s)$, with a_1, a_2 and $\mathcal{J}$ denoting positive, dimensionless material parameters. The scalar function $f(s)$ represents a multiwell potential, favoring special directions of alignment at equilibrium. As a result of the elastic and viscous contributions to the model, the nature of the flow is fully non-Newtonian.

The flow behavior is determined mostly by three nondimensional parameter groups, the *Reynolds* number $\mathcal{R}$, the *Interface* number $\mathcal{I}$, and the *Ericksen* number $\mathcal{E}$. The latter measures the ratio of the viscous torque of the flow with respect to the elastic one. The condition of $\mathcal{E}$ being large corresponds to flow with large pressure gradient, and also to the case of viscous torque dominating the elastic one. The parameter $\mathcal{I}$ is associated with the free energy required to maintain defects in the flow, and it corresponds to the quotient of the bulk elastic energy and the gradient part of the Helmholtz free energy. The quantity $\mathcal{J}^{-1} = \mathcal{I}\mathcal{E}^{-1}$ appears as a coefficient in the free energy, as previously indicated. The derivation of the model studied in this article, the physical and nondimensional parameter groups, can be found in [3], [4], and [5].

We observe that s identically zero is not a solution of the problem. For an arbitrary flow domain, the bulk isotropic state, $s \equiv 0$, can only be realized at equilibrium, if permitted by the boundary conditions. In the case of Poiseuille flow, prescribing a nonzero pressure gradient excludes $s = 0$ from being an equilibrium solution. Consequently, one of the main outcomes of our study is to show that the isotropic state can be nearly realized in an effective sense.

Intuitively, one expects that for large Ericksen number, viscosity effects are dominant and therefore, the molecular alignment (associated with $s \neq 0$) is destroyed. The absence of alignment is indicated by $s = 0$. In this sense, the limit of large Ericksen number should represent the transition from order to disorder. In section 5, we show (Theorem 5.1) that there is indeed a sequence of generalized solutions such that in the limit $\mu \to 0$, s tends to zero uniformly and v becomes the Newtonian velocity field of Poiseuille flow. This alone would indicate a perfect isotropic limit. However, the Young measure generated by ϕ satisfies the additional moment relations, indicating a residual molecular alignment. Although at a macroscopic scale the flow is isotropic, a liquid crystalline microstructure is present. The oscillatory behavior of solutions at the limit of large Ericksen number was numerically detected by the simulations performed in [14]. This provided the motivation for the present study.

In addition to the governing system being singularly perturbed as $\mu \to 0$, it becomes degenerate at $s = 0$. One mathematical difficulty is that standard methods of analysis of singular perturbations [7] cannot be readily applied here. On the other hand, owing to the singularly perturbed nature of the system, a priori bounds for the derivatives are not uniform in μ and therefore, embedding theorems cannot be applied to obtain compactness of s. We overcome these difficulties by constructing tight bounds in terms of supersolutions and subsolutions. They are both solutions of a

classical variational problem for particle motion in a central force field [1]. The upper and lower bounds on s so obtained tend to 0 as $\mu \to 0$, which entails compactness. This leads to partial stability in the sense that s converges to zero uniformly as $\mu \to 0$, while ϕ oscillates with increasing frequency.

Another interesting aspect of the system being driven to degeneracy at the limit $\mu \to 0$ manifests itself in the prescription of boundary conditions. Since the system is of second order with respect to s and ϕ, one expects that both $s(\pm 1)$ and $\phi'(\pm 1)$ (or $\phi(\pm 1)$) should be prescribed. We choose $s(\pm 1) = \mu$. However, in passing to the limit $\mu \to 0$, we find that the boundary conditions $\phi'(\pm 1)$ cannot be freely chosen, but depend on $s(\pm 1)$. In addition, $|\phi'(\pm 1)| \to \infty$ as $\mu \to 0$, in agreement with the fact that our limiting process amounts to creating a boundary defect. The unboundedness of ϕ' causes ϕ to be undefined at the defect location as expected. So, boundary values for the angular variable become redundant at the limit. The choice of $s(\pm 1) = \mu$ does not detract from generality. Our purpose is to study the oscillatory phenomenon on the bulk and avoid boundary layer contributions that would appear in the case that s is not driven to zero at the boundary. These contributions could be incorporated in the current analysis using the techniques previously developed in [3], [4], and [5], but we do not attempt such an analysis here.

The article is organized as follows. Following section 1, the statement of the problem is presented in section 2. We developed a priori bounds and necessary conditions for Young measures in section 3. The main technical result of the paper is stated in section 4, with Theorem 4.3 being the focus of the work. A discussion of the effective equations is presented in section 5, with Theorem 5.1 stating the physical conclusions.

2. Formulation of the problem. We study the following system of differential equations on the interval $I = (-1, 1)$:

$$\mu \left(a_1 s'' - a_2 s(\phi')^2\right) = G_1(s, \phi, x), \tag{2.1}$$

$$\mu a_2 (s^2 \phi')' = G_2(s, \phi, x), \tag{2.2}$$

where

$$G_1(s, \phi, x) = \frac{1}{2}\beta_1(s) g^{-1}(s, \phi) x \sin 2\phi + \frac{1}{\mathcal{J}} \frac{df}{ds}(s), \tag{2.3}$$

$$G_2(s, \phi, x) = \frac{1}{2}(\gamma_1(s) + \gamma_2(s) \cos 2\phi) g^{-1}(s, \phi) x. \tag{2.4}$$

Equations (2.1)–(2.4) model plane Poiseuille flow of nematic liquid crystals with variable degree of orientation. The previously introduced functions $s(x)$ and $\phi(x)$ are the unknown fields of the problem. In (2.1)–(2.4), $'$ denotes derivative with respect to the independent variable $x \in I$. The boundary conditions are prescribed as follows:

$$s(-1) = \mu, \quad s(1) = \mu, \tag{2.5}$$

$$\phi'(-1) = A(\mu), \quad \phi'(1) = B(\mu). \tag{2.6}$$

As mentioned in section 1, $A(\mu)$ and $B(\mu)$ cannot be arbitrarily prescribed for sufficiently small $\mu > 0$. The system (2.1)–(2.4) contains positive scalar parameters μ, $\mathcal{J}$, a_1, and a_2 discussed in section 1. For more details on these quantities, see [3], [4], and [5].

Let $\tau > 0$ be a given material parameter. We suppose that β_1, γ_1, and γ_2 are smooth functions of s on the interval $(-\frac{1}{2}, 1)$, and that $g(s, \phi)$ is also smooth. We

assume that the following hypotheses hold:

$$\beta_1 < 0, \quad \gamma_1 > 0, \tag{2.7}$$

$$\gamma_1(s) = O(s^2), \quad \gamma_2(s) = O(s) \quad \text{for } s \text{ close to } 0, \tag{2.8}$$

$$g(s,\phi) \geq \tau > 0. \tag{2.9}$$

It follows from assumptions (2.8) and (2.9) that $\frac{1}{s}G_2(s)$ is a smooth function of s and satisfies

$$\frac{1}{s}G_2(s) = O(1), \tag{2.10}$$

for $|s|$ small. Equations (2.1)–(2.9) completely describe the problem studied in the paper. The key observation that enables us to overcome the lack of uniform bounds mentioned in section 1 is as follows. If the right-hand sides of (2.1) and (2.2) are replaced with $\frac{1}{\mathcal{J}}df/ds$ and zero, respectively, and, in addition, $a_1 = a_2$, one obtains the classical system modeling motion of a particle in a central force field with potential $-\frac{1}{\mathcal{J}}f(s)$ (see, e.g., [1]). In that case, the variable x is time and $s(x), \phi(x)$ are the radial and angular particle coordinates, respectively. This system has two first integrals: the angular momentum and the total mechanical energy, which make the problem completely integrable. In this paper, this classical system (with different potentials) is used to construct sub- and supersolutions of (2.1), (2.2) (see the proof of Theorem 4.3).

In the remainder of this section we outline the derivation of the governing system from the Leslie–Ericksen equations studied previously in [3], [4], and [5]). The equations considered in these papers are as follows:

$$\mu\left(a_1 s'' - a_2 s(\phi')^2\right) = \beta_1(s)v' \sin\phi\cos\phi + \frac{1}{\mathcal{J}}\frac{df}{ds}(s) \tag{2.11}$$

$$\mu a_2(s^2\phi')' = \left(\gamma_1(s) + \gamma_2(s)\cos 2\phi\right)v' \tag{2.12}$$

$$\frac{1}{\mathcal{R}}(g(s,\phi)v')' = 1, \tag{2.13}$$

Equation (2.13) corresponds to the balance of linear momentum. Its right-hand side represents the prescribed pressure gradient and $\mathcal{R}$ is the Reynolds number. Since we are interested in the behavior of solutions when $\mathcal{E}$ is large and the Reynolds number $\mathcal{R}$ is moderate, there is no loss of generality in assuming $\mathcal{R} = 1$. The function $\beta_1(s)$ is a Leslie coefficient, and $\gamma_1(s), \gamma_2(s)$, and $g(s,\phi)$ are given in terms of $\alpha_i(s)$, $i = 1, \ldots, 5$, as follows:

$$\gamma_1 := \alpha_3 - \alpha_2, \quad \gamma_2 := \alpha_2 + \alpha_3, \tag{2.14}$$

$$\begin{aligned} g(s,\phi) = {} & \frac{1}{2}\alpha_4(s) + \alpha_1(s)\sin^2\phi\cos^2\phi \\ & + \frac{1}{2}(\alpha_5 - \alpha_2)(s)\sin^2\phi + \frac{1}{2}\alpha_3(s)\cos^2\phi. \end{aligned} \tag{2.15}$$

In section 5, we shall make use of the fact (see, e.g., [3], [4], and [5]) that $\alpha_j(0) = 0$ for $j \neq 4$. This implies

$$g(0,\phi) = \frac{1}{2}\alpha_4(0) > 0, \tag{2.16}$$

where $\frac{1}{2}\alpha_4(0)$ is the Newtonian viscosity. The second law of thermodynamics requires that the system be dissipative. This imposes inequality restrictions on the Leslie coefficients $\beta_1(s)$ and $\alpha_i(s)$ (see [12]). As a consequence, we have that $\beta_1 < 0, \gamma_1 > 0$ and

$$g(s,\phi) \geq 0, \quad s \in \left[-\frac{1}{2}, 1\right], \quad \phi \in \mathbf{R}. \tag{2.17}$$

For the forthcoming analysis, we will impose a stronger assumption (2.9), where $\tau > 0$ is a material parameter related to the Newtonian viscosity. Such a strict inequality is an immediate consequence of $\mathcal{E} \neq 0$ [3]. Additional assumptions on Leslie coefficients consistent with kinetic theory of polymers are discussed in [3] and [4].

Next, we solve (2.13) for the velocity v, obtaining

$$v'(x) = (x + C)g^{-1}(s(x), \phi(x)), \tag{2.18}$$

where C is a constant of integration. The boundary conditions $v(\pm 1) = 0$ imply that

$$v(x) = \int_{-1}^{x} tg^{-1}(s(t), \phi(t))\, dt + C \int_{-1}^{x} g^{-1}(s(t), \phi(t))\, dt,$$

and

$$C = -\left(\int_{-1}^{1} g^{-1}(s(t), \phi(t))\, dt\right)^{-1} \int_{-1}^{1} tg^{-1}(s(t), \phi(t))\, dt. \tag{2.19}$$

It should be noted that C is a functional of s, ϕ taking constant values for a given flow. A calculation using the positivity of g yields $-1 < C \leq 1$. Postulating translational invariance of the equations, it is not difficult to show that there exists an interval J (which may be different for different μ) such that $C = 0$ when the equations are considered on J. Calculations involving C are analogous to those with the remaining terms. So, without loss of generality, we will set $C = 0$ in the equations. We also point out that this does not change the qualitative behavior of solutions because of the translational invariance. Combining (2.11), (2.12), and (2.18), we obtain the reduced system (2.1)–(2.4).

3. A priori bounds and necessary conditions for Young measures. In this section we introduce the concept of Young measures generated by sequences of weak solutions of the problem. We derive integral identities and a priori estimates satisfied by weak solutions. Passing to the limit $\mu \to 0$ in the weak formulation of the equations yields algebraic momentum relations satisfied by the Young measures. Such relations can be appropriately interpreted as the effective equations of the system.

A weak solution of the system (2.1)–(2.2) is a pair of functions (s, ϕ) such that $s \in W^{1,2}(I)$, $s\phi \in W^{1,2}(I)$, and for all test functions $h \in C_0^1(I)$, the integral identities

$$-\mu a_1 \int_I s'h'dx - \mu a_2 \int_I s(\phi')^2 h dx = \int_I G_1 h dx, \tag{3.1}$$

$$-\mu a_2 \int_I s^2 \phi' h' dx = \int_I G_2 h dx \tag{3.2}$$

hold, where the right-hand sides G_1 and G_2 are given by (2.3) and (2.4), respectively.

For each $\mu > 0$, existence of weak solutions satisfying $s \in (-\frac{1}{2}, 1), |\phi| \leq \frac{\pi}{2}$ together with the boundary conditions (2.5), (2.6) can obtained using the existence theorem in [2].

3.1. L^2-bounds. The L^2-estimates will be obtained from the following proposition.

PROPOSITION 3.1. *Let s, ϕ be a sufficiently smooth solution of* (2.1) *and* (2.2). *Then the following identities hold:*

$$-\mu a_1 \int_I |s'|^2\,dx + \mu a_1(s's(1) - s's(-1)) - \mu a_2 \int_I s^2(\phi')^2\,dx = \int_I G_1(x,s,\phi)s\,dx, \tag{3.3}$$

$$-\mu a_2 \int_I s^2(\phi')^2 dx + \mu a_2((s^2\phi'\phi)(1) - (s^2\phi'\phi)(-1)) = \int_I G_2(x,s,\phi)\phi dx. \tag{3.4}$$

Proof. Multiplying (2.1) by s and integrating by parts yields

$$\mu a_1 \int_I (s')^2 dx + \mu a_1(s's(1) - s's(-1)) + \mu a_2 \int_I s^2|\phi'|^2 dx = \int_I G_1(x,s,\phi)s dx.$$

Likewise, multiplication of (2.2) by ϕ and integration by parts yields (3.4). ∎

The uniform boundedness of G_1 and G_2 with respect to μ follows from the analogous property of s and ϕ. If, in addition, the boundary terms $b_1(\mu) \equiv \mu a_1(s(1)s'(1) - s(-1)s'(-1))$ and $b_2(\mu) \equiv \mu a_2((s^2\phi'\phi)(1) - (s^2\phi'\phi)(-1))$ are also uniformly bounded, then the identities (3.3) and (3.4) yield the following uniform a priori bounds:

$$\begin{aligned} \mu^{\frac{1}{2}} \|s'\|_{L^2(I)} &\le C, \\ \mu^{\frac{1}{2}} \|s\phi'\|_{L^2(I)} &\le C, \end{aligned} \tag{3.5}$$

where $C > 0$ is independent of μ.

Note that $b_1(\mu)$ and $b_2(\mu)$ vanish when $s(\pm 1) = 0$. Otherwise, if $s'(\pm 1)$ and $\phi'(\pm 1)$ do not grow too fast as $\mu \to 0$, and s and ϕ are uniformly bounded, then $b_1(\mu)$ and $b_2(\mu)$ are also uniformly bounded. Such a statement follows from the estimates on the boundary layer terms valid for a large class of boundary conditions. Indeed, near the boundary $x = \pm 1$, s can be well approximated by a boundary layer term S, satisfying $S(x) = O(e^{\frac{-|x-1|}{\mu^{1/2}}})$, for μ close to 0 (cf. [3] and [4]). Moreover,

$$|s'(\pm 1)| \le \frac{C}{\mu^{1/2}}, \tag{3.6}$$

where C is independent of μ. A related property of the solutions of the governing equations for small μ is the oscillatory behavior of s about $s = 0$ on the interval I; an estimate on the number of oscillations gives $N = O(\mu^{-\frac{1}{2}})$. Moreover, the first and the last zeroes of s in I approach the boundary as $\mu \to 0$ [2]. Numerical simulations of such a behavior are presented in [14].

3.2. Momentum relations for Young measures. The estimates (3.5) allow for some control of the oscillations and yield existence of the microscopic length scale $l \sim \mu^{1/2}|I|$. Unfortunately, the bounds on the derivatives are not uniform in μ, so

it is not possible to extract subsequences convergent weakly in $W^{1,2}(I)$ as $\mu \to 0$. Since s, ϕ are bounded pointwise, we can extract subsequences weak-*convergent in $L^\infty(I)$, and hence weakly convergent in $L^2(I)$. However, this convergence cannot be improved to strong convergence due to high oscillatory behavior of s and ϕ. Such behavior is appropriately encoded in the Young measure generated by sequences of weak solutions [20]. In [17] this measure is defined as follows.

DEFINITION 1. *A Young (parameterized) measure is a family of probability measures $\lambda = \{\lambda_x\}_{x\in\Omega}$ associated with a sequence of functions $f_j : \Omega \subset \mathbf{R}^N \to \mathbf{R}^m$ such that*

(i) $\mathrm{supp}(\lambda_x) \subset \mathbf{R}^m$;

(ii) *λ_x depend measurably on $x \in \Omega$, which means that for any continuous function $\phi : \mathbf{R}^N \to \mathbf{R}$, the function*

$$\bar{\phi}(x) = \int \phi(y) d\lambda_x(y) = \langle \phi(y), \lambda_x(y) \rangle$$

is (Lebesgue) measurable;

(iii) *if the sequence $\phi(f_j)$ converges weakly in $L^p(\Omega)$, $1 \le p < \infty$ (weak-$\star$ in $L^\infty(\Omega)$), then the weak limit is the function*

$$\bar{\phi}(x) = \int \phi(\xi) d\lambda_x(\xi).$$

In what follows, we use the following facts about Young measures. The first [17, Theorem 6.2] is the existence theorem.

THEOREM 3.2. *Let $\Omega \subset \mathbf{R}^N$ be a (Lebesgue) measurable set and let $z_j : \Omega \to \mathbf{R}^m$ be measurable functions such that*

$$\sup_j \int_\Omega g(|z_j|) dx < \infty,$$

where $g : [0, \infty]$ is a continuous, nondecreasing function such that $\lim_{t\to\infty} g(t) = \infty$.

Then there exists a subsequence, not relabeled, and a family of probability measures, $\lambda = \{\lambda_x\}_{x\in\Omega}$ (the associated Young measure) with the property that whenever the sequence $\{\psi(x, z_j(x))\}$ is weakly convergent in $L^1(\Omega)$ for any Caratheodory function $\psi(x, \xi) : \Omega \times \mathbf{R}^m \to [-\infty, \infty]$, the weak limit is the function

$$\bar{\psi}(x) = \int_{\mathbf{R}^m} \psi(x, \xi) d\lambda_x(\xi).$$

The second fact concerns Young measures generated by sequences of vector functions, for which strong convergence holds for some of the components, but not for all of them [17, Proposition 6.13].

THEOREM 3.3. *Let $z_j = (u_j, v_j) : \Omega \to \mathbf{R}^d \times \mathbf{R}^m$ be a bounded sequence in $L^p(\Omega)$ such that $\{u_j\}$ converges strongly to u in $L^p(\Omega)$. If $\lambda = \{\lambda_x\}_{x\in\Omega}$ is the Young measure associated with z_j, then $\lambda_x = \delta_{u(x)} \otimes \nu_x$ for (Lebesgue) almost all $x \in \Omega$, where $\{\nu_x\}_{x\in\Omega}$ is the Young measure corresponding to $\{v_j\}$.*

Young measures provide a description of the weak limits of the nonweakly continuous functions G_1 and G_2 in (2.1) and (2.2). We denote by λ_x the Young measure associated with the sequences of weak solutions (s^μ, ϕ^μ). By definition, for any

continuous function H on $[-1/2,1]\times[-\pi/2,\pi/2]$, the sequence $H(s^\mu,\phi^\mu)$ converges weakly-* to

$$\bar{H}(x)=\int_{-1/2}^{1}\int_{-\pi/2}^{\pi/2}H(s,\phi)d\lambda_x(s,\phi).$$

It is a well known fact [17] that Young measures are in general difficult to compute. In particular, there is no general method for derivation of conditions imposed on Young measures by the nonlinear differential constraints on generating sequences. Available results concern sequences generated by gradients [10]. A generalization of [10] to sequences solving certain constant-coefficient linear partial differential equations is obtained in [9].

In the present case it is possible to derive some necessary conditions for Young measures passing to the limit in the integral identities (3.1) and (3.2). In the next proposition we use notation $a=(a_2/a_1)^{1/2}$.

PROPOSITION 3.4. *Let λ_x be the Young measure generated by a sequence of weak solutions (s^μ,ϕ^μ) satisfying the estimates* (3.5). *Then for any test function $h\in C_0^1(I)$, the following relations hold:*

$$\int\left(G_1(x,y,z)\sin az+\frac{1}{ay}G_2(x,y,z)\cos az\right)d\lambda_x(y,z)h(x)dx=0, \tag{3.7}$$

$$\int\left(G_1(x,y,z)\cos az-\frac{1}{ay}G_2(x,y,z)\sin az\right)d\lambda_x(y,z)h(x)dx=0. \tag{3.8}$$

Proof. Let $w(x)$ be a test function. In the integral identities (3.1), we first replace h with the test function $\sin(a\phi)w$. Next, we formally substitute $\frac{1}{s}a\cos(a\phi)w$ for h in (3.2) and add the resulting identities. This yields,

$$-\mu a_1\int(s\sin a\phi)'w'dx=\int\left(G_1\sin a\phi+\frac{1}{as}G_2\cos a\phi\right)wdx. \tag{3.9}$$

To justify the use of $\frac{1}{s}a\cos(a\phi)w$ as a test function, we point out that $\frac{1}{s}G_2$ is a smooth function of s and satisfies (2.10). Thus, the right-hand side of the previous equation is well defined. This allows us to approximate $\frac{1}{s}$ by a sequence of test functions and then pass to the limit in the resulting integral identities.

The estimates (3.5) on derivatives yield

$$\|\,\mu^{1/2}(s\sin a\phi)'\,\|_{L^2(I)}\le C$$

with C independent of μ. Hence the integral on the left tends to zero as $\mu\to 0$. Using the definition of Young measures to pass to the limit in the right-hand side we obtain the first equation in (3.7).

Next, using $\cos a\phi w$ as a test function in (3.1) and $-\frac{A}{s}\sin a\phi w$ in (3.2) and summing up, we have

$$-\mu a_1\int(s\cos a\phi)'w'\,dx=\int\left(G_1\cos a\phi-\frac{1}{as}G_2\sin a\phi\right)w\,dx. \tag{3.10}$$

Again, the limit of the integral on the left is zero. Passing to the limit in (3.10) yields the second equation in (3.7). ∎

Remark 3.1. The method of proof is based on the following formal procedure. Consider the original system (2.1)–(2.2):

$$\mu(a_1 s'' - a_2 s(\phi')^2) = G_1(x, s, \phi),$$
$$\mu a_2 (s^2 \phi')' = G_2(x, s, \phi).$$

If we multiply the first equation of the system by $\sin a\phi$, the second equation by $1/s \cos a\phi$, and add the resulting equations, we obtain

$$\mu a_1 (s \sin a\phi)'' = G_1 \sin a\phi + \frac{G_2}{s} \cos a\phi. \tag{3.11}$$

Similarly, multiplying the first equation by $\cos a\phi$, the second one by $-1/s \sin a\phi$, and adding the results, we get

$$\mu a_1 (s \cos a\phi)'' = G_1 \cos a\phi - \frac{G_2}{s} \sin a\phi. \tag{3.12}$$

The momentum relations from the proposition are obtained by passing to the limit (in the sense of distributions) in the integral identities corresponding to the system (3.11), (3.12), and thus they are the effective equations for this new system. Moreover, relations (3.7), (3.8) are also effective equations for the original system, since their solutions satisfy the differential equations almost everywhere, and, for each $\mu > 0$, the zero sets of the functions $\cos a\phi$, $\sin a\phi$, and s can be shown to be countable by standard Sturm–Liouville results as in [2].

Remark 3.2. It is interesting to ask to what extent the relations from the proposition characterize Young measures generated by sequences of weak solutions of (2.1) and (2.2). The measures in question satisfy (3.7) and (3.8), but these conditions are far from being sufficient, since they can be obtained for Young measures generated by different equations. These equations may contain, for instance, terms with linear combinations of higher-order derivatives of s and $s\phi'$ multiplied by sufficiently large powers of μ.

4. Generalized solutions and isotropic defects. In this section, we construct generalized solutions s^μ, ϕ^μ from a special class of weak solutions. Generalized solutions are such that s^μ approaches 0, uniformly on I, and ϕ^μ is bounded and presents multiple jump discontinuities. Such jump discontinuities may correspond to isotropic (plane) defects in stationary three dimensional flow.

The numerical experiments carried out in [14] indicate partial stability, which means that $s^\mu \to 0$ as $\mu \to 0$, and ϕ^μ oscillates with a frequency of the order of $\mu^{-1/2}$. In this subsection, we obtain sufficient conditions for such type of behavior. For this, we appeal to results from the theory of ordinary differential equations; the main technical tool is Nagumo's theorem [15], (stated as Theorem 4.1). This theorem provides sufficient conditions for existence of solutions satisfying pointwise upper and lower bounds, constructed from sub- and supersolutions. This result allows us to prove Theorem 4.3 on existence of partially stable solutions.

The strategy of the proof of Theorem 4.3 is as follows. First, we consider arbitrary solutions satisfying boundary conditions that depend on μ (see Theorem 4.3 for the precise formulation). Then, we show that among these, there is a large class of boundary conditions for which s^μ is bounded by a multiple of μ. Nagumo's theorem guarantees existence of such solutions. To apply the theorem, we first show that sub- and supersolutions of the system can be constructed from solutions of a classical variational problem that models the motion of a particle in a central force field.

The solutions of the variational systems depend on the first integral parameters. By appropriate choice of the parameters, we find sub- and supersolutions satisfying the desired bounds. An additional difficulty that we encounter is that Nagumo's theorem concerns classical solutions, while the present system may have singularities near the points where $s = 0$. We deal with this problem by examining solutions only on the set where the absolute value of s^μ is larger than μ, and then use upper and lower bounds to demonstrate that, on this set, $|s^\mu| \le 2\mu$.

Nagumo's theorem concerns solutions of the Dirichlet problem on the interval $[a, b]$ for a single second order equation

$$\begin{aligned} u'' &= F(x, u, u'), \\ u(a) &= u_0, \qquad u(b) = u_1. \end{aligned} \tag{4.1}$$

Assume that F is a continuous function of its arguments satisfying the condition

$$F(x, u, z) = O(|z|^2) \ \text{ as } \quad |z| \to \infty \tag{4.2}$$

for all (x, u) in a rectangle $[a, b] \times [\alpha, \beta]$. The following theorem is due to Nagumo [15].

THEOREM 4.1. *Suppose that F satisfies* (4.2)*, and there exist functions $\alpha(x), \beta(x)$ with the properties*

(i) $\alpha, \beta \in C^2([a, b])$;
(ii) $\alpha(x) \le \beta(x)$;
(iii) $\alpha'' \ge F(x, \alpha, \alpha')$, $\beta'' \le F(x, \beta, \beta')$;
(iv) $\alpha(a) \le u_0 \le \beta(a)$, $\alpha(b) \le u_1 \le \beta(b)$. *Then the problem* (4.1) *has a solution $u(x) \in C^2([a, b])$ such that*

$$\alpha(x) \le u(x) \le \beta(x) \tag{4.3}$$

on $[a, b]$.

This theorem has been used by Howes and Chang [7] to study stability of singularly perturbed ODE. We note that condition (4.2) is not the most general one (see [7] for other types of conditions), but it is sufficient for our purpose. We point out that stability conditions for systems studied in [7] do not apply to the present case, since the vector solution (s, ϕ) is not expected to be stable.

According to this theorem, to show that $s^\mu \to 0$, it is enough to construct sequences of bounds $\alpha^\mu(x), \beta^\mu(x)$ for solutions s^μ of (2.1) which converge to zero uniformly on I. Since we are dealing with a system, the bounds for s must be uniform in ϕ. First, we consider the case when $\frac{1}{\mathcal{J}}$ tends to infinity as μ approaches zero.

THEOREM 4.2. *We assume that hypotheses* (2.7)–(2.9) *are satisfied. Let*

$$\frac{1}{\mathcal{J}(\mu)} \to \infty \quad \text{as} \quad \mu \to 0.$$

Suppose that $f \in C^1(-\frac{1}{2}, 1)$ has an isolated local minimum at $s = 0$ and ϕ is an arbitrary differentiable function on I.

Then, there exist a decreasing sequence $\{\mu_n\} \subset (0, 1]$, $\mu_n \to 0$ as $n \to \infty$ and sequences of real numbers $\mathcal{L}_n$ and $\mathcal{M}_n$, such that for every n, there exists a solution s_n to the scalar boundary value problem

$$\mu_n a_1 s'' = \mu_n a_2 s(\phi')^2 - \frac{1}{2}\beta_1(s)x\frac{\sin 2\phi}{g(s, \cos 2\phi)} + \frac{1}{\mathcal{J}(\mu_n)}\frac{df}{ds}(s), \tag{4.4}$$

$$s(-1) = \mathcal{L}_n, \quad s(1) = \mathcal{M}_n,$$

with the property that $s_n \to 0$, *as* $n \to \infty$, *uniformly on* I. *Moreover, the rate of convergence is independent of* ϕ.

Proof. By the assumptions on f, there exists a fixed interval E, containing zero, such that $\frac{df}{ds} > 0$ for $s > 0$, and $\frac{df}{ds} < 0$ for $s < 0$. Let us choose a strictly decreasing sequence $\{\sigma_n\} \subset E$, $\sigma_n > 0$, and such that $\lim_{n\to\infty} \sigma_n = 0$. Likewise, we select an increasing sequence $\{\alpha_n\} \subset E$, $\alpha_n < 0$, and such that $\lim_{n\to\infty} \alpha_n = 0$. For each positive integer n, we can find a real number $\mu_n > 0$ such that

$$\mu a_2 \sigma_n (\phi')^2 + \frac{1}{\mathcal{J}}(\mu)\frac{df}{ds}(\sigma_n) - \frac{1}{2}\beta_1(\sigma_n)x\frac{\sin 2\phi}{g(s,\cos 2\phi)} \geq 0 \quad \text{and}$$

$$\mu a_2 \alpha_n (\phi')^2 + \frac{1}{\mathcal{J}}(\mu)\frac{df}{ds}(\alpha_n) - \frac{1}{2}\beta_1(\alpha_n)x\frac{\sin 2\phi}{g(s,\cos 2\phi)} \leq 0$$

for all $0 < \mu \leq \mu_n$. Indeed, by assumptions (2.7)–(2.9), the last term on the left-hand side of the previous expressions is bounded by a constant $\frac{1}{2}\frac{m}{\tau}$, where $m > 0$ is a bound on $x\beta_1$. For each $n > 0$, we now choose $\alpha_n < \mathcal{L}_n, \mathcal{M}_n < \sigma_n$. The conclusion follows from Theorem 4.1, if we use the constant functions σ_n and α_n as, respectively, upper and lower bounds. ∎

Remark 4.1. We point out that the condition $\mathcal{J} \to 0$ (i.e., $\frac{1}{\mathcal{J}} \to \infty$) as $\mu \to 0$ represents the Leslie–Ericksen limit of the theory, that is the case when the liquid crystal is described by the director $\mathbf{n}$, with the order parameter taking a constant value corresponding to a critical point of the bulk energy $f(s)$.

Next, we consider the more subtle case of $\frac{1}{\mathcal{J}}$ bounded. We assume that $\frac{1}{\mathcal{J}}$ and f satisfy the following conditions:

(i) $\frac{1}{\mathcal{J}}(\mu)$ is bounded by a constant M for all $\mu \leq 1$,
(ii) $f(s) \in C^2(-\frac{1}{2}, 1)$,
(iii) there exist $a, b \in (-\frac{1}{2}, 1)$ such that $|\frac{df}{ds}(s)| \leq M$ for $s \in [a, b]$, $\frac{df}{ds}(s) > 0$ on $(b, 1)$ and $\frac{df}{ds}(s) < 0$ on $(-\frac{1}{2}, a)$.

We observe that the last condition on f may allow several potential wells between $s = -\frac{1}{2}$ and $s = 1$. This may include pure nematic liquid crystals as well as compounds (see [6, chapter 1]).

THEOREM 4.3. *We assume that hypotheses* (2.7)–(2.9) *are satisfied. Suppose that conditions* (i), (ii), *and* (iii) *are fulfilled. Consider sequences* $\{s^\mu, \phi^\mu\}$ *of weak solutions of* (2.1)–(2.2) *such that*

$$s(-1) = s(1) = \mu \tag{4.5}$$

hold. Then there exist $A(\mu), B(\mu) \in \mathbf{R}$ *such that sequences* $\{s^\mu, \phi^\mu\}$ *additionally satisfying*

$$(\phi^\mu)'(-1) = A(\mu), \quad (\phi^\mu)'(1) = B(\mu), \tag{4.6}$$

have the property that $s^\mu \to 0$ *as* $\mu \to 0$, *uniformly on* I.

Proof. Fix $\mu > 0$ and consider solutions (s, ϕ) of the problem

$$\mu a_1 s'' = \mu a_2 s(\phi')^2 - \frac{\beta_1 x \sin 2\phi}{2g(s,\phi)} + \frac{1}{\mathcal{J}}\frac{df}{ds}(s), \tag{4.7}$$

$$\mu a_2 (s^2\phi')' = \frac{x}{g}(\gamma_1(s) + \gamma_2(s)\cos 2\phi), \tag{4.8}$$

satisfying boundary conditions (4.5) and (4.6), with the values $A(\mu)$ and $B(\mu)$ to be specified later. Although this solution is weak, the estimates (3.5) imply continuity of s.

First, we integrate (4.8) to obtain

$$s^2(x)\phi'(x) = \frac{1}{\mu a_2}(\mu^2 A + p(x)), \tag{4.9}$$

where

$$p(x) = \int_{-1}^{x} \frac{y}{g(s(y),\phi(y))}(\gamma_1(s(y)) + \gamma_2(s(y))\cos 2\phi(y))\, dy. \tag{4.10}$$

Next, we represent I as the union of the sets $E^+ = \{x \in I : s > \mu\}$, $E^- = \{x \in I : s < \mu\}$ and $E_0 = \{x \in I : |s| \le \mu\}$, and define s^+ and s^- to be restrictions of s to E^+, E^-, respectively. Since E^+ is open, it can be represented as $E^+ = \cup_{j=1}^{\infty} I_j$, where $I_j = (a_j, b_j)$ are open intervals, and $s(a_j) = s(b_j) = \mu$. Let us consider the system (4.7), (4.8) on some interval I_j. Since $s \ge \mu$ on I_j, we can obtain s from (4.9) and substitute it into (4.7):

$$\mu a_1 s'' = \frac{1}{\mu a_2 s^3}\left(\mu^2 A + p(x)\right)^2 + \frac{\beta_1(s) x \sin 2\phi}{2g(s,\phi)} + \frac{1}{\mathcal{J}}\frac{df}{ds}(s). \tag{4.11}$$

From now on, the strategy of the proof is as follows.

1. Use Nagumo's theorem to construct, for each choice of ϕ, a solution u of (4.11) on I_j, satisfying the boundary conditions (4.5), and such that $-2\mu \le u \le 2\mu$ on I_j. To obtain existence of u, we shall choose A, B sufficiently large. The same choice of A, B works for all I_j, which provides u on E^+. The analogous argument yields existence of u on E^-.
2. Once the choice of A, B is made, the boundary conditions (4.5), (4.6) are specified, and we obtain a particular solution (S, Φ) of (4.7) and (4.8).
3. We prove that u, corresponding to the choice $\phi = \Phi$ in (4.11), coincides with S on the intervals where $|S| > \mu$. This gives an upper bound on S.
4. For each μ, define $s^\mu = S$, $\phi^\mu = \Phi$.

Now we begin implementing the previously described strategy. First, note that the right-hand side of (4.11) is a continuous function of s when $x \in I_j$, and the Nagumo condition (4.2) is satisfied, since the right-hand side is independent of s'. The bounds on u must be independent of the choice of ϕ, which would allow us to adjust ϕ later on when $A = \phi'(-1)$ is chosen. We begin with the construction of the upper bound. To construct the bound, we use the comparison variational system

$$\mu a_1 q'' = \mu a_2 q(\psi')^2 - \frac{dh}{dq}(q), \tag{4.12}$$

$$\mu a_2 (q^2\psi')' = 0. \tag{4.13}$$

This system is classical (see [1]), at least when $a_1 = a_2$. It describes the motion of a particle in a central force field with the potential h. In that case, x denotes time and q, ψ represent the radial and angular particle coordinates, respectively. The following properties of (4.12) and (4.13) are well known. The system has two first integrals, the angular momentum and the total energy, respectively:

$$M = q^2\psi',$$
$$E = \frac{1}{2}\mu a_1 (q')^2 + \frac{1}{2}\mu a_2 q^2 (\psi')^2 + h(q).$$

Combining these two equations allows us to rewrite

$$E = \frac{1}{2}\mu a_1 (q')^2 + V(q), \quad \text{where}$$
$$V(q) = \frac{1}{2}\mu a_2 \frac{M^2}{q^2} + h(q)$$

is the effective potential energy. The evolution of q is described by the autonomous differential equation

$$\mu a_1 q'' = \mu a_2 \frac{M^2}{q^3} - \frac{dh}{dq}(q). \tag{4.14}$$

We point out that it is possible to choose E, M, and h (nonuniquely) so that $V(q)$ has a minimum at $q_0 = 3/2\mu$, and such that the solution, $q(x)$ remains close to q_0 for all $x \in I_j$; in such a case, there are positive constants c_1, c_2, with the property that $c_1\mu \le q \le c_2\mu$ and $c_1 > 1, c_2 < 2$. The function q is an upper bound for u provided the inequality

$$\mu a_2 \frac{M^2}{q^3} - h'(q) \le \frac{1}{\mu a_2 q^3}\left(\mu^2 A - C_1\right)^2 - \frac{C_2}{\tau} + \frac{1}{\mathcal{J}}\frac{df}{ds}(q) \tag{4.15}$$

holds. Here the constant C_1 denotes an upper bound of $|p(x)|$ and C_2 an upper bound of $|\beta_1(q)x|$. Since $q > 0$, the right-hand side of (4.15) is dominated by that of (4.11). We point out that C_1 and C_2 are independent of s, ϕ, and μ. We also note that, since $q \in (c_1\mu, c_2\mu)$, the second and third terms in the right-hand side of (4.15) are bounded independent of μ. So, in order to satisfy inequality (4.15), it is sufficient to choose $A(\mu)$ large enough so that

$$(\mu^2 A - C_1)^2 - (\mu a_2)^2 M^2 \ge K, \tag{4.16}$$

where $K > 0$ is independent of μ. Then, for $\mu \in (0, 1]$, the first term in the right-hand side is dominant. We observe that the condition (4.16) is independent of the choice of the interval I_j, so it will remain the same when we consider different intervals. This completes the construction of the upper bound.

The construction of the lower bound is similar. We only need to choose q to be negative with the absolute value on the order of μ. Since the power of q in the dominant term is odd, we obtain a lower bound $q \ge -2\mu$. Now by Nagumo's theorem, for each choice of ϕ, there exists u such that the pair u, ϕ solves the (4.11) and such that $|u| \le 2\mu$ holds on each I_j, which means that we have u defined on E^+.

Next, a slight modification of the preceding argument provides existence of the solution u on E^-. Since A has already been determined, we can now determine B using (4.5) and (4.9). Specifically, set

$$B = \frac{1}{\mu^3 a_2}(\mu^2 A + C_1). \tag{4.17}$$

So the boundary conditions A and B of ϕ' are now specified. Thus, solving the system (4.7), (4.8) subject to boundary conditions (4.5) and (4.6) yields a solution (S, Φ).

We observe from relations (4.16) and (4.17) that the quantities A and B are independent of the particular choice of s used in the definition of the level sets $E^{\pm}$.

We now consider the sets $E^{\pm}$ corresponding to the fields (S, Φ). Arguing as above, we construct the solution u, on the new sets $E^{\pm}$, and denote the restrictions of S to

E^+ (E^-) by S^+ (S^-). Next, we show that u and S^+ are equal on I_j. Suppose, otherwise, that $S^+ \geq u$ on some interval $L \subset I_j$ and $S^+ = u$ at the endpoints of L. Consider the function $v = S^+ - u$. Substituting u, S^+ into (4.7) and subtracting the resulting equations we find that

$$\mu a_1 v'' = \mu a_2 v(\Phi')^2 - \frac{1}{2}\left(\frac{\beta_1(S^+)}{g(S^+,\Phi)} - \frac{\beta_1(u)}{g(u,\Phi)}\right) x \sin 2\Phi + \frac{1}{\mathcal{J}}\left(\frac{df}{ds}(S^+) - \frac{df}{ds}(u)\right). \tag{4.18}$$

Since $\frac{\beta_1}{g(s,\Phi)}$ is Lipschitz in s, we note that $|\left(\frac{\beta_1(S^+)}{g(S^+,\Phi)} - \frac{\beta_1(u)}{g(u,\Phi)}\right) x \sin 2\Phi| \leq C_g|v|$, holds, where C_g is independent of the choice of I_j. Furthermore, for μ sufficiently close to zero, because of condition (iii), $\frac{df}{ds}(u)$ belongs to the interval on which $\frac{df}{ds}$ and $\frac{d^2f}{ds^2}$ are bounded. Now if S^+ is also in that interval, then using condition (ii), we can write $|\frac{df}{ds}(S^+) - \frac{df}{ds}(u)| \leq C_f v$, where C_f is independent of I_j. Otherwise, if S^+ is large, $\frac{df}{ds}(S^+) > |\frac{df}{ds}(u)|$, by condition (iii). Hence, the right-hand side of (4.18) is positive, for all nonnegative v. Then (4.18) implies that $v'' \geq 0$ on L. Hence, v is a convex function on L which is nonnegative and satisfies zero boundary conditions. Thus, v must be identically zero on L.

Similarly, we prove that if $u \geq S^+$ on some interval, then in fact $u = S^+$. From this and the continuity of S^+, we conclude that $u = S^+$ on I_j and the original solution S satisfies the inequalities

$$\mu \leq S \leq 2\mu \tag{4.19}$$

on I_j. Since the analogous arguments apply to all intervals in E^+, we obtain that S satisfies the inequality (4.19) in E^+. Moreover, $|S| \leq \mu$ in E_0. It remains to prove that $-2\mu \leq S \leq -\mu$ on E^-. The proof is analogous to that in the case of E^+.

Finally, for each $\mu \in (0, 1]$, define $s^\mu = S$, $\phi^\mu = \Phi$, since $|s^\mu| \leq 2\mu$ on I and the sequence s^μ converges to zero uniformly as $\mu \to 0$. □

Remark 4.2. Equations (4.16) and (4.17) suggest that $A = O(\mu^{-2}), B = O(\mu^{-3})$ as $\mu \to 0$.

Remark 4.3. The second part of the proof provides a partial uniqueness argument for solutions of the system (2.1)–(2.4), with the boundary conditions satisfying (4.16), (4.17). It seems that such an argument can be used to prove uniqueness for each fixed $\mu > 0$. This will be the subject of future research.

The proof of the previous theorem brings out some relevant physical aspects of the problem as well as related mathematical issues. The former arguments strongly reflect the interplay between the mechanisms responsible for the oscillatory behavior of solutions and the degenerate nature of the boundary conditions for s close to 0. (Let us recall that ϕ is undefined when $s = 0$.) In order to illustrate such features, let us consider the boundary value problem with prescribed nonzero boundary conditions on s, and also prescribed values of ϕ'. If we now allow the boundary values of s to approach 0, it is natural to expect that the boundary values of ϕ' cannot be independently chosen. This is indeed the nature of Theorem 4.3. Specifically, the restriction on $\phi'(\pm 1)$ imposed by $s(\pm 1) = \mu$ with μ small is contained in inequality (4.16).

A consequence of the fact that $\phi'(\pm 1)$ is large, as indicated by (4.16), is that ϕ' is positive on I, so ϕ is increasing on the intervals of continuity. Consequently, ϕ may be large in such intervals. In order to ensure boundedness of ϕ, we will make

use of the shift invariance of the system (2.1)–(2.2). Indeed, if s, ϕ is a solution, then $s, \phi + k\pi$ is also a solution for any integer k. Starting with an increasing $\tilde{\phi}$, we can split the interval I into subintervals on which $k\pi \le \tilde{\phi} \le (k+1)\pi$ and, then define ϕ by shifting appropriately on each subinterval. The function ϕ obtained in such a fashion will be bounded, rapidly oscillating and discontinuous. We will refer to solutions with rapidly oscillating discontinuous ϕ as generalized to distinguish them from the weak solutions. The discontinuities of ϕ are associated with liquid crystal defects.

COROLLARY 4.4. *Let $\tilde{s}^\mu, \tilde{\phi}^\mu$ be a sequence of weak solutions constructed in Theorem* 4.3. *Then there exists a sequence of generalized solutions s^μ, ϕ^μ such that*

(i) $s^\mu \to 0$ *uniformly on* I;
(ii) $|\phi^\mu| \le \frac{\pi}{2}$;
(iii) *for each $\mu > 0$, the support of the distribution $(\tilde{\phi}^\mu)' - (\phi^\mu)'$ is a finite set of points;*
(iv) *for any $h \in C_0^1(I)$, s^μ, ϕ^μ satisfies the integral identities*

$$\mu a_1 \int_I (s^\mu)' h'\, dx - \mu a_2 \int_I s^\mu ((\phi^\mu)')^2 h\, dx = \int_I G_1(s^\mu, \phi^\mu) h\, dx, \tag{4.20}$$

$$\mu a_2 \int_I (s^\mu)^2 (\tilde{\phi}^\mu)' h'\, dx = \int_I G_2(s^\mu, \phi^\mu) h\, dx.$$

Proof. Let J be the largest integer such that $\phi(-1) \ge J\pi - \frac{\pi}{2}$. The interval I can be represented as a union of N_μ disjoint intervals I_k, $k = 0, 1, 2, \dots N_\mu$, such that $x \in I_k$ when

$$\pi(J+k) - \frac{\pi}{2} \le \phi(x) < \pi(J+k) + \frac{\pi}{2},$$

holds. Starting with $\tilde{s}^\mu, \tilde{\phi}^\mu$, we define s^μ, ϕ^μ as follows:

$$s^\mu = \tilde{s}^\mu,$$
$$\phi^\mu = \tilde{\phi}^\mu - \pi(J+k)$$

for $x \in I_k$. Note that the distributional derivative of ϕ^μ is not locally integrable. However, we have

$$(\phi^\mu)' = (\tilde{\phi}^\mu)'$$

when both are restricted to the complement of the set of the endpoints of I_k.

Next, we observe that $G_l(s, \phi) = G_l(s, \phi + m\pi)$, where $l = 1, 2$ and m is an integer. Hence, replacing $\tilde{\phi}^\mu$ by ϕ^μ in the right-hand sides of the integral identities for the weak solutions which yields (4.20). ∎

Remark 4.4. Note that, in general, it is not possible to replace $(\tilde{\phi}^\mu)'$ by $(\phi^\mu)'$ in the left-hand side of the second identity.

5. Effective configurations. The limiting process yields effective governing equations and configurations of Newtonian Poiseuille flow, with constant viscosity. However, additional equations associated to microstructural phenomena also arise at the limit. They may be related to the occurrence of remnant ordered states on a microscopic scale.

Let us first recall (2.11)–(2.13) and the definition of g:

$$g(s, \phi) = \frac{1}{2}\alpha_4(s) + \alpha_1(s) \sin^2\phi \cos^2\phi + \frac{1}{2}(\alpha_5 - \alpha_2)(s) \sin^2\phi + \frac{1}{2}\alpha_3(s) \cos^2\phi. \tag{5.1}$$

If $\{s^\mu, \phi^\mu\}$ is a sequence of functions such that $s^\mu \to 0$ uniformly, then by (2.16)

$$g(s^\mu, \phi^\mu) \to \frac{1}{2}\alpha_4(0),$$

uniformly on I, as μ approaches zero. Up to a subsequence, the solution v_μ of (2.13) tends to a limit v_0 strongly in $L^2(I)$, and weakly in $W^{1,2}(I)$. Hence, the product $g(s^\mu, \phi^\mu)v'_\mu$ converges to $\frac{1}{2}\alpha_4(0)v'_0$ weakly in $L^2(I)$. This implies that v_0 satisfies the effective equation

$$\eta_{\text{eff}} v''_0 = 1, \tag{5.2}$$

which is the Newtonian Poiseuille flow with constant effective viscosity $\eta_{\text{eff}} = \frac{1}{2}\alpha_4(0)$. When the Ericksen number is large and the Reynolds number is of order 1, the typical viscosity is much larger than the typical elasticity. It is natural to expect that alignment of the molecules will be destroyed by the diffusion, so that liquid crystal flow is that of an isotropic liquid with a constant viscosity. If that were the case, (5.2) would be the only effective equation of the limiting flow. Rigorous analysis suggests, however, that one should also consider the Young measure ν_x generated by the sequence ϕ^μ. In Remark 5.1, following the proof of Theorem 5.1, we explain that ν_x is nontrivial, which means that it cannot have the form $\delta(z-\bar{\phi}(x))$ for any function $\bar{\phi}(x)$. Since localization of ν_x at $\bar{\phi}$ signals strong convergence, this implies that the sequence ϕ^μ cannot converge strongly. Therefore, ν_x must describe "possible disordered states" compatible with the boundary conditions and macroscopic flow.

Combining Theorem 3.3 and Proposition 3.4 with the partial stability Theorem 4.3, we obtain the following theorem.

THEOREM 5.1. *Let $\tilde{s}^\mu, \tilde{\phi}^\mu$ be a sequence of weak solutions from Theorem* 4.3 *satisfying the a priori estimates*

$$\| \tilde{s}' \|_{L^2(I)} \le C,$$
$$\| \tilde{s}\tilde{\phi}' \|_{L^2(I)} \le C,$$

with C independent of μ. Let (s^μ, ϕ^μ) be a corresponding sequence of generalized solutions. Then, up to a subsequence,

(i) *$s^\mu \to 0$ uniformly on I;*

(ii) *the sequence ϕ^μ generates a Young measure ν_x satisfying moment relations*

$$\int_{-1}^{1}\int_{-\pi/2}^{\pi/2}\left(G_1(0,z,x)\sin az - \frac{1}{a}\frac{G_2}{z}(0,z,x)\cos az\right)d\nu_x(z)h(x)\,dx = 0, \tag{5.3}$$
$$\int_{-1}^{1}\int_{-\pi/2}^{\pi/2}\left(G_1(0,z,x)\cos az + \frac{1}{a}\frac{G_2}{z}(0,z,x)\sin az\right)d\nu_x(z)h(x)\,dx = 0,$$

for each $h \in C_0^1(I)$. In (5.3), *$a = (a_2/a_1)^{1/2}$;*

(iii) *the sequence $s((\tilde{\phi}^\mu)')^2$ converges to a measure ρ in the sense of distributions. Moreover,*

$$\rho = \int_{-\pi/2}^{\pi/2} G_1(0,z,x)d\nu_x(z). \tag{5.4}$$

Proof. Part (i) follows directly from Theorem 4.3. Since s^μ converges to zero in L^2, the corresponding Young measure is $\delta(y)$. Hence, part (ii) follows from Theorem 3.3 and momentum relations in Proposition 3.4.

Next, consider the integral identity

$$-\mu a_1 \int_I s' h' dx + \mu a_2 \int_I s(\phi')^2 h \, dx = \int_I G_1(x, s, \phi) h \, dx.$$

Since $\|s'\|_{L^2(I)} \leq C\mu^{-1/2}$ with C independent of μ, the first integral on the left converges to zero as $\mu \to 0$. The integral in the right-hand side converges to

$$\int_I \int_{-\pi/2}^{\pi/2} G_1(0, z, x) d\nu_x(z) h(x) dx.$$

This yields,

$$\lim_{\mu \to 0} \int_I s(\phi')^2 h \, dx = \int_I \int_{-\pi/2}^{\pi/2} G_1(0, z, x) d\nu_x(z) h(x) \, dx$$

for all $h \in C_0^1(I)$. Since this space is dense in $C_0(I)$, the equality above holds for all $h \in C_0(I)$. Hence, the distributional limit of $s(\phi')^2$ is a Radon measure ρ such that

$$\int_I h d\rho = \int_I h \int_{-\pi/2}^{\pi/2} G_1(0, z, x) d\nu_x(z) dx$$

for all $h \in C_0(I)$. ∎

Remark 5.1. Strong convergence of ϕ^μ is incompatible with the moment relations (5.3). If strong convergence takes place, then $\nu_x(z) = \delta(x - \bar{\phi}(x))$ for some function $\bar{\phi}$ (the strong limit). If that were the case, then from (5.3) we would obtain

$$G_1(0, \bar{\phi}, x) \sin a\bar{\phi} - \frac{1}{a} \left(\frac{G_2}{s} \right) (0, \bar{\phi}, x) \cos a\bar{\phi} = 0, \tag{5.5}$$

$$G_1(0, \bar{\phi}, x) \cos a\bar{\phi} + \frac{1}{a} \left(\frac{G_2}{s} \right) (0, \bar{\phi}, x) \sin a\bar{\phi} = 0, \tag{5.6}$$

and thus

$$G_1(0, \bar{\phi}, x) = 0, \tag{5.7}$$

$$\left(\frac{G_2}{s} \right) (0, \bar{\phi}, x) = 0. \tag{5.8}$$

This means that $s = 0, \phi = \bar{\phi}$ is an equilibrium solution of (2.1, 2.2). Direct computation shows that $s = 0$ cannot be an equilibrium solution for any $\bar{\phi}$.

6. Conclusions. We study the oscillatory behavior of the solutions of the governing equations modeling Poiseuille flow of liquid crystals with variable degree of orientation, at the limit of large Ericksen number $\mathcal{E}$. The governing equations are singularly perturbed and highly degenerate. The oscillations of s occur about the isotropic value $s = 0$, and at points where s vanishes the angle of alignment ϕ is discontinuous and ϕ' becomes unbounded. This situation corresponds to the presence of defects in the flow. We obtain necessary condition for the Young measures generated

by sequences of solutions and show that the persistent oscillatory behavior is encoded in the Young measure generated by ϕ. We prove a partial stability result establishing uniform convergence of s to 0 as $\mu = \mathcal{E}^{-1} \to 0$ and the increasingly oscillatory behavior of ϕ. Compactness of s allows us to pass to the limit in the governing system and obtain the effective equations. The latter consist of the Newtonian flow equation together with the algebraic relations for the Young measure generated by ϕ. This suggest that macroscopically the flow is isotropic and Newtonian with a remaining liquid crystalline microstructure.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, New York, 1989.

[2] M. C. Calderer and C. Liu, *Liquid crystal flow: Dynamic and static configurations*, SIAM J. Appl. Math., 60 (2000), pp. 1925–1949.

[3] M. C. Calderer and B. Mukherjee, *Chevron patterns in liquid crystal flows*, Phys. D, 98 (1996), pp. 201–224.

[4] M. C. Calderer and B. Mukherjee, *On Poiseuille flow of liquid crystals*, Liquid Crystals, 22 (1997), pp. 121–136.

[5] M. C. Calderer and B. Mukherjee, *Mathematical issues in the modeling of flow behavior of polymeric liquid crystals*, J. Rheol., 42 (1998), pp. 1519–1536.

[6] S. Chandrasekhar, *Liquid Crystals*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.

[7] K. W. Chang and F. A. Howes, *Nonlinear Singular Perturbation Phenomena: Theory and Application*, Springer-Verlag, New York, 1984.

[8] J. L. Ericksen, *Liquid crystals with variable degree of orientation*, Arch. Ration. Mech. Anal., 113 (1990), pp. 97–120.

[9] I. Fonseca and S. Müller, *A-quasiconvexity, lower semicontinuity, and Young measures*, SIAM J. Math. Anal., 30 (1999), pp. 1355–1390.

[10] D. Kinderlehrer and P. Pedregal, *Characterization of Young measures generated by gradients*, Arch. Ration. Mech. Anal., 115 (1991), pp. 329–365.

[11] R. G. Larson and D. W. Mead, *Development of orientation and texture during shearing of liquid-crystalline polymers*, Liquid Crystals, 12 (1992), pp. 751–768.

[12] F. Leslie, *Some constitutive equations for liquid crystals*, Arch. Ration. Mech. Anal., 28 (1968), pp. 265–283.

[13] G. Marrucci and P. L. Maffettone, *Description of the liquid crystalline phase of rodlike polymer at high shear rates*, Amer. Chem. Soc., 22 (1989), pp. 4076–4082.

[14] B. Mukherjee, S. Mazumder, and M. C. Calderer, *Poiseuille flow of liquid crystals: Highly oscillatory regimes*, J. non-Newtonian Fluid Mech., 99 (2001), pp. 37–55.

[15] M. Nagumo, *Über die Differentialgleichung y"=f(x, y, y')*, Proc. Phys. Math. Soc. Japan, 19 (1937), pp. 861–866.

[16] S. Onogi and T. Asada, *Rheology and rheoptics of polymer liquid crystals*, in Rheology 1. Principles, Plenum, 1980, pp. 127–147.

[17] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser-Verlag, Boston, 1997.

[18] L. Walker and N. Wagner, *Rheology of region one flow in a lyotropic liquid-crystal polymer: The effects of defect texture*, J. Rheology, 38 (1994), pp. 1525–1547.

[19] K. F. Wissbrun, *A model for domain flow of liquid-crystal polymer*, Faraday Discuss. Chem. Soc., 79 (1985), pp. 161–173.

[20] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, Comptes Rendus de la Sociéte des Sciences et des Lettres de Varsovie, 30 (1937), pp. 212–234.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1660–1693

STABILITY OF COMPRESSIBLE VORTEX SHEETS IN STEADY SUPERSONIC EULER FLOWS OVER LIPSCHITZ WALLS*

GUI-QIANG CHEN†, YONGQIAN ZHANG‡, AND DIANWEN ZHU§

Abstract. We are concerned with the stability of compressible vortex sheets in two-dimensional steady supersonic Euler flows over Lipschitz walls under a BV boundary perturbation, since steady supersonic Euler flows are important in many physical situations. It is proved that steady compressible vortex sheets in supersonic flow are stable in structure globally, even under the BV perturbation of the Lipschitz walls. In order to achieve this, we develop a modified Glimm difference scheme and identify a Glimm-type functional to obtain the required BV estimates by incorporating the Lipschitz boundary and the strong vortex sheets naturally and by tracing the interaction not only between the boundary and weak waves but also between the strong vortex sheets and weak waves. Then these estimates are employed to establish the convergence of the approximate solutions to a global entropy solution and the corresponding approximate strong vortex sheets to a strong compressible vortex sheet of the entropy solution. The asymptotic stability of entropy solutions in the flow direction is also established.

Key words. compressible vortex sheets, stability, BV perturbation, supersonic Euler flows, Riemann solutions, Glimm scheme, nonlinear interaction, global existence, adiabatic Euler equations, isentropic Euler equations

AMS subject classifications. 35B40, 35A05, 35B35, 76Y05, 35L65, 85A05

DOI. 10.1137/050642976

1. Introduction.

We are concerned with the stability of compressible vortex sheets in steady supersonic Euler flows over Lipschitz walls under a BV boundary perturbation. The two-dimensional steady supersonic Euler flows are generally governed by

$$\begin{cases} (\rho u)_x + (\rho v)_y = 0, \\ (\rho u^2 + p)_x + (\rho uv)_y = 0, \\ (\rho uv)_x + (\rho v^2 + p)_y = 0, \\ (\rho u(E + p/\rho))_x + (\rho v(E + p/\rho))_y = 0, \end{cases} \tag{1.1}$$

where (u, v) is the velocity, ρ the density, p the scalar pressure, and

$$E = \frac{1}{2}(u^2 + v^2) + e(\rho, p)$$

*Received by the editors October 18, 2005; accepted for publication (in revised form) August 22, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/sima/38-5/64297.html

†Institute of Mathematics and Key Laboratory of Mathematics for Nonlinear Science, Fudan University, Shanghai 200433, People's Republic of China; Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208 (gqchen@math.northwestern.edu). The research of this author was supported in part by the National Science Foundation under grants DMS-0505473, DMS-0244473, DMS-0204225, and INT-9987378 and by an Alexander von Humboldt Foundation Fellowship.

‡Institute of Mathematics and Key Laboratory of Mathematics for Nonlinear Science, Fudan University, Shanghai 200433, People's Republic of China (yongqianz@fudan.edu.cn). The research of this author was supported in part by NSFC projects 10531020, 10271108, 10111120248, and 10311120437, and by the National Science Foundation under grant INT-9987378.

§Center for Scientific Computation and Mathematical Modeling (CSCAMM), 4125 CSIC BLDG 406, University of Maryland, College Park, MD 20742-3289 (zhu@cscamm.umd.edu). The research of this author was supported in part by the National Science Foundation under grants DMS-0244473, DMS-0204225, and INT-9987378.

the total energy with internal energy e, a given function of (ρ, p) defined through thermodynamical relations. The other two thermodynamic variables are the temperature T and the entropy S. If (ρ, S) are chosen as the independent variables, then we have the constitutive relations

$$(e, p, T) = (e(\rho, S), p(\rho, S), T(\rho, S)), \tag{1.2}$$

governed by

$$TdS = de - \frac{p}{\rho^2} d\rho. \tag{1.3}$$

For an ideal gas,

$$p = R\rho T, \qquad e = c_v T, \qquad \gamma = 1 + \frac{R}{c_v} > 1, \tag{1.4}$$

and

$$p = p(\rho, S) = \kappa \rho^\gamma \exp(S/c_v), \qquad e = \frac{\kappa}{\gamma - 1} \rho^{\gamma - 1} \exp(S/c_v) = \frac{RT}{\gamma - 1}, \tag{1.5}$$

where R, κ, and c_v are all positive constants.

If the flow is isentropic (i.e., $S = \text{const.}$), then the pressure p is a function of density ρ, $p = p(\rho)$, and the flow is governed by the following simpler isentropic Euler equations:

$$\begin{cases} (\rho u)_x + (\rho v)_y = 0, \\ (\rho u^2 + p)_x + (\rho uv)_y = 0, \\ (\rho uv)_x + (\rho v^2 + p)_y = 0. \end{cases} \tag{1.6}$$

For polytropic isentropic gases, by scaling, the pressure-density relation can be expressed as

$$p(\rho) = \rho^\gamma / \gamma, \qquad \gamma > 1. \tag{1.7}$$

For the isothermal flow, $\gamma = 1$. The quantity

$$c = \sqrt{p_\rho(\rho, S)}$$

is defined as the sonic speed and, for polytropic gases, $c = \sqrt{\gamma p/\rho}$.

System (1.1) or (1.6) governing a supersonic flow (i.e., $u^2 + v^2 > c^2$) has all real eigenvalues and is hyperbolic in the flow direction, while system (1.1) or (1.6) governing a subsonic flow (i.e., $u^2 + v^2 < c^2$) has complex eigenvalues and is elliptic-hyperbolic mixed and composite (cf. [4, 5]).

We are interested in whether steady compressible vortex sheets in supersonic flow are always stable in supersonic flow under the BV perturbation of the Lipschitz walls. Multidimensional steady supersonic Euler flows are important in many physical situations (cf. Courant and Friedrichs [8]). In particular, when the upstream flow is a uniform steady flow above the plane wall in $x < 0$ all the time, the flow downstream above a Lipschitz wall in $x > 0$ is governed by a steady Euler flow after a sufficiently long time. Furthermore, since steady Euler flows are asymptotic states and may be global attractors of the corresponding unsteady Euler flows, it is important to establish

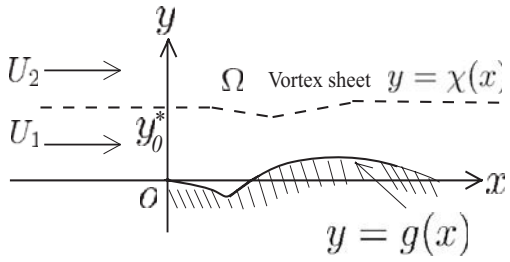


FIG. 1. *Stability of the compressible vortex sheet in supersonic flow.*

the existence of steady Euler flows and understand their qualitative behavior to shed light on the long-time asymptotic behavior of the unsteady compressible Euler flows, one of the most fundamental problems in mathematical fluid dynamics which is still widely open. In this paper we focus on the existence and behavior of such global supersonic Euler flows, especially the nonlinear stability of strong compressible vortex sheets in the supersonic Euler flows, under the BV perturbation of the Lipschitz walls.

For concreteness, we will analyze the problem in the region over the Lipschitz wall for the supersonic Euler flows governed by system (1.1) for $U = (u, v, p, \rho)$ and by (1.6) for $U = (u, v, \rho)$, respectively. Then we have the following:

(i) There exists a Lipschitz function $g \in Lip(\mathbb{R}_+; \mathbb{R})$ with $g(0) = 0$, $g'(0+) = 0$, $g'_\infty = \lim_{x\to\infty} g'(x+)$, and $g' \in BV(\mathbb{R}_+; \mathbb{R})$ such that

$$TV(g'(\cdot)) < \varepsilon \qquad \text{for some constant } \varepsilon > 0, \tag{1.8}$$

$$\Omega = \{(x, y) : y > g(x),\, x \geq 0\}, \quad \Gamma = \{(x, y) : y = g(x),\, x \geq 0\},$$

and $\mathbf{n}(x\pm) = \frac{(-g'(x\pm),1)}{\sqrt{(g'(x\pm))^2+1}}$ are the outer normal vectors to Γ at the points $x\pm$, respectively (see Figure 1).

(ii) The upstream flow consists of one straight vortex sheet $y = y_0^* > 0$ and two constant vectors U_1 when $0 < y < y_0^*$ and U_2 when $y > y_0^* > 0$ satisfying

$$v_1 = v_2 = 0, \qquad u_i > c_i > 0,\ i = 1, 2,$$

where $c_i = \sqrt{\gamma p_i/\rho_i}$ is the sonic speed of state U_i, $i = 1, 2$.

With this setup, the vortex sheet problem can be formulated into the following problem of initial-boundary value type for system (1.1).

Cauchy condition:

$$U|_{x=0} = \begin{cases} U_1, & 0 < y < y_0^*, \\ U_2, & y > y_0^*. \end{cases} \tag{1.9}$$

Boundary condition:

$$(u, v) \cdot \mathbf{n} = 0 \qquad \text{on } \Gamma. \tag{1.10}$$

The main theorem of this paper is the following.

MAIN THEOREM (existence and stability). *There exist $\varepsilon_0 > 0$ and $C > 0$ such that if* (1.8) *holds for $\varepsilon \leq \varepsilon_0$, then there exists a pair of functions*

$$U \in BV_{loc}(\Omega) \cap L^\infty(\Omega), \qquad \chi \in Lip(\mathbb{R}_+; \mathbb{R}_+)$$

with $\chi(0) = y_0^*$ *such that the following hold:*

(i) U *is a global entropy solution of system* (1.1)*, or* (1.6)*, in* Ω *with the initial-boundary data* (1.9)–(1.10) *that is satisfied in the trace sense and*

$$TV\{U(x,\cdot) : [g(x),\infty)\} \le C\,TV(g'(\cdot)) \qquad \text{for every } x \in [0,\infty). \tag{1.11}$$

(ii) *The curve* $\{y = \chi(x)\}$ *is a strong vortex sheet with* $\chi(x) > g(x)$ *for any* $x > 0$*,*

$$\sup_{g(x)<y<\chi(x)} |U(x,y) - U_1| \le C\varepsilon, \qquad \sup_{y>\chi(x)} |U(x,y) - U_2| \le C\varepsilon, \tag{1.12}$$

and

$$\lim_{x\to\infty} \sup\{|v(x,y)/u(x,y) - g'_\infty| : y > g(x)\} + \lim_{x\to\infty} |\chi'(x) - g'_\infty| = 0. \tag{1.13}$$

(iii) *There exists a constant* $p_\infty > 0$ *such that*

$$\lim_{x\to\infty} \sup\{|p(x,y) - p_\infty| : y > g(x)\} = 0. \tag{1.14}$$

This theorem indicates that the strong vortex sheets in supersonic flow are nonlinearly stable globally in structure.

In order to establish this theorem, we first develop a modified Glimm scheme whose mesh grids are designed to follow the slope of the Lipschitz boundary, which is different from the standard case with rectangular mesh grids, so that the lateral Riemann building blocks contain only one wave emanating from the mesh points on the boundary. For the adiabatic Euler flow determined by (1.1), the essential estimate is on the strength δ_1 of the reflected 1-wave when the weak 4-wave with strength α_4 hits the strong vortex sheet from below. We obtain that the key bound on the reflection coefficient in front of α_4 in the estimate of δ_1 is strictly less than one, i.e.,

$$\delta_1 = \beta_1 + K_{11}\alpha_4 + O(1)|\beta_1|(|\alpha_2| + |\alpha_3|), \qquad |K_{11}| < 1.$$

The estimate on the interaction between the boundary and weak waves is also crucial.

Based on the construction of the modified Glimm scheme and the new interaction estimates, we successfully identify a Glimm-type functional by incorporating the Lipschitz wall and the strong vortex sheet naturally and by tracing the interactions not only between the boundary and weak waves but also between the strong vortex sheet and weak waves, so that the Glimm-type functional monotonically decreases in the flow direction. Another essential estimate is to trace the approximate strong vortex sheets in order to establish the nonlinear stability and asymptotic behavior of the strong vortex sheet in supersonic flow in the downstream direction under the BV boundary perturbation. Condition (1.8) may be relaxed so that the total variation of $g'(x)$ is allowed to be relatively larger than the small L^∞ norm of $g'(x)$ by combining the analysis in this paper with the arguments in [21, 22].

We observe that the stability of contact discontinuities for the Cauchy problem for strictly hyperbolic systems under a BV perturbation has been studied by Sablé-Tougeron [18] and Corli and Sablé-Tougeron [9]. In particular, the reflection coefficients, such as K_{11} here, are required to be less than one, which is the stability condition for the mixed problem in the strip $(0,1)$ in the earlier works; see, e.g., Sablé-Tougeron [18]. The essential difference between system (1.1) as analyzed in sections 2–5 here and strictly hyperbolic systems as considered in [9, 18] is that two of the four characteristic eigenvalues coincide and have two linearly independent

eigenvectors which determine the compressible vortex sheets so that two independent parameters are required to describe precisely these vortex sheets. Furthermore, our physical boundary condition for the initial-boundary value problem is the slip-wall boundary condition on the nonflat boundary whose slope function is allowed to be discontinuous, i.e., the characteristic boundary condition on the *nonflat, nonsmooth* boundary, which is different from the initial-boundary value problem over the *flat* boundary considered in [18]. For our case, additional nonlinear waves are produced by the boundary vertices, A_k, so that careful estimates on boundary reflection and interaction for nonlinear waves are required to match the slip-wall boundary condition and to be incorporated into the functional naturally.

We further remark that our Glimm-type functional has additional ingredients, in comparison with that used in Corli and Sablé-Tougeron [9] and Sablé-Tougeron [18]. More precisely, in this paper, the linear part of the functional has a new term, $L_0(J)$, to control the new waves produced by the flow moving around the boundary vertices, A_k. The linear part of the functional is shown to be decreasing near the vortex sheet and the boundary, in addition to controlling the quadratic part. The proof of the decreasing of the linear part makes full use of the following facts:

(i) $L_0(J)$ is strictly decreasing along the boundary.

(ii) If a 1-wave interacts with the boundary or the vortex sheet, it disappears and a new wave is produced, which implies the strict decreasing of L_1^i near the boundary and vortex sheet. The same is true when a 2-wave interacts with the vortex sheet.

Then, combining these two facts with the estimates on the coefficients K_{ij} and K_{bi} in the wave interactions and reflections, we succeed in assigning different weights to the strengths of the waves so that the linear part is strictly decreasing. With the careful design of the linear part, which is different from that in [18, 9], we simply use the same quadratic part as that used in Glimm [11] (see also [14, 15, 19]) in our Glimm-type functional, which sufficiently controls the linear part in the regions (1) and (2). For more details, see section 4.

It would be interesting and important to clarify the connections between the stability of steady compressible vortex sheets and long-time asymptotic stability of unsteady compressible vortex sheets in supersonic flow, since the nonlinear instabilities of compressible vortex sheets may develop in a short-time or some intermediate-time regime under certain situations. For this, we refer the reader to Miles [17], Artola and Majda [1, 2, 3], Woodward [24], and the references cited therein.

In this paper we first focus on the adiabatic Euler flows in sections 2–5 and then study the isentropic Euler flows in section 6. In section 2, we study the lateral and classical Riemann problems and analyze the properties of the Riemann solutions to the adiabatic Euler equations (1.1), which are essential for the interaction estimates among the nonlinear waves and the wall in section 3 and for the existence and behavior of entropy solutions of the problem in sections 4–5. In section 3, we make estimates on the wave interactions and reflections on the boundary and the strong vortex sheet, respectively. In section 4, we develop a modified Glimm scheme to construct the approximate solutions and establish necessary estimates for them in the approximate domains. In section 5, we establish the convergence of approximate solutions to a global entropy solution and prove the nonlinear stability and asymptotic behavior of the strong vortex sheet over the Lipschitz wall. In section 6, we extend the analysis and approach to establish the existence and behavior of two-dimensional steady supersonic flows governed by the isentropic Euler equations.

2. Riemann problems and Riemann solutions. In this section, we study the lateral and classical Riemann problems and analyze the properties of the Riemann solutions to the adiabatic Euler system (1.1), which are essential not only for the interaction estimates among the nonlinear waves and the Lipschitz wall but also for the existence and behavior of solutions of our problem in sections 3–5.

2.1. Euler equations. The Euler system can be written in the following conservation form:

$$W(U)_x + H(U)_y = 0, \tag{2.1}$$

where $U = (u, v, p, \rho)$, $W(U) = (\rho u, \rho u^2 + p, \rho uv, \rho u(h + \frac{u^2+v^2}{2}))$, and $H(U) = (\rho v, \rho uv, \rho v^2 + p, \rho v(h + \frac{u^2+v^2}{2}))$ with $h = \frac{\gamma p}{(\gamma-1)\rho}$. For a smooth solution $U(x, y)$, system (2.1) is equivalent to

$$\nabla_U W(U)U_x + \nabla_U H(U)U_y = 0. \tag{2.2}$$

Then the eigenvalues of (2.1) are the roots of the fourth order polynomial,

$$\det(\lambda \nabla_U W(U) - \nabla_U H(U)), \tag{2.3}$$

that is, the solutions of the equation

$$(v - \lambda u)^2 \left((v - \lambda u)^2 - c^2(1 + \lambda^2)\right) = 0,$$

where $c = \sqrt{\gamma p/\rho}$ is the sonic speed. If the flow is supersonic (i.e., $u^2 + v^2 > c^2$), system (1.1) is hyperbolic. In particular, when $u > c$, system (1.1) has four eigenvalues in the x-direction,

$$\lambda_j = \frac{uv + (-1)^j c\sqrt{u^2 + v^2 - c^2}}{u^2 - c^2}, \ \ j = 1, 4; \qquad \lambda_i = v/u, \ \ i = 2, 3, \tag{2.4}$$

and the four corresponding linearly independent eigenvectors:

$$\begin{aligned} &\mathbf{r}_j = \kappa_j(-\lambda_j, 1, \rho(\lambda_j u - v), \rho(\lambda_j u - v)/c^2)^\top, \quad j = 1, 4, \\ &\mathbf{r}_2 = (u, v, 0, 0)^\top, \qquad \mathbf{r}_3 = (0, 0, 0, \rho)^\top, \end{aligned} \tag{2.5}$$

where κ_j are chosen so that $\mathbf{r}_j \cdot \nabla\lambda_j = 1$ since the jth-characteristic fields are genuinely nonlinear, $j = 1, 4$. Note that the second and third characteristic fields are always linearly degenerate: $\mathbf{r}_j \cdot \nabla\lambda_j = 0, j = 2, 3$. We also point out that, at the unperturbed states $U_i = (u_i, 0, p_i, \rho_i)$ for $i = 1, 2$,

$$\lambda_2(U_i) = \lambda_3(U_i) = 0, \quad \lambda_1(U_i) = -c_i/\sqrt{u_i^2 - c_i^2} = -\lambda_4(U_i) < 0 \ \text{ for } i = 1, 2.$$

Definition 2.1 (entropy solutions). *A BV function $U = U(x, y)$ is called an entropy solution of problem* (1.1) *and* (1.9)–(1.10) *provided that*

(i) *U is a weak solution of* (1.1) *and satisfies* (1.9)–(1.10) *in the trace sense;*

(ii) *U satisfies the entropy inequality*

$$(\rho uS)_x + (\rho vS)_y \geq 0 \tag{2.6}$$

in the sense of distributions in Ω including the boundary.

Remark 2.1. The entropy inequality (2.6) for our steady case directly follows from the Clausius inequality in the time-dependent case,

$$(\rho S)_t + (\rho u S)_x + (\rho v S)_y \geq 0,$$

under the assumption that the solution (u, v, p, ρ) is independent of time t. Furthermore, by a direct but tedious calculation, when $u > c$, the function $\eta = -\rho u S$ as an entropy function of system (1.1) is a strictly convex function of the conserved variables

$$w = (w_1, w_2, w_3, w_4) = (\rho u, \rho u^2 + p, \rho u v, \rho u(E + p/\rho)).$$

2.2. Wave curves in the phase space. In this subsection, we analyze some basic properties of nonlinear waves. We consider the problem for $u > c$ in the state space, especially in the neighborhoods of U_1 and U_2. We first seek the self-similar solutions to (1.1),

$$(u, v, p, \rho)(x, y) = (u, v, p, \rho)(\xi), \qquad \xi = y/x,$$

which connect to a state $U_0 = (u_0, v_0, p_0, \rho_0)$. Then we have

$$\det\left(\xi \nabla_U W(U) - \nabla_U H(U)\right) = 0, \tag{2.7}$$

which implies

$$\xi = \lambda_i = v/u, \quad i = 2, 3, \quad \text{or} \quad \xi = \lambda_j = \frac{uv + (-1)^j c\sqrt{u^2 + v^2 - c^2}}{u^2 - c^2}, \quad j = 1, 4.$$

Plugging $\xi = \lambda_i$ into (2.7), we obtain

$$dp = 0, \qquad v du - u dv = 0,$$

which yields the vortex sheet curves $C_i(U_0)$ in the phase space:

$$C_i(U_0): \quad p = p_0, \quad w = v/u = v_0/u_0, \qquad i = 2, 3, \tag{2.8}$$

which describe compressible vortex sheets. More precisely, we have

$$C_2(U_0): U = (u_0 e^{\sigma_2}, v_0 e^{\sigma_2}, p_0, \rho_0)^\top \tag{2.9}$$

with strength σ_2 and slope v_0/u_0, which is determined by

$$\begin{cases} \frac{dU}{d\sigma_2} = \mathbf{r}_2(U), \\ U|_{\sigma_2 = 0} = U_0, \end{cases}$$

and

$$C_3(U_0): U = (u_0, v_0, p_0, \rho_0 e^{\sigma_3})^\top \tag{2.10}$$

with strength σ_3 and slope v_0/u_0, which is determined by

$$\begin{cases} \frac{dU}{d\sigma_3} = \mathbf{r}_3(U), \\ U|_{\sigma_3 = 0} = U_0. \end{cases}$$

Remark 2.2. The full Euler system (1.1) has two contact discontinuities that coincide as a single vortex sheet in the physical xy-plane. However, in the phase space

$U = (u, v, p, \rho)$, to describe this vortex sheet precisely, it requires two independent parameters, since there are two linearly independent eigenvectors corresponding to the repeated eigenvalues $\lambda_2 = \lambda_3 = v/u$ of the two linearly degenerate fields.

Plugging $\xi = \lambda_j$ into (2.7), we get the jth-rarefaction wave curve $R_j(U_0), j = 1, 4$, in the phase space through U_0:

(2.11)
$$R_j(U_0): \ dp = c^2 d\rho, \ du = -\lambda_j dv, \ \rho(\lambda_j u - v)dv = dp \quad \text{for } \rho < \rho_0, \ u > c, \quad j = 1, 4.$$

Similarly, the Rankine–Hugoniot conditions for (1.1) are

$$s[\rho u] = [\rho v], \tag{2.12}$$

$$s[\rho u^2 + p] = [\rho uv], \tag{2.13}$$

$$s[\rho uv] = [\rho v^2 + p], \tag{2.14}$$

$$s\left[\rho u\left(h + \frac{u^2 + v^2}{2}\right)\right] = \left[\rho v\left(h + \frac{u^2 + v^2}{2}\right)\right], \tag{2.15}$$

where the jump symbol $[\cdot]$ stands for the value of the quantity of the front state minus that of the back state.

Then we have

$$(v_0 - su_0)^2\left((v_0 - su_0)^2 - \overline{c}^2(1 + s^2)\right) = 0,$$

where $\overline{c}^2 = \frac{c_0^2}{b}\frac{\rho}{\rho_0}$ and $b = \frac{\gamma+1}{2} - \frac{\gamma-1}{2}\frac{\rho}{\rho_0}$. This implies

$$s = s_i = v_0/u_0, \quad i = 2, 3, \tag{2.16}$$

or

$$s = s_j = \frac{u_0 v_0 + (-1)^j \overline{c}\sqrt{u_0^2 + v_0^2 - \overline{c}^2}}{u_0^2 - \overline{c}^2}, \quad j = 1, 4, \tag{2.17}$$

where $u_0 > \overline{c}$ for small shocks.

Plugging $s_i, i = 2, 3$, into (2.12)–(2.15), we get the same $C_i(U_0), i = 2, 3$, as defined in (2.9)–(2.10); while plugging $s_j, j = 1, 4$, into (2.12)–(2.15), we obtain the jth-shock wave curve $S_j(U_0), j = 1, 4$, through U_0:

$$S_j(U_0): \ [p] = \frac{c_0^2}{b}[\rho], \ [u] = -s_j[v], \ \rho_0(s_j u_0 - v_0)[v] = [p] \quad \text{for } \rho > \rho_0, \ u > c, \quad j = 1, 4,$$

where $\rho_0 < \rho$ is equivalent to the entropy condition (2.6) on the shock wave.

Note that $S_j(U_0)$ contacts with $R_j(U_0)$ at U_0 up to second order.

2.3. Lateral Riemann problem. It has been shown in [8] that when the angle between the flow direction of the front state and the wall at a boundary vertex is smaller than π and larger than the extreme angle determined by the incoming flow state and $\gamma \geq 1$, then a unique 4-shock forms, which separates the front state from the supersonic back state; when the angle between the flow direction of the front state and the wall at a boundary vertex is larger than π and less than an extreme angle, then a 4-rarefaction wave forms, which emanates from the vertex (see Figure 2). This indicates that when the angle between the flow direction of the front state and the wall at a boundary vertex is near π, the lateral Riemann problem is always uniquely solvable. For the detail, see Figure 3 and Proposition 3.3; also see Courant and Friedrichs [8].

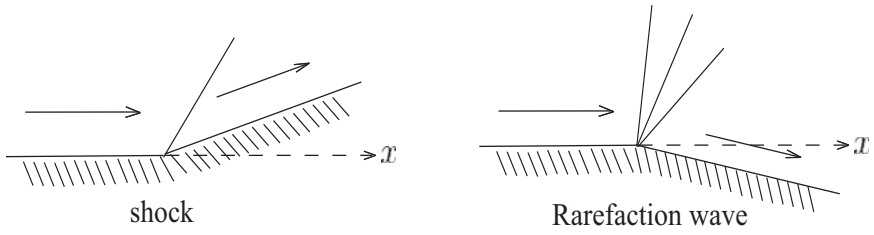


FIG. 2. *Lateral Riemann solutions.*

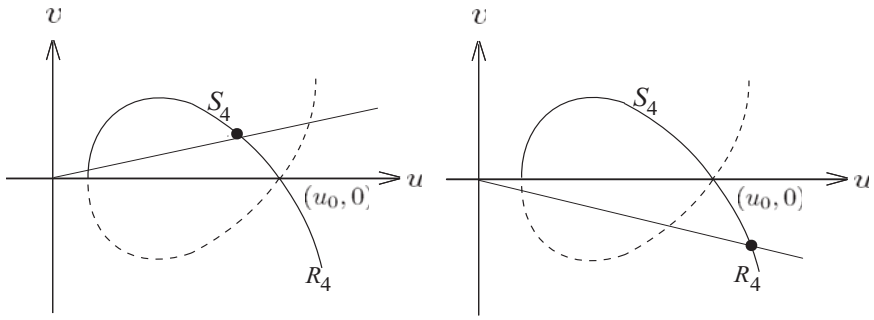


FIG. 3. *Wave curves in the (u, v)-plane for the lateral Riemann problem.*

2.4. Riemann problem involving only weak waves. Consider the Riemann problem for (2.1):

$$U|_{x=x_0} = \begin{cases} U_a, & y > y_0, \\ U_b, & y < y_0, \end{cases} \tag{2.18}$$

where U_a and U_b are the constant states which are regarded as the *above* state and *below* state with respect to the line $y = y_0$, respectively.

Following Lax [13], we can parameterize any physically admissible wave curve in a neighborhood of any constant state U_0, $O_\varepsilon(U_0)$, by $\alpha_j \mapsto \Phi_j(\alpha_j; U_b)$, with $\Phi_j \in C^2$ and $\Phi_j|_{\alpha_j=0} = U_b$ and $\partial_{\alpha_j}\Phi_j|_{\alpha_j=0} = \mathbf{r}_j(U_b)$. For $j = 1, 4$, the case $\alpha_j > 0$ corresponds to a rarefaction wave, while the case $\alpha_j < 0$ corresponds to a shock wave. Notice that, for system (1.1), we have the explicit formulas for Φ_2 and Φ_3 to describe the vortex sheets by two independent parameters (σ_2, σ_3):

$$\Phi_2(\sigma_2, U_b) = (u_b e^{\sigma_2}, v_b e^{\sigma_2}, p_b, \rho_b), \qquad \Phi_3(\sigma_3, U_b) = (u_b, v_b, p_b, \rho_b e^{\sigma_3}).$$

For simplicity, we set $\Phi(\alpha_4, \alpha_3, \alpha_2, \alpha_1; U_b) := \Phi_4(\alpha_4; \Phi_3(\alpha_3; \Phi_2(\alpha_2; \Phi_1(\alpha_1; U_b))))$. We also denote by $O_\varepsilon(W)$ a universal neighborhood that is a ball with radius $M\varepsilon > 0$ and center W, where $M > 0$ is a universal constant depending only on the parameters in the system and possibly the boundary function $g(x)$ (starting section 4.2), which may be different at each occurrence. Then we have the following lemma.

LEMMA 2.2. *There exists $\varepsilon > 0$ such that, for any states $U_a, U_b \in O_\varepsilon(U_0)$, the Riemann problem* (2.18) *admits a unique admissible solution consisting of four elementary waves. In addition, the state U_a can be represented by*

$$U_a = \Phi(\alpha_4, \alpha_3, \alpha_2, \alpha_1; U_b)$$

with $\Phi|_{\alpha_1=\alpha_2=\alpha_3=\alpha_4=0} = U_b$ and $\partial_{\alpha_i}\Phi|_{\alpha_1=\alpha_2=\alpha_3=\alpha_4=0} = \mathbf{r}_i(U_b), i = 1, 2, 3, 4.$

Furthermore, we find that the renormalization factors $\kappa_j(U), j = 1, 4$, in (2.5) are positive in a neighborhood $O_\varepsilon(U_0)$ of any state $U_0 = (u_0, 0, p_0, \rho_0)$ with $u_0 > c_0$, which is either U_1 or U_2 below.

LEMMA 2.3. *At any state $U_0 = (u_0, 0, p_0, \rho_0)$ with $u_0 > c_0$,*

$$\kappa_1(U_0) = \kappa_4(U_0) = 1/\left(\nabla_U \lambda_j \cdot (-\lambda_j, 1, \rho u \lambda_j, \rho u \lambda_j / c^2)|_{U=U_0}\right) > 0,$$

which implies $\kappa_j(U) > 0$ for any $U \in O_\varepsilon(U_0)$ with small $\varepsilon > 0$ since $\kappa_j(U)$ are continuous for $j = 1, 4$.

In fact, at the state $U_0 = (u_0, 0, p_0, \rho_0)$, it is straightforward to see that

$$\nabla_U \lambda_j \cdot (-\lambda_j, 1, \rho u \lambda_j, \rho u \lambda_j / c^2)|_{U=U_0} > 0, \qquad j = 1, 4.$$

Therefore, we have $\kappa_1(U_0) = \kappa_4(U_0) > 0$, which implies that $\kappa_j(U) > 0$, $j = 1, 4$, for any state $U \in O_\varepsilon(U_0)$ with small $\varepsilon > 0$.

2.5. Riemann problem involving the strong vortex sheets. For simplicity, we use the notation $\{U_b, U_a\} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ to denote $U_a = \Phi(\alpha_4, \alpha_3, \alpha_2, \alpha_1; U_b)$ throughout the paper. For any $U_b \in O_\varepsilon(U_1)$ and $U_a \in O_\varepsilon(U_2)$, we also use $\{U_b, U_a\} = (0, \sigma_2, \sigma_3, 0)$ to denote the strong vortex sheet that connects U_b and U_a with strength (σ_2, σ_3). That is,

$$U_m = \Phi_2(\sigma_2; U_b) := (u_b e^{\sigma_2}, v_b e^{\sigma_2}, p_b, \rho_b)^\top, \; U_a = \Phi_3(\sigma_3; U_m) := (u_m, v_m, p_m, \rho_m e^{\sigma_3})^\top.$$

In particular, we note that

$$U_2 = (u_2, 0, p_2, \rho_2)^\top = (u_1 e^{\sigma_{20}}, 0, p_1, \rho_1 e^{\sigma_{30}})^\top.$$

We write $G(\sigma_3, \sigma_2; U_b) = \Phi_3(\sigma_3; \Phi_2(\sigma_2; U_b))$ for short for any $U_b \in O_\epsilon(U_1)$. Then we have the following.

LEMMA 2.4. *The vector function $G(\sigma_3, \sigma_2; U_b)$ satisfies*

$$G_{\sigma_2}(\sigma_3, \sigma_2; U_b) = (u_b e^{\sigma_2}, v_b e^{\sigma_2}, 0, 0)^\top, \qquad G_{\sigma_3}(\sigma_3, \sigma_2; U_b) = (0, 0, 0, \rho_b e^{\sigma_3})^\top,$$

and

$$\nabla_U G(\sigma_3, \sigma_2; U_b) = \operatorname{diag}(e^{\sigma_2}, e^{\sigma_2}, 1, e^{\sigma_3}).$$

These can be easily obtained from direct calculations and are thus omitted.

The following lemma is essential to estimate the strengths of reflected weak waves in the interaction between the strong vortex sheet and weak waves (see the proofs for Propositions 3.4–3.5).

LEMMA 2.5. *For the plane vortex sheet with the lower state $U_1 = (u_1, 0, p_1, \rho_1)$, upper state $U_2 = (u_2, 0, p_1, \rho_2)$, and strength $(\sigma_{20}, \sigma_{30})$,*

$$\det(\mathbf{r}_4(U_2), G_{\sigma_3}(\sigma_{30}, \sigma_{20}; U_1), G_{\sigma_2}(\sigma_{30}, \sigma_{20}; U_1), \nabla_U G(\sigma_{30}, \sigma_{20}; U_1) \cdot \mathbf{r}_1(U_1)) > 0.$$

This can be obtained by the following direct calculations:

$$\begin{aligned}
&\det(\mathbf{r}_4(U_2), G_{\sigma_3}(\sigma_{30}, \sigma_{20}; U_1), G_{\sigma_2}(\sigma_{30}, \sigma_{20}; U_1), \nabla_U G(\sigma_{30}, \sigma_{20}; U_1) \cdot \mathbf{r}_1(U_1)) \\
&= \kappa_1(U_1)\kappa_4(U_2) \begin{vmatrix} -\lambda_4(U_2) & 0 & u_1 e^{\sigma_{20}} & e^{\sigma_{20}}(-\lambda_1(U_1)) \\ 1 & 0 & 0 & e^{\sigma_{20}} \\ \rho_2 u_2 \lambda_4(U_2) & 0 & 0 & \rho_1 u_1 \lambda_1(U_1) \\ \frac{\rho_2 u_2 \lambda_4(U_2)}{c_2^2} & \rho_1 e^{\sigma_{30}} & 0 & e^{\sigma_{30}} \frac{\rho_1 u_1 \lambda_1(U_1)}{c_1^2} \end{vmatrix} \\
&= \kappa_1(U_1)\kappa_4(U_2)\rho_1^2 u_1^2 e^{\sigma_{20}+\sigma_{30}}(\lambda_4(U_2)e^{2\sigma_{20}+\sigma_{30}} + \lambda_4(U_1)) > 0.
\end{aligned}$$

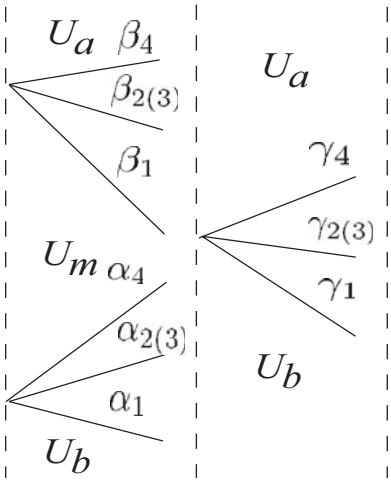


FIG. 4. *Weak wave interactions.*

3. Estimates on wave interactions and reflections. We now make estimates on wave interactions, especially near the strong vortex sheet, and wave reflections on the wall.

3.1. Estimates on weak wave interactions. We first estimate the interactions among weak waves. We will use the following elementary identity, whose proof is straightforward.

LEMMA 3.1. *If $f \in C^2(\mathbb{R}^2)$, then, for any $(x,y) \in \mathbb{R}^2$,*

$$f(x,y) - f(x,0) - f(0,y) + f(0,0) = xy \int_0^1 \int_0^1 f_{xy}(rx, sy) dr ds. \tag{3.1}$$

PROPOSITION 3.2. *Suppose that U_b, U_m, and U_a are three states in a small neighborhood of U_0 with $\{U_b, U_m\} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, $\{U_m, U_a\} = (\beta_1, \beta_2, \beta_3, \beta_4)$, and $\{U_b, U_a\} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ (cf. Figure 4). Then*

$$\gamma_i = \alpha_i + \beta_i + O(1)\triangle(\alpha, \beta), \tag{3.2}$$

where $\triangle(\alpha,\beta) = |\alpha_4||\beta_1| + |\alpha_3||\beta_1| + |\alpha_2||\beta_1| + |\alpha_4||\beta_2| + |\alpha_4||\beta_3| + \sum_{j=1,4} \triangle_j(\alpha,\beta)$ with

$$\triangle_j(\alpha,\beta) = \begin{cases} 0, & \alpha_j \geq 0,\ \beta_j \geq 0; \\ |\alpha_j||\beta_j|, & \text{otherwise.} \end{cases}$$

The proof of this proposition is standard, whose simple proof can be found in Temple [20] or Chen, Zhang, and Zhu [7]; also see Dafermos [10].

3.2. Estimates on the weak wave reflections on the boundary. Denote by $\{C_k\}_{k=0}^{\infty}$ the points $\{(a_k, b_k)\}_{k=0}^{\infty}$ in the xy-plane with $a_{k+1} > a_k > 0$ (cf. Figure 5). Set

$$\begin{aligned} &\omega_{k,k+1} = \arctan\left(\frac{b_{k+1} - b_k}{a_{k+1} - a_k}\right), \quad \omega_k = \omega_{k,k+1} - \omega_{k-1,k}, \quad \omega_{-1,0} = 0, \\ &\Omega_{k+1} = \{(x,y) : x \in [a_k, a_{k+1}],\ y > b_k + (x - a_k)\tan(\omega_{k,k+1})\}, \\ &\Gamma_{k+1} = \{(x,y) : x \in (a_k, a_{k+1}),\ y = b_k + (x - a_k)\tan(\omega_{k,k+1})\}, \end{aligned} \tag{3.3}$$

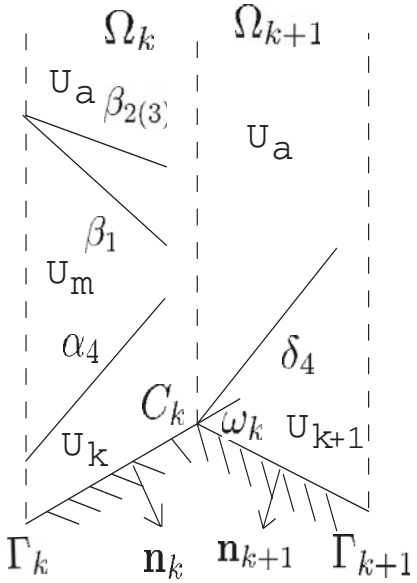


FIG. 5. *Weak wave reflections on the boundary.*

and the outer normal vector to Γ_k:

$$\mathbf{n}_{k+1} = \frac{(b_{k+1}-b_k, a_k - a_{k+1})}{\sqrt{(b_{k+1}-b_k)^2+(a_{k+1}-a_k)^2}} = (\sin(\omega_{k,k+1}), -\cos(\omega_{k,k+1})). \tag{3.4}$$

We consider the initial-boundary value problem

$$\begin{cases} (2.1) & \text{in } \Omega_{k+1}, \\ U|_{x=a_k} = U_a, & \\ (u,v)\cdot \mathbf{n}_{k+1} = 0 & \text{on } \Gamma_{k+1}, \end{cases} \tag{3.5}$$

where U_a is a constant state.

PROPOSITION 3.3. *Let* $\{U_m, U_a\} = (\beta_1, \beta_2, \beta_3, 0)$ *and* $\{U_k, U_m\} = (0,0,0,\alpha_4)$ *with*

$$(u_k, v_k)\cdot \mathbf{n}_k = 0.$$

Then there exists a unique solution U_{k+1} *of problem* (3.5) *such that*

$$\{U_{k+1}, U_a\} = (0,0,0,\delta_4)$$

and $U_{k+1}\cdot(\mathbf{n}_{k+1},0,0) = 0$. *Furthermore,*

$$\delta_4 = \alpha_4 + K_{b1}\beta_1 + K_{b2}\beta_2 + K_{b3}\beta_3 + K_{b0}\omega_k,$$

where K_{b1}, K_{b2}, K_{b3}, *and* K_{b0} *are* C^2*-functions of* β_3, β_2, β_1, α_4, ω_{k+1}, *and* U_a *satisfying*

$$K_{b1}|_{\{\omega_k=\alpha_4=\beta_1=\beta_2=\beta_3=0, U_a=U_1\}} = 1, \qquad K_{bi}|_{\{\omega_k=\alpha_4=\beta_1=\alpha_2=\beta_3=0, U_a=U_1\}} = 0,\ i=2,3,$$

and K_{b0} *is bounded. In particular,* $K_{b0} < 0$ *at the origin.*

Proof. First, by the implicit function theorem, similar to Lemma 2.2, we have $U_k = \Psi(\alpha_1, \alpha_2, \alpha_3, \alpha_4; U_a)$ for the C^2-function $\Psi = \Phi^{-1}$ and $\Psi_{\alpha_j}|_{\alpha_1=\alpha_2=\alpha_3=\alpha_4=0} = -\mathbf{r}_j(U_a)$, $j = 1, 2, 3, 4$. Then it suffices to find the solution to

$$\Psi(0,0,0,\alpha_4; \Psi(\beta_1,\beta_2,\beta_3,0;U_a)) \cdot (\mathbf{n}_k, 0, 0) = \Psi(0,0,0,\delta_4;U_a)\cdot(\mathbf{n}_{k+1},0,0). \tag{3.6}$$

Since

$$\begin{aligned}&\partial_{\delta_4}(\Psi(0,0,0,\delta_4;U_a)\cdot(\mathbf{n}_{k+1},0,0))|_{\{\delta_4=0,U_a=U_1,\omega_{k,k+1}=0\}}\\ &= -\kappa_1(U_1)(-\lambda_1(U_1), 1, \rho_1\lambda_1(U_1)u_1, \rho_1\lambda_1(U_1)u_1/c_1^2)\cdot(0,1,0,0) < 0,\end{aligned}$$

we know from the implicit function theorem that δ_4 can be solved as a C^2-function of β_3, β_2, β_1, α_4, ω_k, $\omega_{k,k+1}$, and U_a. Since $\omega_{k,k+1}$ and U_a are constant here, we write $\delta_4 = \delta_4(\omega_k, \beta_3, \beta_2, \beta_1, \alpha_4)$ without specific indication of the dependence on U_a and $\omega_{k,k+1}$.

Again, from (3.1), we can obtain

$$\begin{aligned}&\delta_4(\omega_k,\beta_3,\beta_2,\beta_1,\alpha_4)\\ &= \delta_4(\omega_k,\beta_3,\beta_2,\beta_1,\alpha_4) - \delta_4(0,\beta_3,\beta_2,\beta_1,\alpha_4) + \delta_4(0,\beta_3,\beta_2,\beta_1,\alpha_4)\\ &\quad - \delta_4(0,0,\beta_2,\beta_1,\alpha_4) + \delta_4(0,0,\beta_2,\beta_1,\alpha_4) - \delta_4(0,0,0,\beta_1,\alpha_4)\\ &\quad + \delta_4(0,0,0,\beta_1,\alpha_4) - \delta_4(0,0,0,0,\alpha_4) + \delta_4(0,0,0,0,\alpha_4)\\ &= K_{b0}\omega_k + K_{b3}\beta_3 + K_{b2}\beta_2 + K_{b1}\beta_1 + \alpha_4.\end{aligned}$$

Differentiating (3.6) with respect to β_1, β_2, and β_3, respectively, and letting $\omega_k = \beta_3 = \beta_2 = \beta_1 = \alpha_4 = 0$ and $U_a = U_1$, we have

$$\mathbf{r}_1(U_1)\cdot(0,1,0,0) = \partial_{\beta_1}\delta_4\,\mathbf{r}_4(U_1)\cdot(0,1,0,0),$$

$$\mathbf{r}_2(U_1)\cdot(0,1,0,0) = \partial_{\beta_2}\delta_4\,\mathbf{r}_4(U_1)\cdot(0,1,0,0),$$

and

$$\mathbf{r}_3(U_1)\cdot(0,1,0,0) = \partial_{\beta_3}\delta_4\,\mathbf{r}_4(U_1)\cdot(0,1,0,0).$$

Hence we have

$$K_{b1}|_{\{\omega_k=\beta_3=\beta_2=\beta_1=\alpha_4=0,\ U_a=U_1\}} = 1, \qquad K_{b2}|_{\{\omega_k=\beta_3=\beta_2=\beta_1=\alpha_4=0,\ U_a=U_1\}} = 0,$$

and $K_{b3}|_{\{\omega_k=\beta_3=\beta_2=\beta_1=\alpha_4=0,\ U_a=U_1\}} = 0$. It is clear that $K_{b0} = \partial_{\omega_k}\delta_4$ is bounded. Differentiating (3.6) with respect to ω_k, we find that, at the origin, $K_{b0} = \partial_{\omega_k}\delta_4| = -u_1/\kappa_4(U_1) < 0$. This completes the proof. ∎

3.3. Estimates on the interaction between the strong vortex sheet and weak waves. There are two cases depending on how the strong vortex sheet and the weak waves interact. The first case is when the weak waves approach the strong vortex sheet from below, as in Figure 6. Then we have the following.

PROPOSITION 3.4. *Let $U_b, U_m \in O_\epsilon(U_1)$ and $U_a \in O_\epsilon(U_2)$ with*

$$\{U_b, U_m\} = (0, \alpha_2, \alpha_3, \alpha_4), \qquad \{U_m, U_a\} = (\beta_1, \sigma_2, \sigma_3, 0).$$

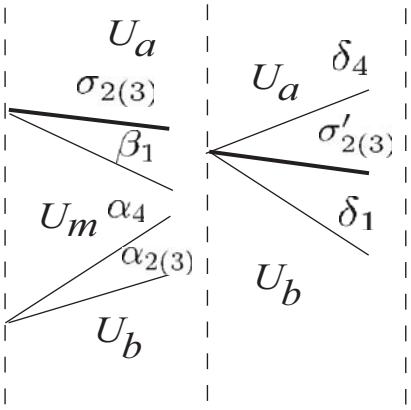


FIG. 6. *Weak waves approach the strong vortex sheet from below.*

Then there exists a unique $(\delta_1, \sigma_2', \sigma_3', \delta_4)$ *such that the Riemann problem* (2.18) *admits an admissible solution that consists of a weak* 1*-wave of strength* δ_1*, a strong vortex sheet of strength* (σ_2', σ_3')*, and a weak* 4*-wave of strength* δ_4*:*

$$\{U_b, U_a\} = (\delta_1, \sigma_2', \sigma_3', \delta_4).$$

Moreover,

$$\begin{aligned}
&\delta_1 = \beta_1 + K_{11}\alpha_4 + O(1)\Delta', \quad \delta_4 = K_{14}\alpha_4 + O(1)\Delta',\\
&\sigma_2' = \sigma_2 + \alpha_2 + K_{12}\alpha_4 + O(1)\Delta', \quad \sigma_3' = \sigma_3 + \alpha_3 + K_{13}\alpha_4 + O(1)\Delta',
\end{aligned}$$

where $\Delta' = |\beta_1|(|\alpha_2| + |\alpha_3|)$*,* $|K_{11}|_{\{\alpha_4=\alpha_3=\alpha_2=0, \sigma_2'=\sigma_{20}, \sigma_3'=\sigma_{30}\}} < 1$*, and* $\sum_{j=2}^4 |K_{1j}|$ *is bounded.*

Proof. We need to find the solution $(\delta_1, \sigma_2', \sigma_3', \delta_4)$ as a function of α_2, α_3, α_4, β_1, σ_2, σ_3, and U_b to

$$\Phi_4(\delta_4; G(\sigma_3', \sigma_2'; \Phi_1(\delta_1; U_b))) = G(\sigma_3, \sigma_2; \Phi_1(\beta_1, \Phi(\alpha_4, \alpha_3, \alpha_2, 0; U_b))). \tag{3.7}$$

Lemma 2.5 implies

$$\begin{aligned}
&\det\left(\frac{\partial \Phi_4(\delta_4; G(\sigma_3', \sigma_2'; \Phi_1(\delta_1; U_b)))}{\partial(\delta_4, \sigma_3', \sigma_2', \delta_1)}\right)\Bigg|_{\{\delta_4=\delta_1=0, \sigma_2'=\sigma_{20}, \sigma_3'=\sigma_{30}\}}\\
&= \det(\mathbf{r}_4(U_2), G_{\sigma_3}(\sigma_{30}, \sigma_{20}; U_1), G_{\sigma_2}(\sigma_{30}, \sigma_{20}; U_1), \nabla_U G(\sigma_{30}, \sigma_{20}; U_1) \cdot \mathbf{r}_1(U_1)) > 0.
\end{aligned}$$

Therefore, $(\delta_1, \sigma_2', \sigma_3', \delta_4)$ can be solved as a C^2-function of $(\alpha_2, \alpha_3, \alpha_4, \beta_1, \sigma_2, \sigma_3; U_b)$ uniquely. That is,

$$\sigma_i' = \sigma_i'(\alpha_2, \alpha_3, \alpha_4, \beta_1, \sigma_2, \sigma_3), \ i = 2, 3; \qquad \delta_j = \delta_j(\alpha_2, \alpha_3, \alpha_4, \beta_1, \sigma_2, \sigma_3), \ j = 1, 4,$$

where we have omitted U_b for simplicity. Using Lemma 3.1(i), we have

$$\begin{aligned}
\delta_1 &= \delta_1(\alpha_2, \alpha_3, \alpha_4, \beta_1, \sigma_2, \sigma_3) - \delta_1(\alpha_2, \alpha_3, 0, \beta_1, \sigma_2, \sigma_3) + \delta_1(\alpha_2, \alpha_3, 0, \beta_1, \sigma_2, \sigma_3)\\
&= K_{11}\alpha_4 + \beta_1 + O(1)(|\beta_1||\alpha_2| + |\beta_1||\alpha_3|),\\
\delta_4 &= \delta_4(\alpha_2, \alpha_3, \alpha_4, \beta_1, \sigma_2, \sigma_3) - \delta_4(\alpha_2, \alpha_3, 0, \beta_1, \sigma_2, \sigma_3) + \delta_4(\alpha_2, \alpha_3, 0, \beta_1, \sigma_2, \sigma_3)\\
&= K_{14}\alpha_4 + O(1)(|\beta_1||\alpha_2| + |\beta_1||\alpha_3|),\\
\sigma_i' &= \sigma_i'(\alpha_2, \alpha_3, \alpha_4, \beta_1, \sigma_2, \sigma_3) - \sigma_i'(\alpha_2, \alpha_3, 0, \beta_1, \sigma_2, \sigma_3) + \sigma_i'(\alpha_2, \alpha_3, 0, \beta_1, \sigma_2, \sigma_3)\\
&= K_{1i}\alpha_4 + \alpha_i + \sigma_i + O(1)(|\beta_1||\alpha_2| + |\beta_1||\alpha_3|), \qquad i = 2, 3,
\end{aligned}$$

where

$$K_{1i} = \int_0^1 \partial_{\alpha_4}\sigma_i'(\alpha_2, \alpha_3, \theta\alpha_4, \beta_1, \sigma_2, \sigma_3)d\theta, \quad i = 2, 3,$$
$$K_{1j} = \int_0^1 \partial_{\alpha_4}\delta_j(\alpha_2, \alpha_3, \theta\alpha_4, \beta_1, \sigma_2, \sigma_3)d\theta, \quad j = 1, 4.$$

When $\beta_1 = \alpha_4 = \alpha_3 = \alpha_2 = 0$, $\sigma_2 = \sigma_{20}$, and $\sigma_3 = \sigma_{30}$, it is clear that $|\partial_{\alpha_4}(\delta_1, \delta_4)|$ and $|\partial_{\alpha_4}(\sigma_2', \sigma_3')|$ are bounded. We can further claim that $|\partial_{\alpha_4}\delta_1| < 1$, which can be seen as follows:

Differentiate (3.7) with respect to α_4 and let $\beta_1 = \alpha_4 = \alpha_3 = \alpha_2 = 0$, $\sigma_2 = \sigma_{20}$, and $\sigma_3 = \sigma_{30}$. We obtain

$$\begin{aligned}
&\nabla_U G(\sigma_{30}, \sigma_{20}; U_1) \cdot \mathbf{r}_4(U_1)\\
&= \partial_{\alpha_4}\delta_4\, \mathbf{r}_4(U_2) + \partial_{\alpha_4}\sigma_3'\, G_{\sigma_3}(\sigma_{30}, \sigma_{20}; U_1)\\
&\quad + \partial_{\alpha_4}\sigma_2'\, G_{\sigma_2}(\sigma_{30}, \sigma_{20}; U_1) + \partial_{\alpha_4}\delta_1\, \nabla_U G(\sigma_{30}, \sigma_{20}; U_1) \cdot \mathbf{r}_1(U_1).
\end{aligned}$$

By Lemma 2.5, we have

$$\begin{aligned}
&|\partial_{\alpha_4}\delta_1|\\
&= \left|\frac{\det(\mathbf{r}_4(U_2), G_{\sigma_3}(\sigma_{30}, \sigma_{20}; U_1), G_{\sigma_2}(\sigma_{30}, \sigma_{20}; U_1), \nabla_U G(\sigma_{30}, \sigma_{20}; U_1) \cdot \mathbf{r}_4(U_1))}{\det(\mathbf{r}_4(U_2), G_{\sigma_3}(\sigma_{30}, \sigma_{20}; U_1), G_{\sigma_2}(\sigma_{30}, \sigma_{20}; U_1), \nabla_U G(\sigma_{30}, \sigma_{20}; U_1) \cdot \mathbf{r}_1(U_1))}\right|\\
&= \left|\frac{\kappa_4(U_2)\kappa_4(U_1)\rho_1^2 u_1^2 e^{\sigma_{20}+\sigma_{30}}(\lambda_4(U_2)e^{2\sigma_{20}+\sigma_{30}} - \lambda_4(U_1))}{\kappa_4(U_2)\kappa_1(U_1)\rho_1^2 u_1^2 e^{\sigma_{20}+\sigma_{30}}(\lambda_4(U_2)e^{2\sigma_{20}+\sigma_{30}} - \lambda_1(U_1))}\right|\\
&= \left|\frac{\lambda_4(U_2)e^{2\sigma_{20}+\sigma_{30}} - \lambda_4(U_1)}{\lambda_4(U_2)e^{2\sigma_{20}+\sigma_{30}} + \lambda_4(U_1)}\right| < 1.
\end{aligned}$$

This completes the proof. □

Remark 3.1. The essential feature of system (1.1) is that the reflection coefficient K_{11} is less than one, which also appears as the stability condition in [9, 18] for the strictly hyperbolic case.

The second case is when the weak waves approach the strong vortex sheet from above, as in Figure 7. By symmetry, we can easily obtain the following.

PROPOSITION 3.5. *Let $U_b \in O_\epsilon(U_1)$ and $U_m, U_a \in O_\epsilon(U_2)$ with*

$$\{U_b, U_m\} = (0, \sigma_2, \sigma_3, \alpha_4), \qquad \{U_m, U_a\} = (\beta_1, \beta_2, \beta_3, 0).$$

Then there exists a unique $(\delta_1, \sigma_2', \sigma_3', \delta_4)$ such that the Riemann problem (2.18) admits an admissible solution that consists of a weak 1-wave of strength δ_1, a strong vortex sheet of strength (σ_2', σ_3'), and a weak 4-wave of strength δ_4, that is,

$$\{U_b, U_a\} = (\delta_1, \sigma_2', \sigma_3', \delta_4).$$

Moreover,

$$\begin{aligned}
&\delta_1 = K_{21}\beta_1 + O(1)\Delta'', \quad \sigma_2' = \sigma_2 + \beta_2 + K_{22}\beta_1 + O(1)\Delta'',\\
&\sigma_3' = \sigma_3 + \beta_3 + K_{23}\beta_1 + O(1)\Delta'', \quad \delta_4 = \alpha_4 + K_{24}\beta_1 + O(1)\Delta'',
\end{aligned}$$

where $\sum_{j=1}^4 |K_{2j}|$ is bounded and $\Delta'' = |\alpha_4|(|\beta_2| + |\beta_3|)$.

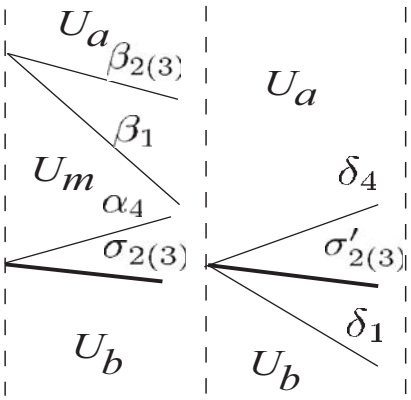


FIG. 7. *Weak waves approach the vortex sheet from above.*

4. Approximate solutions. In this section, we develop a modified Glimm difference scheme to construct a family of approximate solutions and establish their necessary estimates for the initial-boundary value problem (1.1) and (1.9)–(1.10) in the corresponding approximate domains $\Omega_{\Delta x}$.

4.1. A modified Glimm scheme. To define the scheme more clearly, we first use the fact that the boundary is a perturbation of the straight wall, that is,

$$\sup_{x\geq 0}|g'(x)|<\varepsilon \qquad \text{for sufficiently small } \varepsilon>0.$$

For any $\Delta x\geq 0$, set $a_k:=k\Delta x$ and $b_k:=y_k=g(k\Delta x)$ in (3.3) and (3.4), and follow the notation in subsection 3.2 (also see Figure 5). Then

$$m:=\sup_{k>0}\left\{\frac{y_k-y_{k-1}}{\Delta x}\right\}<\varepsilon. \tag{4.1}$$

Define

$$\Omega_{\Delta x}=\bigcup_{k\geq 1}\Omega_{\Delta x,k},$$

where $\Omega_{\Delta x,k}=\{(x,y):(k-1)\Delta x<x\leq k\Delta x, y>y_{k-1}+(x-(k-1)\Delta x)\tan(\omega_{k-1,k})\}$. Choose $\Delta y>0$ such that the Courant–Friedrichs–Lewy-type condition holds:

$$\frac{\Delta y-m\Delta x}{\Delta x}<\max_{j=1,4}\left(\sup_{U\in O_\varepsilon(U_1)\cup O_\varepsilon(U_2)}|\lambda_j(U)|\right).$$

Define

$$b_{k,n}=(2n+1+\theta_k)\Delta y+y_k,$$

where θ_k is randomly chosen in $(-1,1)$. Then we choose

$$P_{k,n}=(k\Delta x,b_{k,n}),\qquad k\geq 0,\quad n=0,1,2,\dots,$$

to be the mesh points and define the approximate solutions $U_{\Delta x,\theta}$ in $\Omega_{\Delta x}$ for any $\theta=(\theta_0,\theta_1,\theta_2,\dots)$ in an inductive way:

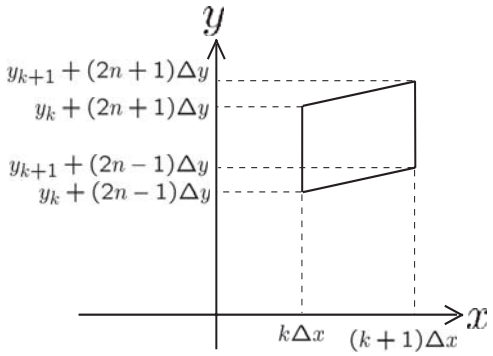


FIG. 8. *Diamond* $T_{k,n}$.

For $k=0$, we define $U_{\Delta x,\theta}$ in $\{0\le x<\Delta x\}\cap\Omega_{\Delta x}$ starting from

$$U_{\Delta x,\theta}|_{x=0,y>0}=\begin{cases} U_1, & 0<y<y_0^*,\\ U_2, & y>y_0^*.\end{cases}$$

Assume that $U_{\Delta x,\theta}$ has been constructed in $\{0\le x<k\Delta x\}$. Denoting, for $n\ge 1$,

$$U^0_{k,n}(y):=U_{\Delta x,\theta}(k\Delta x-,b_{k,n})\qquad \text{if } y\in(y_k+2n\Delta y,y_k+(2n+2)\Delta y),$$

then we define $U_{\Delta x,\theta}$ in $\{k\Delta x\le x<(k+1)\Delta x\}$ as follows: We first solve the following lateral Riemann problem in the diamond $T_{k,0}$, whose vertices are $((k+1)\Delta x,y_{k+1})$, $((k+1)\Delta x,y_{k+1}+\Delta y)$, $(k\Delta x,y_k)$, and $(k\Delta x,y_k+\Delta y)$:

$$\begin{cases} W(U_k)_x+H(U_k)_y=0 & \text{in } T_{k,0},\\ U_k|_{x=k\Delta x}=U^0_{k,n}, & \\ (u_k,v_k)\cdot\mathbf{n}_k=0 & \text{on } \Gamma_k.\end{cases}$$

We thus obtain the lateral Riemann solution U_k in $T_{k,0}$ as constructed in section 2.3 and define

$$U_{\Delta x,\theta}=U_k\qquad\text{in } T_{k,0}.$$

Then we solve the following Riemann problem in each diamond $T_{k,n}$ for $n\ge 1$, whose vertices are $((k+1)\Delta x,y_{k+1}+(2n-1)\Delta y)$, $((k+1)\Delta x,y_{k+1}+(2n+1)\Delta y)$, $(k\Delta x,y_k+(2n-1)\Delta y)$, and $(k\Delta x,y_k+(2n+1)\Delta y)$ (see Figure 8):

$$\begin{cases} W(U_k)_x+H(U_k)_y=0 & \text{in } T_{k,n},\\ U_k|_{x=k\Delta x}=U^0_{k,n}.\end{cases}$$

We obtain the Riemann solution $U_k(x,y)$ in $T_{k,n}$ as constructed in sections 2.4–2.5 and define

$$U_{\Delta x,\theta}=U_k\qquad\text{in } T_{k,n},\ n\ge 1.$$

In this way, we have constructed the approximate solutions $U_{\Delta x,\theta}(x,y)$ globally provided that we can obtain a uniform bound of the approximate solutions.

4.2. Glimm-type functional and its bounds. In this section, we prove that the approximate solutions can be well defined in $\Omega_{\Delta x}$ indeed via the steps in section 4.1 by providing a uniform bound for them. First, we introduce the following lemma.

LEMMA 4.1.

(i) *If* $\{U_b, U_a\} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ *with* $U_b, U_a \in O_\varepsilon(U_i)$ *for fixed* $i = 1$ *or* 2, *then*

$$|U_b - U_a| \le B_1(|\alpha_1| + |\alpha_2| + |\alpha_3| + |\alpha_4|)$$

with $B_1 = \max_{i=1,2,1\le j\le 4} \big(\sup_{U\in O_\varepsilon(U_i)} |\partial_{\alpha_j}\Phi(\alpha_4, \alpha_3, \alpha_2, \alpha_1; U)|\big)$.

(ii) *For any* $\sigma_j \in O_{\hat\varepsilon}(\sigma_{j0})$ *so that* $G(\sigma_3, \sigma_2, U_b) \subset O_\varepsilon(U_2)$ *when* $U_b \in O_\varepsilon(U_1)$ *for some* $\hat\varepsilon = \hat\varepsilon(\varepsilon) \to 0$ *as* $\varepsilon \to 0$,

$$|G(\sigma_3, \sigma_2, U_b) - G(\sigma_{30}, \sigma_{20}, U_b)| \le \tilde B(|\sigma_3 - \sigma_{30}| + |\sigma_2 - \sigma_{20}|)$$

with $\tilde B = \max_{j=2,3} \big(\sup_{\sigma_j\in O_{\hat\varepsilon}(\sigma_{j0})} |G'_{\sigma_j}(\sigma_3, \sigma_2, U_b)|\big)$.

Next, we show that $U_{\Delta x,\theta}$ can be globally defined. Assume that $U_{\Delta x,\theta}$ has been defined in $\{x < k\Delta x\} \cap \Omega_{\Delta x}$ by the steps in section 4.1 and assume that the following conditions are satisfied:

$$C_1(k-1): \begin{cases} \text{In each } \Omega_{\Delta x,j} \text{ for } 0 \le j \le k-1, \text{ there is a strong vortex sheet } \chi^{(j)} \\ \text{in } U_{\Delta x,\theta} \text{ with strength } (\sigma_2^{(j)}, \sigma_3^{(j)}) \text{ so that } \sigma_i^{(j)} \in O_{\hat\varepsilon}(\sigma_{i0}), \text{ which} \\ \text{divides } \Omega_{\Delta x,j} \text{ into two parts: } \Omega^{(1)}_{\Delta x,j} \text{ and } \Omega^{(2)}_{\Delta x,j}, \text{ where } \Omega^{(1)}_{\Delta x,j} \text{ is} \\ \text{the part bounded by } \chi^{(j)} \text{ and } \Gamma_j, \text{ where } \Omega_{\Delta x,j} \text{ and } \Gamma_j \text{ are defined} \\ \text{by (3.3)}; \end{cases}$$

$$C_2(k-1): U_{\Delta x,\theta}|_{\Omega^{(1)}_{\Delta x,j}} \in O_\varepsilon(U_1) \text{ and } U_{\Delta x,\theta}|_{\Omega^{(2)}_{\Delta x,j}} \in O_\varepsilon(U_2),\ 0 \le j \le k-1;$$

$$C_3(k-1): \begin{cases} \{\chi^{(j)}\}_{j=0}^{k-1} \text{ forms } \chi_{\Delta x,\theta} : y = \chi_{\Delta x,\theta}(x), \text{ called an approximate} \\ \text{vortex sheet, emanating from point } (0, y_0^*). \end{cases}$$

Here and in what follows, we always denote by $\chi^{(j)}$ the strong vortex sheet front with strength $(\sigma_2^{(j)}, \sigma_3^{(j)})$. Then we prove that $U_{\Delta x,\theta}$ can be defined in $\Omega_{\Delta x,k}$ and satisfies $C_1(k), C_2(k)$, and $C_3(k)$.

From the construction steps in section 4.1, we first define $U_{\Delta x,\theta}$ and the strong vortex sheet $\chi^{(k)}$ in $\Omega_{\Delta x,k}$ so that there exists a diamond $\Lambda_{k,n(k)}$ such that $\chi^{(k-1)}$ enters $\Lambda_{k,n(k)}$ and $\chi^{(k)}$ emanates from the center of $\Lambda_{k,n(k)}$. We extend $\chi_{\Delta x,\theta}$ to $\Omega_{\Delta x,k}$ such that $\chi_{\Delta x,\theta} = \chi^{(k)}$ in $\Omega_{\Delta x,k}$ and define $\Omega^{(1)}_{\Delta x,j}$ and $\Omega^{(2)}_{\Delta x,j}$ in the same way as in $C_1(k-1)$. Then it suffices to impose some conditions so that $C_2(k-1)$ holds and $\sigma_i^{(k)} \in O_{\hat\varepsilon}(\sigma_{i0}), i = 2, 3$.

To achieve this, we will establish the bound on the total variation of $U_{\Delta x,\theta}$ on a class of space-like curves. Denote

$$N(\theta_{k+1}, n) = \begin{cases} P_{k+1,n} & \text{if } \theta_{k+1} \le 0, \\ P_{k+1,n-1} & \text{if } \theta_{k+1} > 0, \end{cases} \qquad S(\theta_k, n) = \begin{cases} P_{k-1,n-1} & \text{if } \theta_k \le 0, \\ P_{k-1,n} & \text{if } \theta_k > 0. \end{cases}$$

Then we introduce the following.

DEFINITION 4.2. *A* j*-mesh curve* J *is defined to be an unbounded space-like curve lying in the strip* $\{(j-1)\Delta x \le x \le (j+1)\Delta x\}$ *and consisting of the segments of the form* $P_{k,n-1}N(\theta_{k+1}, n)$, $P_{k,n-1}S(\theta_k, n)$, $S(\theta_k, n)P_{k,n}$, *and* $N(\theta_{k+1}, n)P_{k,n}$.

This means that we connect the mesh point $P_{k,n}$ by two line segments to the two mesh points $P_{k-1,n-1}$ and $P_{k-1,n}$ if $\theta_k \le 0$, or we connect the mesh point $P_{k,n}$ by two line segments to the two mesh points $P_{k-1,n}$ and $P_{k-1,n+1}$ if $\theta_k > 0$ (see Figure 9).

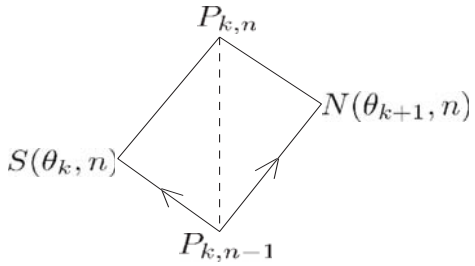


FIG. 9. *Interaction diamond $\Lambda_{k,n}$ and orientation of the segments.*

Clearly, for any $k > 0$, each k-mesh curve I divides the plane $\mathbb{R}^2$ into part I^+ and part I^-, where I^- is the one containing the set $\{x < 0\}$. As in Glimm [11], we also partially order these mesh curves by saying $J > I$ if every point of the mesh curve J is either on I or contained in I^+, and we call J an immediate successor to I if $J > I$ and every mesh point of J except one is on I.

With such mesh curves J, we associate the Glimm-type functional $F_s(J)$ on J.

DEFINITION 4.3. *We define*

$$F_s(J) = C^*(|\sigma_2^J - \sigma_{20}| + |\sigma_3^J - \sigma_{30}|) + F(J)$$

with

$$\begin{aligned}
&F(J) = L(J) + KQ(J),\\
&L(J) = L^1(J) + L^2(J),\\
&L^1(J) = K_1^* L_0(J) + K_{11}^* L_1^1(J) + K_{12}^* L_2^1(J) + K_{13}^* L_3^1(J) + L_4^1(J),\\
&L^2(J) = K_{21}^* L_1^2(J) + K_{22}^* L_2^2(J) + K_{23}^* L_3^2(J) + K_{24}^* L_4^2(J),\\
&Q(J) = \sum\{|\alpha_i||\beta_j| \ : \ \textit{both } \alpha_i \textit{ and } \beta_j \textit{ cross } J \textit{ and approach}\},
\end{aligned}$$

and $L_0(J) = \sum\{|\omega(C_k)| \ : \ C_k \in \Omega_J\}$,

$$L_j^i(J) = \sum\{|\alpha_j| \ : \ \alpha_j \textit{ crosses } J \textit{ in region (i)}\}, \quad i = 1, 2, \ j = 1, 2, 3, 4,$$

where K and C^* will be defined later, while Ω_J is the set of the corner points C_k lying in J^+, i.e.,

$$\Omega_J = \{C_k \in J^+ \cap \partial\Omega_{\Delta x} \ : \ k \geq 0\},$$

(σ_2^J, σ_3^J) stands for the strength of the strong vortex sheet crossing J, and $K_1^*, K_{11}^*, K_{12}^*$, and K_{13}^* for the constants associated with region (1), and $K_{21}^*, K_{22}^*, K_{23}^*$, and K_{24}^* for the constants associated with region (2) which satisfy the following conditions:

$$K_1^* > |K_{b0}|, \quad K_{1j}^* > |K_{bj}|, \qquad j = 1, 2, 3,$$

and

$$K_{24}^* < \frac{1 - |K_{11}|K_{11}^*}{|K_{14}|}, \quad K_{21}^* > |K_{21}|K_{11}^* + |K_{24}|K_{24}^*,$$

while K_{12}^*, K_{13}^*, K_{22}^*, and K_{23}^* are arbitrarily large positive constants. These conditions can be achieved from our discussions of the properties of K_{bj}, K_{1j}, and K_{2j}, $j = 1, 2, 3, 4$, as in Propositions 3.3–3.5 in section 3.

From now on, we denote by $M > 0$ a universal constant, depending only on the parameters in the system and the boundary function $g(x)$, which may be different at each occurrence, and $O(1)$ is the quantity that is bounded by M. Now we prove the decreasing property of our functional F_s. We first have the following.

PROPOSITION 4.4. *Suppose that the boundary function $g(x)$ satisfies* (4.1)*, and I and J are two k-mesh curves such that J is an immediate successor of I. Suppose that*

$$\left|U_{\Delta x,\theta}|_{I\cap(\,\Omega^{(i)}_{\Delta x,k-1}\cup\,\Omega^{(i)}_{\Delta x,k})}-U_i\right|<\varepsilon,\ i=1,2;\qquad |\sigma_j^I-\sigma_{j0}|<\hat{\varepsilon}(\varepsilon),\ j=2,3,$$

for some $\hat{\varepsilon}(\varepsilon)>0$ determined in Lemma 4.1*. Then there exist constants $\tilde{\varepsilon}>0, K>0$, and $C^*>1$, depending only on the parameters in system* (1.6) *and states U_1 and U_2, such that if $F_s(I)\le\tilde{\varepsilon}$, then*

$$F_s(J)\le F_s(I),$$

and hence

$$\left|U_{\Delta x,\theta}|_{J\cap(\,\Omega^{(i)}_{\Delta x,k-1}\cup\,\Omega^{(i)}_{\Delta x,k})}-U_i\right|<\varepsilon,\ i=1,2;\quad |\sigma_j^J-\sigma_{j0}|<\hat{\varepsilon}(\varepsilon),\ j=2,3.$$

Proof. Let Λ be the diamond that is formed by I and J. We can always assume that $I=I_0\cup I'$ and $J=J_0\cup J'$ such that $\partial\Lambda=I'\cup J'$. We divide our proof into four cases depending on the location of the diamond.

Case 1 (interior weak-weak interaction). Λ lies in the interior of $\Omega_{\Delta x}$ and does not touch $\chi_{\Delta x,\theta}$ so that only weak waves enter Λ. We prove only the case when Λ is in between $\partial\Omega_{\Delta x}$ and $\chi_{\Delta x,\theta}$ since the proof is similar when Λ is above $\chi_{\Delta x,\theta}$. Denote $Q(\Lambda)=\Delta(\alpha,\beta)$, where $\Delta(\alpha,\beta)$ is as defined in Proposition 3.2.

Then

$$L^1(J)-L^1(I)\le(1+K_{11}^*+K_{12}^*+K_{13}^*)MQ(\Lambda),$$

and

$$\begin{aligned}&Q(J)-Q(I)\\&=\left(Q(I_0)+\sum_{i=2}^4Q(\gamma_i,I_0)\right)-\left(Q(I_0)+Q(\Lambda)+\sum_{i=2}^4Q(\alpha_i,I_0)+\sum_{i=2}^4Q(\beta_i,I_0)\right)\\&=Q(O(1)Q(\Lambda),I_0)-Q(\Lambda)=(O(1)L(I_0)-1)Q(\Lambda)\le-\frac{1}{2}Q(\Lambda),\end{aligned}$$

since $L(I_0)\ll 1$. Hence

$$\begin{aligned}F(J)-F(I)&=L(J)-L(I)+K(Q(J)-Q(I))\\&\le\left((1+K_{11}^*+K_{12}^*+K_{13}^*)M-\frac{K}{2}\right)Q(\Lambda)\\&\le-\frac{1}{4}Q(\Lambda)\end{aligned}$$

by choosing suitably large K.

Case 2 (near the boundary). The diamond Λ touches the approximate boundary $\partial\Omega_{\Delta x}$, but not the strong vortex sheet $\chi_{\Delta x,\theta}$. Then $\Omega_J=\Omega_I\backslash\{C_k\}$ for certain k and $\sigma_j^I=\sigma_j^J, j=2,3$.

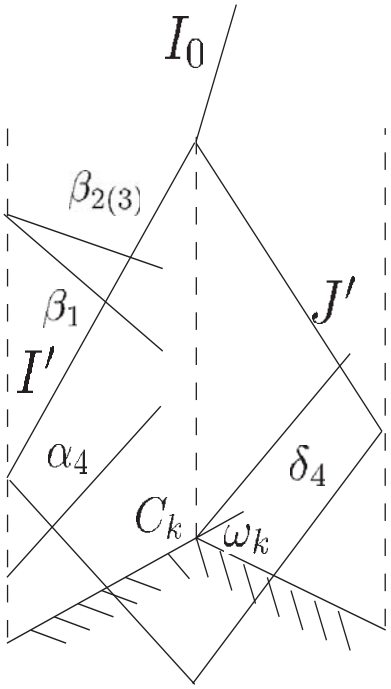


FIG. 10. *Case* 2: *near the boundary.*

Let δ_4 be the weak 4-wave going out of Λ through J', and let $\alpha_4, \beta_1, \beta_2$, and β_3 be the weak waves entering Λ through I', as shown in Figure 10. Then

$$
\begin{aligned}
&L_0(J) - L_0(I) = -|\omega_k|,\\
&L_i^1(J) - L_i^1(I) = \sum_{\gamma_i \text{ crosses } I_0} |\gamma_i| - \left(|\beta_i| + \sum_{\gamma_i \text{ crosses } I_0} |\gamma_i| \right) = -|\beta_i|, \quad i = 1,2,3,\\
&L_4^1(J) - L_4^1(I) = \left(|\delta_4| + \sum_{\gamma_4 \text{ crosses } I_0} |\gamma_4| \right) - \left(|\alpha_4| + \sum_{\gamma_4 \text{ crosses } I_0} |\gamma_4| \right)\\
&\qquad\qquad = |\delta_4| - |\beta_4| \le |K_{b1}||\beta_1| + |K_{b2}||\beta_2| + |K_{b3}||\beta_3| + |K_{b0}||\omega_{k+1}|,
\end{aligned}
$$

where the last step is from Proposition 3.3. Therefore, we have

$$L(J) - L(I) \le (|K_{b0}| - K_1^*)|\omega_{k+1}| + (|K_{b1}| - K_{11}^*)|\beta_1| + (|K_{b2}| - K_{12}^*)|\beta_2| + (|K_{b3}| - K_{13}^*)|\beta_3|.$$

Then, from our requirement in Definition 4.3, we have

$$L(J) \le L(I).$$

Notice that $F_s(I) \le \epsilon'$ implies $L(I) \le \epsilon'$. Hence, the higher order term $Q(I)$ can always be bounded by the linear term $L(I)$, and we can easily conclude $F(J) \le F(I)$.

Case 3. The diamond Λ covers a part of $\partial\Omega_{\Delta x}$, and $S_*(\sigma_{(k-1)})$ emanates from C_{k-1} and enters Λ.

Case 3.1 (near the strong vortex sheet I). We first investigate the situation when Λ lies in region (1) and the strong vortex sheet $\chi^{(k-1)}$ enters Λ. Then $\chi^{(k)}$ is generated from the inside of Λ, $\sigma_j^I = \sigma_j^{(k-1)}$, and $\sigma_j^J = \sigma_j^{(k)}$, $j = 2, 3$.

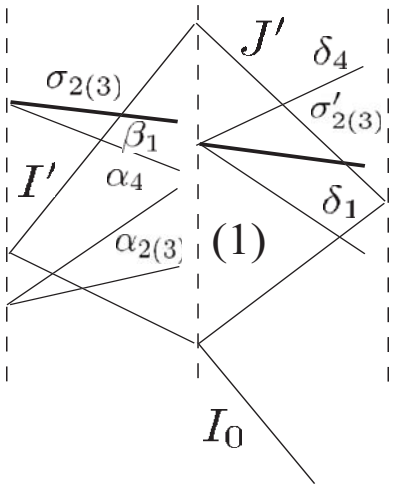


FIG. 11. *Case* 3.1: *near the strong vortex sheet* (I).

Let δ_4 and δ_1 be the weak waves going out of Λ through J', and let $\beta_1, \alpha_4, \alpha_3$, and α_2 be the weak waves entering Λ through I', as shown in Figure 11. Then

$$\begin{aligned}
L_4^2(J) - L_4^2(I) &= \left(|\delta_4| + \sum_{\gamma_4 \text{ crosses } I_0} |\gamma_4| \right) - \sum_{\gamma_4 \text{ crosses } I_0} |\gamma_4| \\
&= |\delta_4| \le |K_{14}||\alpha_4| + O(1)|\beta_1|(|\alpha_2| + |\alpha_3|), \\
L_1^1(J) - L_1^1(I) &= \left(|\delta_1| + \sum_{\gamma_1 \text{ crosses } I_0} |\gamma_1| \right) - \left(|\beta_1| + \sum_{\gamma_1 \text{ crosses } I_0} |\gamma_1| \right) \\
&= |\delta_1| - |\beta_1| \le |K_{11}||\alpha_4| + O(1)|\beta_1|(|\alpha_2| + |\alpha_3|), \\
L_j^1(J) - L_j^1(I) &= \sum_{\gamma_j \text{ crosses } I_0} |\gamma_j| - \left(|\alpha_j| + \sum_{\gamma_j \text{ crosses } I_0} |\gamma_j| \right) = -|\alpha_j|, \quad j = 2, 3, 4.
\end{aligned}$$

The above inequalities are from Propositions 3.4–3.5. Hence,

$$L(J) - L(I) \le (|K_{14}|K_{14}^* + |K_{11}|K_{11}^* - 1)|\alpha_4| - (K_{12}^* + O(1)|\beta_1|)|\alpha_2| - (K_{13}^* + O(1)|\beta_1|)|\alpha_3|$$

with $|K_{14}|K_{24}^* + |K_{11}|K_{11}^* - 1 < 0$ by our appropriate choice of the constants.

Furthermore, since the higher order term $Q(I)$ can always be bounded by the linear term $L(I)$ and

$$|\sigma_2^J - \sigma_2^I| + |\sigma_3^J - \sigma_3^I| \le (|K_{12}| + |K_{13}|)|\alpha_4| + (1 + O(1)|\beta_1|)|\alpha_2| + (1 + O(1)|\beta_1|)|\alpha_3|$$

with $|K_{12}|$ and $|K_{13}|$ bounded, we can choose C^* suitably small such that

$$F_s(J) - F_s(I) \le C^*(|\sigma_2^J - \sigma_2^I| + |\sigma_3^J - \sigma_3^I|) + F(J) - F(I) \le 0.$$

Therefore, we have $F_s(J) \le F_s(I)$.

Case 3.2 (near the strong vortex sheet II). Now, we investigate the second situation when Λ lies in region (2) and the strong vortex sheet $\chi^{(k-1)}$ enters Λ. Then $\chi^{(k)}$ is generated from the interior of Λ, $\sigma_j^I = \sigma_j^{(k-1)}$, and $\sigma_j^J = \sigma_j^{(k)}$.

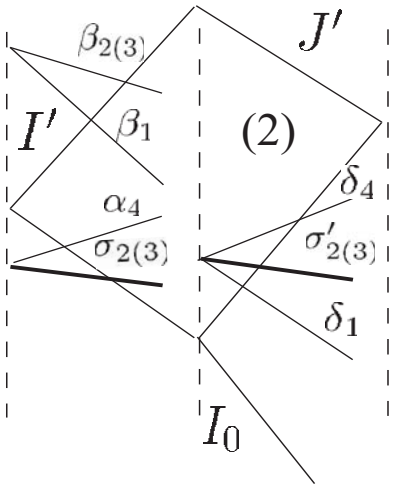


FIG. 12. *Case* 3.2: *near the strong vortex sheet* (II).

Let δ_4 and δ_1 be the weak waves going out of Λ through J', and let β_1, β_2, and β_3 be the weak waves entering Λ through I', as shown in Figure 12. Then

$$\begin{aligned}
L_1^1(J)-L_1^1(I) &= \left(|\delta_1|+\sum_{\gamma_1 \text{ crosses } I_0}|\gamma_1|\right)-\sum_{\gamma_1 \text{ crosses } I_0}|\gamma_1| \\
&= |\delta_1| \le |K_{21}||\beta_1|+O(1)(|\alpha_4||\beta_2|+|\alpha_4||\beta_3|), \\
L_4^2(J)-L_4^2(I) &= \left(|\delta_4|+\sum_{\gamma_4 \text{ crosses } I_0}|\gamma_4|\right)-\left(|\alpha_4|+\sum_{\gamma_4 \text{ crosses } I_0}|\gamma_4|\right) \\
&= |\delta_4|-|\alpha_4| \le |K_{24}||\beta_1|+O(1)(|\alpha_4||\beta_2|+|\alpha_4||\beta_3|), \\
L_j^2(J)-L_j^2(I) &= \sum_{\gamma_j \text{ crosses } I_0}|\gamma_j|-\left(|\beta_j|+\sum_{\gamma_j \text{ crosses } I_0}|\gamma_j|\right)=-|\beta_j|, \quad j=1,2,3.
\end{aligned}$$

The above inequalities are from Propositions 3.4–3.5. Hence

$$L(J)-L(I) \le (|K_{21}|K_{11}^*+|K_{24}|K_{24}^*-K_{21}^*)|\beta_1|-(K_{22}^*+O(1)|\alpha_4|)|\beta_2|-(K_{23}^*+O(1)|\alpha_4|)|\beta_3|$$

with $|K_{21}|K_{11}^*+|K_{24}|K_{24}^*-K_{21}^*<0$ by our appropriate choice of the constants.

Similar to the analysis for Case 3.1, we again have $F_s(J)\le F_s(I)$.

Then, from Lemma 4.1, there exists $\tilde\epsilon>0$ such that, when $F(I)<\tilde\epsilon$, we have $|U-U_1|<\epsilon$, or $|U-U_2|<\epsilon$. □

Let I_k be the k-mesh curve lying in $\{(j-1)\Delta x\le x\le j\Delta x\}$. From Proposition 4.4, we obtain the following theorem for any $k\ge 1$.

THEOREM 4.5. *Suppose that the boundary function $g(x)$ satisfies* (4.1). *Let ε, $\tilde\varepsilon$, $\hat\varepsilon(\varepsilon)$, K, and C^* be the constants specified in Proposition* 4.4. *If the induction hypotheses $C_1(k-1)$, $C_2(k-1)$, and $C_3(k-1)$ hold and if $F_s(I_{k-1})\le\tilde\varepsilon$, then*

$$\left|U_{\Delta x,\theta}\big|_{\Omega^{(i)}_{\Delta x,k}}-U_i\right|<\varepsilon,\ i=1,2; \qquad U_{\Delta x,\theta}|_{x<0}=\begin{cases} U_1, & y<y_0^*,\\ U_2, & y>y_0^*,\end{cases}$$

$$|\sigma_j^{(k)}-\sigma_{j0}|<\hat\varepsilon(\varepsilon),\ j=2,3,$$

and

$$F_s(I_k) \le F_s(I_{k-1}). \tag{4.2}$$

Moreover, we obtain the following.

THEOREM 4.6. *There exists $\varepsilon > 0$ such that if $TV(g'(\cdot)) < \varepsilon$, then, for any $\theta \in \prod_{k=0}^{\infty}(-1,1)$ and every $\Delta x > 0$, the modified Glimm scheme defines a family of global approximate solutions $U_{\Delta x,\theta}$ and the corresponding approximate strong vortex sheet fronts $\chi_{\Delta x,\theta}$ in $\Omega_{\Delta x,\theta}$ which satisfy $C_1(k-1)$, $C_2(k-1)$, $C_3(k-1)$, and (4.2) for any $k \ge 1$. In addition,*

$$TV\{U_{\Delta x,\theta}(k\Delta x-,\cdot) : [y_k,\infty)\} < M\,TV(g'(\cdot))$$

for any $k \ge 0$ and

$$|\chi_{\Delta x,\theta}(x+h) - \chi_{\Delta x,\theta}(x)| \le M|h| + 2\Delta x$$

for any $x \ge 0$ and $h > 0$.

4.3. Estimates on the approximate vortex sheets. We use the notation and estimates in the previous section and define

$$s_{\Delta x,\theta}(x) = \left.\frac{v_{\Delta x,\theta}}{u_{\Delta x,\theta}}\right|_{\{y=\chi_{\Delta x,\theta}\}} = s^{(k)}|_{\{y=\chi_{\Delta x,\theta}\}} \qquad \text{if } x \in (k\Delta x,(k+1)\Delta x],$$

$$E_{\Delta x,\theta}(\Lambda) \begin{cases} Q(\Lambda) & \text{for Case 1,} \\ |\omega_k| + |\beta_1| + |\beta_2| + |\beta_3| & \text{for Case 2,} \\ |\alpha_4| & \text{for Case 3.1,} \\ |\beta_1| & \text{for Case 3.2.} \end{cases}$$

Let $L_{\Delta x,\theta}(\Gamma_b)$ be the sum of the strengths of 4-waves leaving Γ_b, and let $L^2_{\Delta x,\theta}$ and $L^1_{\Delta x,\theta}$ be the sum of the strengths of all 4-waves and all 1-waves, respectively, leaving the vortex sheet $\chi_{\Delta x,\theta}$. Then, by Proposition 4.4, we have the following lemma.

LEMMA 4.7. *There exists a constant M_1, independent of $\Delta x, \theta$, and $U_{\Delta x,\theta}$, such that*

$$\sum_{\Lambda} E_{\Delta x,\theta}(\Lambda) \le M_1,$$

where the summation is over all the diamonds.

Then we have the following lemma.

LEMMA 4.8. *There exist constants M_2 and M_3, independent of $\Delta x, \theta$, and $U_{\Delta x,\theta}$, such that*

$$L^2_{\Delta x,\theta} + L^1_{\Delta x,\theta} \le M_2, \qquad TV(s_{\Delta x,\theta}) \le M_3.$$

5. Global entropy solutions. In this section we establish the convergence of the approximate solutions to a global entropy solution and prove the nonlinear stability and asymptotic behavior of the strong vortex sheet under the BV perturbation.

5.1. Convergence of the approximate solutions. Let the line $x = a > 0$ intersect $\partial\Omega_{\Delta x} = \cup_{k\geq 1}\{\overline{C_{k-1}C_k}\}$ at point $(a, p_a^{\Delta x})$. Similarly to [26], by Theorem 4.6, we have the following lemma.

LEMMA 5.1. *For any $h > 0$ and $x \geq 0$, there exists a constant M independent of $\Delta x, \theta$, and h such that*

$$\int_0^\infty |U_{\Delta x,\theta}(x+h, y+p_{x+h}^{\Delta x}) - U_{\Delta x,\theta}(x, y+p_x^{\Delta x})|dy \leq M|h|.$$

Denote

$$J(\theta, \Delta x, \phi) = \sum_{k=1}^{\infty}\int_0^\infty \phi(k\Delta x, y+y_k)\cdot(U_{\Delta x,\theta}(k\Delta x+, y+y_k) - U_{\Delta x,\theta}(k\Delta x-, y+y_k))dy,$$

where $\phi \in C_0^\infty(\mathbb{R}^2;\mathbb{R}^4)$. Following the steps in [11], we have the following.

LEMMA 5.2. *There exist a null set $N \subset \Pi_{k=0}^{\infty}(-1,1)$ and a subsequence $\{\Delta x_j\}_{j=1}^{\infty} \subset \{\Delta x\}$, which tends to 0, such that*

$$J(\theta, \Delta x_j, \phi) \longrightarrow 0 \qquad \text{when } \Delta x_j \to 0$$

for any $\theta \in \Pi_{k=0}^{\infty}(-1,1)\backslash N$ and $\phi \in C_0^\infty(\mathbb{R}^2;\mathbb{R}^4)$.

To establish the main theorem, we need to estimate the slope of the approximate strong vortex sheet fronts. Let

$$\tilde{d}_k = \frac{s^{(k-1)}\Delta x - (y_k - y_{k-1})}{\Delta y} \quad \text{and} \quad d_k = \begin{cases} \tilde{d}_k - 1 & \text{if } \tilde{d}_k > 0, \\ \tilde{d}_k + 1 & \text{if } \tilde{d}_k < 0. \end{cases}$$

Then, by the choice of Δx and $\{y_k\}$ and by Lemma 4.8, we find that $d_k \in (-1,1)$, which depends only on $\{\theta_l\}_{l=1}^{k-1}$. Thus we define

$$I(x, \Delta x, \theta) = \sum_{k=1}^{[x/\Delta x]} I_k(\Delta x, \theta),$$

where $I_k(\Delta x, \theta) = \mathbf{1}_{(-1,d_k)}(\theta_k)(d_k - 1)\Delta y + \mathbf{1}_{(d_k,1)}(\theta_k)(d_k+1)\Delta y$, $\mathbf{1}_A$ denotes the characteristic function of the set A, and $[x/\Delta x]$ denotes the largest integer less than or equal to $x/\Delta x$. Notice that $I_k(\Delta x, \theta)$ is the jump of the function $y = \chi_{\Delta x,\theta}(x)$ at $x = k\Delta x$ and is a measurable function of $(\Delta x, \theta)$, which depends only on $U_{\Delta x,\theta}|_{\{0\leq x\leq k\Delta x\}}$ and $\{\theta_l\}_{l=0}^{k}$.

LEMMA 5.3.

(i) *For any $x \geq 0$, $\Delta x > 0$, and $\theta \in \Pi_{k=0}^{\infty}(-1,1)$, we have*

$$\chi_{\Delta x,\theta}(x) = -I(x, \Delta x, \theta) + \int_0^x s_{\Delta x,\theta}(s)ds.$$

(ii) *There is a null set N_1 and a subsequence $\{\Delta_l\}_{l=1}^{\infty} \subset \{\Delta x_j\}_{j=1}^{\infty}$ such that*

$$\int_0^\infty e^{-x}|I(x, \Delta_l, \theta)|^2 dx \longrightarrow 0 \qquad \text{when } \Delta_l \to 0$$

for any $\theta \in \Pi_{k=0}^{\infty}(-1,1)\backslash N_1$.

Proof. Part (i) can be obtained by a direct calculation. We will focus only on part (ii). As in [11], let $d\theta = \Pi_{k=0}^{\infty}(d\theta_k/2)$. Then, for any $l > j$, we have

$$\int I_l I_j d\theta = \int \left(I_j \int I_l d\theta_l \right) \Pi_{i=1}^{l-1} d\theta_i = 0.$$

Therefore, we can deduce

$$\int |I(x, \Delta x, \theta)|^2 d\theta = \sum_{k=1}^{[x/\Delta x]} \int |I_k(\Delta x, \theta)|^2 d\theta \le 4 \left| \frac{\Delta y}{\Delta x} \right|^2 x \Delta x.$$

Then, by choosing a subsequence $\{\Delta_l\}_{l=1}^{\infty} \subset \{\Delta x_j\}_{j=1}^{\infty}$ with $\sum_{l=0}^{\infty} \Delta_l < \infty$ as in Lemma 5.2, we arrive at (ii). ∎

Then, by Theorem 4.6 and Lemmas 5.1–5.2, we have the following theorem.

THEOREM 5.4 (existence and stability). *There exist $\varepsilon > 0$ and $C > 0$ such that if* (1.8) *holds, then, for each $\theta \in (\Pi_{k=0}^{\infty}(-1, 1))\backslash(N \cup N_1)$, there exists a subsequence $\{\Delta_l\}$ of mesh sizes with $\Delta_l \to 0$ as $l \to \infty$ and a pair of functions $U_\theta \in BV(\Omega; O_\varepsilon(U_1) \cup O_\varepsilon(U_2))$ and $\chi_\theta \in Lip(\mathbb{R}_+; \mathbb{R}_+)$ with $\chi_\theta(0) = y_0^*$ such that*

(i) *$U_{\Delta_l,\theta}(x, \cdot)$ converges to $U_\theta(x, \cdot)$ in $L^1(g(x), \infty)$ for every $x > 0$, and U_θ is a global entropy solution of problem* (1.6) *and* (1.9)–(1.10) *in Ω and satisfies* (1.11) *and the initial-boundary data* (1.9)–(1.10) *in the trace sense;*

(ii) *$\chi_{\Delta_l,\theta}$ converges to χ_θ uniformly in any bounded x-interval such that* (1.12) *holds;*

(iii) *$s_{\Delta_l,\theta}$ converges to $s_\theta \in BV([0, \infty))$ a.e. with $|s_\theta| \le \hat{\varepsilon} \le \epsilon$ and*

$$\chi_\theta'(x) = s_\theta(x).$$

In addition, if θ is equidistributed, then $\chi_\theta(x) > g(x)$ for any $x > 0$ and the Rankine–Hugoniot conditions hold a.e. along the curve $\{y = \chi_\theta(x)\}$.

The proofs of (i) and (ii) and the convergence proof of $s_{\Delta_l,\theta}$ in (iii) can be carried out in the same way as in the standard cases (see [6, 11, 12, 25]) by using the structure of the approximate solutions. In particular, for any $\varphi \in C_0^{\infty}(\mathbb{R}^2; \mathbb{R})$,

$$\begin{aligned} &\int_{\Omega_{\Delta_x,\theta}} (\rho^{\Delta x,\theta} u^{\Delta x,\theta} \varphi_x + \rho^{\Delta x,\theta} v^{\Delta x,\theta} \varphi_y) dx dy \\ &= \int_{\Omega} \chi_{\Omega_{\Delta x,\theta}} (\rho^{\Delta x,\theta} u^{\Delta x,\theta} \varphi_x + \rho^{\Delta x,\theta} v^{\Delta x,\theta} \varphi_y) dx dy \end{aligned}$$

weak-star converges, and hence the initial condition is satisfied by the trace theorem for BV functions (cf. [23]). Similarly, the boundary condition can be shown to be satisfied. The equality in (iii) can be deduced from section 4.3 and the result on the convergence of $\{\chi_{\Delta_l,\theta}\}$ and $\{s_{\Delta_l,\theta}\}$.

5.2. Asymptotic behavior of the strong vortex sheet in the flow direction. As in Theorem 5.4, let $\theta \in (\Pi_{k=0}^{\infty}(-1, 1))\backslash(N \cup N_1)$ be equidistributed, and let U_θ be the solution and χ_θ its vortex sheet, respectively. By Theorem 5.4, we conclude that the solution U_θ contains at most countable vortex sheets, shocks, and points of wave interactions. Moreover, we can modify the solution U_θ such that U_θ is continuous except on the vortex sheets, shocks, and points of wave interactions (see [10, 12, 16]). Then we have the following lemma.

LEMMA 5.5. *The total variation of $(p_\theta, v_\theta/u_\theta)$ is sufficiently small above the Lipschitz wall and*

$$\lim_{x\to\infty} TV\{(p_\theta(x,\cdot), v_\theta(x,\cdot)/u_\theta(x,\cdot)) : (g(x),\infty)\} = 0.$$

Proof. Let $\{\Delta_l\}$ be the sequence given as in Theorem 5.4, and let $E_{\Delta_l,\theta}(\Lambda)$ and $Q_{\Delta_l,\theta}(\Lambda)$ be the quantities defined in Lemma 4.8. As in [12], we denote by $dE_{\Delta_l,\theta}$ the measures assigning to $E_{\Delta_l,\theta}(\Lambda)$ to the center of Λ.

The boundedness of $E_{\Delta_l,\theta}(\Lambda)$ in Lemma 4.7 implies the compactness of $\{dE_{\Delta_l,\theta}\}$. Then we can select its subsequence (still denoted by itself) so that, when $\Delta_l \to 0$, the limit

$$dE_{\Delta_l,\theta} \to dE_\theta$$

exists in the weak-star topology in the measure space and is finite on Ω. Therefore, for any $\delta > 0$, we can choose $x_\delta > 0$ independent of $\{U_{\Delta_l,\theta}\}$ and $\{\Delta_l\}$ such that, for any $l > 0$,

$$\sum_{k\geq[x_\delta/\Delta_l]} E_{\lambda,\theta}(\Lambda_{k,n}) < \delta.$$

Moreover, let $X^1_\delta = (x_\delta, y^1_\delta)$ (*or* $X^4_\delta = (x_\delta, y^4_\delta)$) be the point lying in $\chi_{\Delta_l,\theta}$ (or $\partial\Omega_{\Delta_k}$, respectively). Let $\chi^4_{\Delta_l,\theta}$ be the minimum approximate 4-characteristics in $U_{\Delta_l,\theta}$ emanating from the point X^1_δ, and $\chi^1_{\Delta_l,\theta}$ the maximum approximate 1-characteristics in $U_{\Delta_l,\theta}$ emanating from the point X^4_δ. From the construction of approximate solutions, we have

$$|\chi^j_{\Delta_l,\theta}(x+h) - \chi^j_{\Delta_l,\theta}(x)| \leq B(|h| + \Delta_l), \qquad j = 1, 4,$$

for some constant B independent of Δx and θ. Then, for $\theta \in \prod_{k=0}^{\infty}[-1,1]\backslash N$, we can select a subsequence (still denoted by) $\{\Delta_l\}$ such that

$$\chi^j_{\Delta_l,\theta} \to \chi^j_\theta \qquad \text{uniformly on every bounded interval as } \Delta_l \to 0$$

for some $\chi^j_\theta \in Lip$ with $(\chi^j_\theta)'$ bounded.

Let the characteristics $y = \chi^4_\theta(x)$ and $y = \chi^1_\theta(x)$ intersect $\partial\Omega$ and $y = \chi_\theta(x)$, respectively, at $(t^4_\delta, \chi_\theta(t^4_\delta))$ and $(t^1_\delta, \chi_\theta(t^1_\delta))$ for some t^4_δ and t^1_δ. Then, since the velocity ratio v/u and the pressure p are invariant across the vortex sheet, by the approximate conservation laws for the weak 1-waves and 4-waves, we can deduce in the same way as in [12] that

$$TV\{v_{\Delta_l,\theta}(x-,\cdot)/u_{\Delta_l,\theta}(x-,\cdot) \ : \ (g_l(x), \chi_{\Delta_l,\theta}(x))\} < M\,\delta$$

and

$$TV\{p_{\Delta_l,\theta}(x-,\cdot) \ : \ (g_l(x), \chi_{\Delta_l,\theta}(x))\} < M\,\delta$$

for $x > 2(t^1_\delta + t^4_\delta)$, where M is independent of δ, x, $U_{\Delta_l,\theta}$, and Δ_l.

Thus, taking the limit as $\Delta_l \to 0$ and using Theorem 5.4 and the regularity of U_θ yields that, for $x > 2(t^1_\delta + t^4_\delta)$,

$$TV\{v_\theta(x-,\cdot)/u_\theta(x-,\cdot) \ : \ (g(x), \chi_\theta(x))\} < M\,\delta,$$

$$TV\{p_\theta(x-,\cdot)\ :\ (g(x),\chi_\theta(x))\} < M\,\delta.$$

The corresponding estimates above the strong vortex sheet can be obtained similarly. This completes the proof. ∎

THEOREM 5.6.

(i) *Let* $g'_\infty = \lim_{x\to\infty} g'(x+)$. *Then*

$$\lim_{x\to\infty} \sup\{|v_\theta(x,y)/u_\theta(x,y) - g'_\infty|\ :\ y > g(x)\} + \lim_{x\to\infty} |\chi'_\theta(x) - g'_\infty| = 0.$$

(ii) *There exists a constant* p_∞ *such that*

$$\lim_{x\to\infty} \sup\{|p_\theta(x,y) - p_\infty|\ :\ y > g(x)\} = 0.$$

Proof. Let $U_{l,\theta} = U_{\Delta_l,\theta}$, $\sigma_{l,\theta} = \sigma_{\Delta_l,\theta}$, and $\chi_{l,\theta} = \chi_{\Delta_l,\theta}$, where Δ_l is chosen as in the proof of Lemma 5.5. Following the construction of approximate solutions, we conclude that, for every $x > 0$,

$$v_{l,\theta}/u_{l,\theta}|_{\Gamma_k} = (y_{k+1} - y_k)/\Delta x = g'(\eta_k)$$

for some $\eta_k \in (k\Delta x, (k+1)\Delta x)$. Then, choosing x_δ so that $|g'(x+) - g'(\infty)| < \delta$ for $x > x_\delta$, we have

$$\begin{aligned}&\sup\{|v_{l,\theta}(x,y)/u_{l,\theta}(x,y) - g'_\infty|\ :\ g(x) < y < \chi_\theta(x)\}\\ &\le TV\{v_{l,\theta}(x,\cdot)/u_{l,\theta}(x,\cdot)\ :\ (g(x),\chi_\theta(x))\} + M\delta \qquad \text{for } x > 2x_\delta.\end{aligned}$$

Therefore, taking the limit as $\Delta_l \to 0$, by Theorem 5.4 and Lemma 5.5, and by the regularity of U_θ, we can deduce part (i) for $g(x) < y < \chi_\theta(x)$. The case $y > \chi_\theta(x)$ can be proved similarly.

Part (ii) can be obtained similarly by using Lemma 5.5. ∎

6. Isentropic Euler flows over Lipschitz walls. In this section, we study the isentropic Euler equations (1.6) for steady supersonic flows, which can be written in the following conservation form:

$$W(U)_x + H(U)_y = 0, \qquad U = (u, v, \rho), \tag{6.1}$$

with

$$W(U) = (\rho u, \rho u^2 + p, \rho uv), \quad H(U) = (\rho v, \rho uv, \rho v^2 + p).$$

As in section 1, the problem of supersonic Euler flows governed by (6.1) over Lipschitz walls can be formulated as problem (1.9)–(1.10) for system (6.1) in the region above the wall.

DEFINITION 6.1 (entropy solutions). *A BV function* $U = U(x,y)$ *is called an entropy solution of problem* (6.1) *and* (1.9)–(1.10) *provided that*

(i) U *is a weak solution of* (6.1) *and satisfies* (1.9)–(1.10) *in the trace sense;*

(ii) U *satisfies the entropy inequality*

$$(\rho u(E + p(\rho)/\rho))_x + (\rho v(E + p(\rho)/\rho))_y \le 0 \tag{6.2}$$

in the sense of distributions in Ω *including the boundary.*

Remark 6.1. The entropy inequality (6.2) for the steady case directly follows from the energy inequality in the time-dependent isentropic case,

$$(\rho E)_t + (\rho u(E + p(\rho)/\rho))_x + (\rho v(E + p(\rho)/\rho))_y \le 0,$$

under the assumption that the solution (u, v, p, ρ) is independent of time t. Furthermore, by a straightforward calculation, the function $\eta = \rho u(E + p(\rho)/\rho)$ as an entropy function of system (6.1) is a strictly convex function of the conserved variables $W = (w_1, w_2, w_3) = (\rho u, \rho u^2 + p, \rho uv)$ when $u > c$.

6.1. Riemann problems and Riemann solutions. When $u > c$, the eigenvalues of system (6.1) are

$$\lambda_j = \frac{uv + (-1)^{\frac{j+1}{2}} c\sqrt{u^2 + v^2 - c^2}}{u^2 - c^2}, \quad j = 1, 3; \qquad \lambda_2 = v/u, \tag{6.3}$$

where $c^2 = \sqrt{p'(\rho)}$.

When the flow is supersonic (i.e., $u^2 + v^2 > c^2$), system (6.1) is hyperbolic and, when $u > c$, the corresponding eigenvectors are

$$\mathbf{r}_2 = (u, v, 0)^\top, \quad \mathbf{r}_j = \kappa_j(-\lambda_j, 1, \rho(\lambda_j u - v)/c^2)^\top, \ j = 1, 3,$$

where κ_j are chosen so that $\mathbf{r}_j \cdot \nabla\lambda_j = 1$ since the j-characteristic fields are genuinely nonlinear, $j = 1, 3$. Note that the second characteristic field is always linearly degenerate: $\mathbf{r}_2 \cdot \nabla\lambda_2 = 0$.

6.1.1. Wave curves in the phase space. Similarly as in section 2, the vortex sheet curve $C_2(U_0)$ through U_0 with strength σ_2 is

$$C_2(U_0) : U = (u, v, \rho)^\top = (u_0 e^{\sigma_2}, v_0 e^{\sigma_2}, \rho_0)^\top, \tag{6.4}$$

determined by

$$\begin{cases} \frac{dU}{d\sigma_2} = \mathbf{r}_2(U), \\ U|_{\sigma_2 = 0} = U_0, \end{cases}$$

which describes a compressible vortex sheet. Moreover, the jth-rarefaction wave curve $R_j(U_0)$ in the phase space through U_0 is

$$R_j(U_0) : \quad du = -\lambda_j dv, \quad \rho(\lambda_j u - v)dv = dp \qquad \text{for } \rho < \rho_0,\ u > c, \qquad j = 1, 3. \tag{6.5}$$

It is easy to check that $\frac{d\lambda_j}{d\rho}$ along $R_j(U_0), j = 1, 3$, satisfy

$$\frac{d\lambda_1}{d\rho}\bigg|_{R_1(U_0)} < 0, \quad \frac{d\lambda_3}{d\rho}\bigg|_{R_3(U_0)} > 0.$$

Similarly, the Rankine–Hugoniot conditions for (6.1) are

$$s[\rho u] = [\rho v], \tag{6.6}$$

$$s[\rho u^2 + p] = [\rho uv], \tag{6.7}$$

$$s[\rho uv] = [\rho v^2 + p]. \tag{6.8}$$

Then we have

$$s = s_2 = v_0/u_0, \qquad s = s_j = \frac{u_0v_0 + (-1)^{\frac{j+1}{2}}\overline{c}\sqrt{u_0^2 + v_0^2 - \overline{c}^2}}{u_0^2 - \overline{c}^2},\ j = 1, 3,$$

where $\overline{c}^2 = \frac{\rho}{\rho_0}\frac{[p]}{[\rho]}$.

Plugging s_2 into (6.6)–(6.8), we get the same $C_2(U_0)$, as defined in (6.4); while plugging $s_j, j = 1, 3$, into (6.6)–(6.8), we get the jth-shock wave curve $S_j(U_0)$ through U_0:

$$S_j(U_0): \quad [u] = -s_j[v], \quad \rho_0(s_ju_0 - v_0)[v] = [p] \quad \text{for } \rho > \rho_0,\ u > c, \quad j = 1, 3.$$

Notice that $S_j(U_0)$ contacts with $R_j(U_0)$ at U_0 up to second order and

$$\frac{ds_1}{d\rho}\bigg|_{S_1(U_0)} < 0, \qquad \frac{ds_3}{d\rho}\bigg|_{S_3(U_0)} > 0.$$

LEMMA 6.2. *At the state $U_0 = (u_0, 0, \rho_0)$ with $u_0 > c_0$, $\kappa_1(U_0) = \kappa_3(U_0) > 0$, which implies $\kappa_j(U) > 0$ for any state $U \in O_\varepsilon(U_0)$ with small $\varepsilon > 0$, since κ_j are continuous for $j = 1, 3$.*

6.1.2. Riemann problem involving strong vortex sheets. For any $U_b \in O_\varepsilon(U_1)$ and $U_a \in O_\varepsilon(U_2)$, we also use $\{U_b, U_a\} = (0, \sigma_2, 0)$ to denote the strong vortex sheet that connects U_b with U_a with strength σ_2. Then

$$U_a = \Phi_2(\sigma_2; U_b) = (u_be^{\sigma_2}, v_be^{\sigma_2}, \rho_b)^\top.$$

LEMMA 6.3. *The vector function $\Phi_2(\sigma_2; U_b)$ satisfies*

$$\partial_{\sigma_2}\Phi_2(\sigma_2; U_b) = (u_be^{\sigma_2}, v_be^{\sigma_2}, 0)^\top \quad \text{and} \quad \nabla_U\Phi_2(\sigma_2; U_b) = \operatorname{diag}(e^{\sigma_2}, e^{\sigma_2}, 1).$$

Furthermore, we have the following essential lemma for any 2-straight vortex sheet with below state $U_1 = (u_1, 0, \rho_1)$, above state $U_2 = (u_2, 0, \rho_1)$, and strength σ_{20}.

LEMMA 6.4. $\det(\mathbf{r}_3(U_2), \partial_{\sigma_2}\Phi_2(\sigma_{20}; U_1), \nabla_U\Phi_2(\sigma_{20}; U_1)\cdot\mathbf{r}_1(U_1)) > 0$.

This can be seen as follows:

$$\begin{aligned}
&\det(\mathbf{r}_3(U_2), \partial_{\sigma_2}\Phi_2(\sigma_{20}; U_1), \nabla_U\Phi_2(\sigma_{20}; U_1)\cdot\mathbf{r}_1(U_1))\\
&= \kappa_1(U_1)\kappa_3(U_2)\begin{vmatrix} -\lambda_3(U_2) & u_1e^{\sigma_{20}} & e^{\sigma_{20}}(-\lambda_1(U_1))\\ 1 & 0 & e^{\sigma_{20}}\\ \frac{\rho_1u_2\lambda_3(U_2)}{c_2^2} & 0 & \frac{\rho_1u_1\lambda_1(U_1)}{c_1^2}\end{vmatrix}\\
&= \frac{1}{c_1^2}\rho_1u_1^2e^{\sigma_{20}}\kappa_1(U_1)\kappa_3(U_2)(\lambda_3(U_2)e^{2\sigma_{20}} + \lambda_3(U_1)) > 0.
\end{aligned}$$

6.2. Estimates on wave interactions and reflections. Now we make essential estimates as in section 3. The interaction estimates are similar and the corresponding Figures 5–7 are the same except that the 2, 3-vortex sheets and 4-wave in section 3 are now replaced by a 2-vortex sheet and 3-wave, respectively.

6.2.1. Estimates on the weak wave reflections on the boundary. We use the same notation as in section 3.2 for $C_k(a_k, b_k)$ with $a_{k+1} > a_k > 0$, $\omega_{k,k+1}$, ω_k, Ω_k, Γ_k, and the outer normal vector $\mathbf{n}_k$ to Γ_k (cf. Figure 5). Then we consider the initial-boundary value problem with constant state U_a:

$$\begin{cases} (6.1) & \text{in } \Omega_{k+1}, \\ U|_{x=a_k} = U_a, & \\ (u, v) \cdot \mathbf{n}_{k+1} = 0 & \text{on } \Gamma_{k+1}. \end{cases}$$

PROPOSITION 6.5. *Let* $\{U_a, U_m\} = (\alpha_3, \alpha_2, 0)$ *and* $\{U_m, U_k\} = (0, 0, \beta_1)$ *with*

$$(u_k, v_k) \cdot \mathbf{n}_k = 0.$$

Then there exists U_{k+1} *such that*

$$\{U_a, U_{k+1}\} = (0, 0, \delta_3) \qquad \text{and} \qquad (u_{k+1}, v_{k+1}) \cdot \mathbf{n}_{k+1} = 0.$$

Furthermore,

$$\delta_3 = \beta_1 + K_{b3}\alpha_3 + K_{b2}\alpha_2 + K_{b0}\omega_k,$$

where K_{b3}, K_{b2}, *and* K_{b0} *are* C^2*-functions of* $(\alpha_3, \alpha_2, \beta_1, \omega_k; U_a)$ *satisfying*

$$K_{b3}|_{\{\omega_k=\alpha_3=\alpha_2=\beta_1=0, U_a=U_1\}} = 1, \qquad K_{b2}|_{\{\omega_k=\alpha_3=\alpha_2=\beta_1=0, U_a=U_1\}} = 0,$$

and K_{b0} *is bounded.*

6.2.2. Estimates on the interaction between the strong vortex sheet and weak waves. Again, we have to consider the two cases depending on how the vortex sheet and the weak waves interact. See Figures 6 and 7.

The first case is when the weak waves approach the vortex sheet from below. We have the following proposition.

PROPOSITION 6.6. *Let* $U_b, U_m \in O_\epsilon(U_1)$ *and* $U_a \in O_\epsilon(U_2)$ *with*

$$\{U_b, U_m\} = (0, \alpha_2, \alpha_3), \qquad \{U_m, U_a\} = (\beta_1, \sigma_2, 0).$$

Then there exists a unique $(\delta_1, \sigma_2', \delta_3)$ *such that the Riemann problem* (2.18) *admits an admissible solution that consists of a weak* 1*-wave of strength* δ_1*, a strong vortex sheet of strength* σ_2'*, and a weak* 3*-wave of strength* δ_3*, that is,*

$$\{U_b, U_a\} = (\delta_1, \sigma_2', \delta_3).$$

Moreover,

$$\begin{aligned} &\delta_1 = \beta_1 + K_{11}\alpha_3 + O(1)|\beta_1||\alpha_2|, \qquad \delta_3 = K_{13}\alpha_3 + O(1)|\beta_1||\alpha_2|, \\ &\sigma_2' = \sigma_2 + \alpha_2 + K_{12}\alpha_3 + O(1)|\beta_1||\alpha_2|, \end{aligned}$$

where

$$|K_{11}|_{\{\alpha_3=\alpha_2=0, \sigma_2'=\sigma_{20}\}} = \left|\frac{\lambda_3(U_2)e^{\sigma_{20}} - \lambda_3(U_1)}{\lambda_3(U_2)e^{\sigma_{20}} + \lambda_3(U_1)}\right| < 1, \quad K_{12} \text{ and } K_{13} \text{ are bounded.}$$

We need to find the solution $(\delta_1, \sigma_2', \delta_3)$ as a function of $(\alpha_2, \alpha_3, \beta_1, \sigma_2; U_b)$ to

$$\Phi_2(\sigma_2; \Phi_1(\beta_1; \Phi_3(\alpha_3, \alpha_2, 0; U_b))) = \Phi_3(\delta_3; \Phi_2(\sigma_2'; \Phi_1(\delta_1; U_b))). \tag{6.9}$$

Lemma 6.4 implies

$$\det\left(\frac{\partial\Phi_3(\delta_3,\Phi_2(\sigma_2';\Phi_1(\delta_1,U_b)))}{\partial(\delta_3,\sigma_2',\delta_1)}\right)\Bigg|_{\{\delta_1=0,\sigma_2'=\sigma_{20}\}}$$
$$=\det(\mathbf{r}_3(U_2),\partial_{\sigma_2}\Phi_2(\sigma_{20};U_1),\nabla_U\Phi_2(\sigma_{20};U_1)\cdot\mathbf{r}_1(U_0))>0.$$

Therefore, $(\delta_3,\sigma_2',\delta_1)$ can be solved as a C^2-function of $(\alpha_2,\alpha_3,\beta_1,\sigma_2;U_b)$ uniquely, i.e., $\delta_i=\delta_i(\alpha_2,\alpha_3,\beta_1,\sigma_2;U_b)$, $i=1,3$, and $\sigma_2'=\sigma_2'(\alpha_2,\alpha_3,\beta_1,\sigma_2;U_b)$. Then we can follow the proof of Proposition 3.3 to obtain our desired result.

The second case is when the weak waves approach the vortex sheet from above. By symmetry, we can easily obtain the following.

PROPOSITION 6.7. *Let $U_b\in O_\epsilon(U_1)$ and $U_a,U_m\in O_\epsilon(U_2)$ with*

$$\{U_b,U_m\}=(0,\sigma_2,\alpha_3),\qquad \{U_m,U_a\}=(\beta_1,\beta_2,0).$$

Then there exists a unique $(\delta_1,\sigma_2',\delta_3)$ such that the Riemann problem (2.18) *admits an admissible solution that consists of a weak* 1*-wave of strength δ_1, a vortex sheet of strength σ_2', and a weak* 3*-wave of strength δ_3, that is,*

$$\{U_b,U_a\}=(\delta_1,\sigma_2',\delta_3).$$

Moreover,

$$\delta_1=K_{21}\beta_1+O(1)|\alpha_3||\beta_2|,\quad \delta_3=\alpha_3+K_{23}\beta_1+O(1)|\alpha_3||\beta_2|,$$
$$\sigma_2'=\sigma_2+\beta_2+K_{22}\beta_1+O(1)|\alpha_3||\beta_2|,$$

where $\sum_{j=1}^3|K_{2j}|$ is bounded.

6.3. Approximate solutions. Similarly to section 4, we can construct globally defined, modified Glimm approximate solutions $U_{\Delta x,\theta}$ in the approximate domains (see Figure 8). We need to estimate $U_{\Delta x,\theta}$ on a class of space-like curves, i.e., j-mesh curves J as introduced in Definition 4.2. To achieve this, we now define the Glimm-type functional.

DEFINITION 6.8. *We define*

$$F_s(J)=C^*|\sigma_2^J-\sigma_{20}|+F(J)$$

with

$$F(J)=L(J)+KQ(J),$$
$$L(J)=L^1(J)+L^2(J),$$
$$L^1(J)=K_1^*L_0(J)+K_{11}^*L_1^1(J)+K_{12}^*L_2^1(J)+L_3^1(J),$$
$$L^2(J)=K_{21}^*L_1^2(J)+K_{22}^*L_2^2(J)+K_{23}^*L_3^2(J),$$
$$Q(J)=\sum\{|\alpha_i||\beta_j|\ :\ \textit{both } \alpha_i \textit{ and } \beta_j \textit{ cross } J \textit{ and approach}\},$$

and $L_0(J)=\sum\{|\omega(C_k)|\ :\ C_k\in\Omega_J\}$,

$$L_j^i(J)=\sum\{|\alpha_j|\ :\ \alpha_j \textit{ crosses } J \textit{ in region (i)}\},\quad i=1,2;\ j=1,2,3,$$

where K and C^* are chosen later, while Ω_J is the set of the corner points C_k lying in J^+, that is,

$$\Omega_J=\{C_k\in J^+\cap\partial\Omega_{\Delta x}\ :\ k\ge 0\},$$

σ_2^J stands for the strength of the strong vortex sheet crossing J, K_1^*, K_{11}^*, and K_{12}^* are the constants associated with region (1)*, K_{21}^*, K_{22}^*, and K_{23}^* are the constants associated with region* (2)*, which satisfy the conditions*

$$K_1^* > |K_{b0}|, \qquad K_{1i}^* > |K_{bi}|, \quad i = 1, 2,$$

and

$$K_{23}^* < \frac{1 - |K_{11}|K_{11}^*}{|K_{13}|}, \quad K_{21}^* > |K_{21}|K_{11}^* + |K_{23}|K_{23}^*,$$

while K_{12}^* and K_{22}^* are arbitrarily large positive constants. These conditions can be achieved from our discussions of the properties of K_{bj} and K_{sj}, $j = 1, 2$, as in Propositions 6.5–6.7 *in section* 6.2*.*

The rest can be done in the same way as in the previous sections by choosing suitable K and C^*, so we omit the details here.

Acknowledgment. The second author thanks the Department of Mathematics of Northwestern University for its hospitality during his visit.

REFERENCES

[1] M. Artola and A. Majda, *Nonlinear development of instabilities in supersonic vortex sheets* I: *The basic kink modes*, Phys. D, 28 (1987), pp. 253–281.
[2] M. Artola and A. Majda, *Nonlinear development of instabilities in supersonic vortex sheets* II: *Resonant interaction among kink modes*, SIAM J. Appl. Math., 49 (1989), pp. 1310–1349.
[3] M. Artola and A. Majda, *Nonlinear kink modes for supersonic vortex sheets*, Phys. Fluids A, 1 (1989), pp. 583–596.
[4] T. Chang and L. Hsiao, *The Riemann Problem and Interaction of Waves in Gas Dynamics*, Longman Scientific & Technical, Essex, UK, 1989.
[5] G.-Q. Chen, *Overtaking of shocks of the same kind in the isentropic steady supersonic plane flow*, Acta Math. Sci. (English Ed.), 7 (1987), pp. 311–327.
[6] G.-Q. Chen and D. Wagner, *Global entropy solutions to exothermically reacting, compressible Euler equations*, J. Differential Equations, 191 (2003), pp. 277–322.
[7] G.-Q. Chen, Y. Q. Zhang, and D. Zhu, *Existence and stability of supersonic Euler flows past Lipschitz wedges*, Arch. Rational Mech. Anal., 181 (2006), pp. 261–310.
[8] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Wiley Interscience, New York, 1948.
[9] A. Corli and M. Sablé-Tougeron, *Stability of contact discontinuities under perturbations of bounded variation*, Rend. Sem. Mat. Univ. Padova, 97 (1997), pp. 35–60.
[10] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, 2000.
[11] J. Glimm, *Solution in the large for nonlinear systems of conservation laws*, Comm. Pure Appl. Math., 18 (1965), pp. 695–715.
[12] J. Glimm and P. D. Lax, *Decay of Solutions of Systems of Hyperbolic Conservation Laws*, Mem. Amer. Math. Soc., 101 (1970).
[13] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[14] W. Lien and T. P. Liu, *Nonlinear stability of a self-similar* 3*-dimensional gas flow*, Commun. Math. Phys., 204 (1999), pp. 525–549.
[15] T. P. Liu, *The deterministic version of the Glimm scheme*, Commun. Math. Phys., 57 (1977), pp. 135–148.
[16] T. P. Liu, *Large-time behavior of initial and initial-boundary value problems of a general system of hyperbolic conservation laws*, Commun. Math. Phys., 55 (1977), pp. 163–177.
[17] J. W. Miles, *On the disturbed motion of a plane vortex sheet*, J. Fluid Mech., 4 (1958), pp. 538–552.
[18] M. Sablé-Tougeron, *Méthode de Glimm et problème mixte*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 423–443.

[19] J. Smoller, *Shock Waves and Reaction-Diffusion Equations,* Springer-Verlag, New York, 1983.
[20] B. Temple, *Solutions in the large for the nonlinear hyperbolic conservation laws of gas dynamics*, J. Differential Equations, 41 (1981), pp. 96–161.
[21] B. Temple and R. Young, *The large time existence of periodic solutions for the compressible Euler equations*, Mat. Contemp., 11 (1996), pp. 171–190.
[22] B. Temple and R. Young, *The large time stability of sound waves*, Commun. Math. Phys., 179 (1996), pp. 417–466.
[23] A. I. Volpert, *The space BV and quasilinear equations*, Mat. Sb. (N.S.), 73 (1967), pp. 255–302 (in Russian); Math. USSR Sb., 2 (1967), pp. 225–267 (in English).
[24] P. R. Woodward, *Simulation of the Kelvin–Helmholtz instability of a supersonic slip surface with the piecewise-parabolic method (PPM)*, in Numerical Methods for the Euler Equations of Fluid Dynamics (Rocquencourt, 1983), SIAM, Philadelphia, 1985, pp. 493–508.
[25] Y. Zhang, *Global existence of steady supersonic potential flow past a curved wedge with piecewise smooth boundary*, SIAM J. Math. Anal., 31 (1999), pp. 166–183.
[26] Y. Zhang, *Steady supersonic flow past an almost straight wedge with large vertex angle*, J. Differential Equations, 192 (2003), pp. 1–46.

SIAM J. MATH. ANAL.
Vol. 38, No. 5, pp. 1694–1713

GLOBAL EXISTENCE OF CLASSICAL SOLUTIONS FOR A HAPTOTAXIS MODEL*

CHRISTOPH WALKER† AND GLENN F. WEBB†

Abstract. A system of nonlinear partial differential equations modeling haptotaxis is investigated. The model arises in cell migration processes involved in tumor invasion. The existence of unique global classical solutions is proved.

Key words. haptotaxis, diffusion, global existence, uniqueness, classical solutions

AMS subject classifications. 92C17, 92D25, 35G25

DOI. 10.1137/060655122

1. Introduction.

Models of complex dynamic biological processes frequently involve systems of nonlinear partial differential equations for production, growth, decay, interaction, and spatial movement. For models that include spatial movement the equations typically contain both diffusion and taxis terms. Our goal in this paper is to examine the issues of global existence and uniqueness for a model involving spatial movement and haptotaxis. The term *haptotaxis* originated with S. B. Carter in 1965: "... the movement of a cell is controlled by the relative strengths of its peripheral adhesions, and that movements directed in this way, together with the influence of patterns of adhesion on cell shape are responsible for the arrangement of cells into complex and ordered tissues" [8]. Cell movement in morphogenesis, inflammation, wound healing, tumor invasion, and other migrations are the result of haptotactic responses of cells to differential adhesion strengths [8, 9].

The haptotaxis model we investigate here is a simplified version of a model proposed by Anderson [5] in 2005 to describe tumor invasion into surrounding tissue (see also [6]). The model involves four key components of the process: tumor cells, surrounding tissue macromolecules, degradative enzymes, and oxygen. The model in [5] hybridizes continuum partial differential equations and cellular automata formulations to incorporate cell cycle elements, and a similar model in [7] uses continuum cell age structure for the same purpose. Both of these investigations model other features of tumor invasion, including the role of quiescent cells and the evolution of mutated cell lines of increasingly invasive aggressiveness. Our objective is to investigate the simplified system of four nonlinear partial differential equations which underlie the models in [5] and [7]. The mathematical formulation of haptotaxis is similar to that of more familiar chemotaxis processes for which we refer to the survey article [16] and the extensive list of references therein. Haptotaxis in tumor growth, however, possesses unique features in that the movement of tumor cells is directed to the bound (i.e., nondiffusible) extracellular environment, which supplies essential oxygen and available space, as it is degraded by the tumor-produced degradative enzyme. The mathematical difficulty in treating haptotaxis in this context is that the

*Received by the editors March 24, 2006; accepted for publication (in revised form) August 21, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/sima/38-5/65512.html
†Department of Mathematics, Vanderbilt University, 1326 Stevenson Center, Nashville, TN 37240 (christoph.walker@vanderbilt.edu, glenn.f.webb@vanderbilt.edu). The work of the second author was supported by PHS-NIH grant 1P50CA113007-01.

haptotaxis term is nonlinearly dependent on the tumor cells through the diffusion of the degradative enzyme produced by these cells.

We make the following assumptions: The tumor is contained in a region of tissue Ω. The dependent variables of the model are as follows: $p(x,t)$ is the density of tumor cells at $x \in \Omega$ at time t, $m(x,t)$ is the concentration of matrix degradative enzyme (MDE) at $x \in \Omega$ at time t, $f(x,t)$ is the density of extracellular matrix macromolecules at $x \in \Omega$ at time t, and $w(x,t)$ is the concentration of oxygen at $x \in \Omega$ at time t. The equations of the model are as follows:

$$(H_1) \qquad \partial_t f = -\underbrace{a(x)\, m\, f}_{\text{degradation}},$$

$$(H_2) \qquad \partial_t m = \underbrace{\alpha\, \Delta m}_{\text{diffusion}} + \underbrace{d(x)\, p}_{\text{production}} - \underbrace{b(x)\, m}_{\text{decay}},$$

$$(H_3) \qquad \partial_t p = \underbrace{\beta\, \Delta p}_{\text{cell motility}} - \underbrace{\nabla \cdot (p\, \chi(f)\, \nabla f)}_{\text{haptotaxis}} - \underbrace{\theta(x,w)\, p}_{\text{cell death}} + \underbrace{\varrho(x,w)\, p}_{\text{cell division}},$$

$$(H_4) \qquad \partial_t w = \underbrace{\gamma\, \Delta w}_{\text{diffusion}} + \underbrace{g(x)\, f}_{\text{production}} - \underbrace{\omega(x,p)\, w}_{\text{uptake}} - \underbrace{e(x)\, w}_{\text{decay}}$$

for $(t,x) \in (0,\infty) \times \Omega$ supplemented with Neumann boundary conditions

$$(H_5) \qquad \partial_\nu m = \partial_\nu p - p\chi(f)\partial_\nu f = \partial_\nu w = 0 \quad \text{on } \partial\Omega$$

and initial conditions

$$(H_6) \qquad f(0) = f^0, \quad m(0) = m^0, \quad p(0) = p^0, \quad w(0) = w^0.$$

It seems that (H_1)–(H_6) have not been considered analytically thus far, but rather related equations of the form

$$\partial_t f = h(p,f), \tag{1}$$

$$\partial_t p = \beta\, \Delta p - \nabla \cdot \big(p\, \chi(f)\, \nabla f\big) \tag{2}$$

have attracted attention, in particular the case

$$h(p,f) = \sigma\, p\, f^r \quad \text{with} \quad \sigma = \pm 1 \quad \text{and} \quad r > 0. \tag{3}$$

We refer to [13] for the case of general functions h satisfying suitable hypotheses. As for (1), (2) with h of the form (3), we refer to [10, 11, 12, 17, 19, 20, 21, 23], where existence of solutions and phenomena such as blowup or stability of steady states are investigated depending on the sign of σ, on r, and on the sensitivity χ.

We point out that our model differs from (1), (2), (3) with $\sigma = -1$, in that the ordinary differential equation in (H_1) is coupled to (H_3) via the "intermediate" equation (H_2).

Solving (H_3) or (2) classically requires that f have second order derivatives (with respect to x) in some L_q-space. In our model the regularity of f is determined by m for which one has a smoothing effect due to (H_2). This induces the regularity that allows us to derive an L_q-bound on p, which is sufficient to deduce global existence for $n \le 3$ and without smallness assumptions on the initial data. As for (1)–(3), the

regularity of f is determined by p. Thus, the second order derivatives of p should also be in L_q. Local existence and uniqueness of "smooth" solutions for (1)–(3) can be obtained using maximal regularity for the p-equation (2). However, global existence then requires estimates on p that are stronger than L_q-estimates and which are far from obvious.

In order to state our main result regarding the solvability of (H_1)–(H_6) we assume that Ω is a bounded and smooth domain in $\mathbb{R}^n$, $n \le 3$, and that the diffusion coefficients α, β, and γ are positive constants. Concerning the data in (H_1)–(H_4) we assume throughout that there exists some $s > 0$ such that

$$\begin{cases} a \in W^2_\infty(\Omega) \ , \quad \partial_\nu a = 0 \text{ on } \partial\Omega \ , \\ b \, , \, d \in C^s(\bar\Omega) \ , \quad g \, , \, e \in L_\infty(\Omega) \ , \end{cases} \tag{4}$$

and that all functions are nonnegative. We also assume that

$$\chi \in C^1(\mathbb{R}^+) \ , \qquad \chi \ge 0 \ , \qquad \chi \text{ and } \chi' \text{ are globally Lipschitz continuous.} \tag{5}$$

Furthermore, regarding ϱ, θ, $\omega \in C(\bar\Omega \times \mathbb{R}, \mathbb{R}^+)$ we suppose that, for some $c_0 > 0$,

$$|\phi(x,\eta) - \phi(x,\bar\eta)| \le c_0 \, |\eta - \bar\eta| \ , \quad x \in \Omega \ , \quad \eta \, , \, \bar\eta \in \mathbb{R} \ , \quad \phi \in \{\varrho, \theta, \omega\} \ . \tag{6}$$

Note that this implies, for some $c > 0$,

$$|\phi(x,\eta)| \le c \, (1 + |\eta|) \ , \quad x \in \Omega \ , \quad \eta \in \mathbb{R} \ , \quad \phi \in \{\varrho, \theta, \omega\} \ . \tag{7}$$

To simplify the notation we put $\vartheta := \varrho - \theta$.

For brevity of notation we set $L_q := L_q(\Omega)$ and $W^\tau_q := W^\tau_q(\Omega)$ for $1 \le q \le \infty$ and $\tau \ge 0$. Moreover, we denote by $W^\tau_{q,\mathcal{B}} := W^\tau_{q,\mathcal{B}}(\Omega)$ the Sobolev–Slobodeckii spaces including the Neumann boundary conditions, that is,

$$W^\tau_{q,\mathcal{B}} := \begin{cases} \left\{ u \in W^\tau_q \, ; \, \partial_\nu u = 0 \right\} \ , & \tau > 1 + 1/q \ , \\ W^\tau_q \ , & 0 \le \tau < 1 + 1/q \ . \end{cases}$$

If $J \subset \mathbb{R}^+$ is an interval containing 0, we set $\dot J := J \setminus \{0\}$.

We shall prove then the following result.

THEOREM 1.1. *Let assumptions* (4)–(6) *be satisfied, and let* $(1 \vee n/2) < q < \infty$ *and* $2\delta \in (0,2) \setminus \{1 + 1/q\}$. *Given any nonnegative initial value*

$$(f^0, m^0, p^0, w^0) \in W^2_{q,\mathcal{B}} \times W^{2\delta}_{q,\mathcal{B}} \times L_q \times L_q$$

there exists a global nonnegative solution (f, m, p, w) *to* (H_1)–(H_6) *such that*

$$\begin{aligned} & f \in C(\mathbb{R}^+, W^2_{q,\mathcal{B}}) \cap C^1(\dot{\mathbb{R}}^+, W^2_{q,\mathcal{B}}) \ , \\ & m \in C(\mathbb{R}^+, W^{2\delta}_{q,\mathcal{B}}) \cap C(\dot{\mathbb{R}}^+, W^2_{q,\mathcal{B}}) \cap C^1(\dot{\mathbb{R}}^+, L_q) \ , \\ & p \in C(\mathbb{R}^+, L_q) \cap C(\dot{\mathbb{R}}^+, W^2_{q,\mathcal{B}}) \cap C^1(\dot{\mathbb{R}}^+, L_q) \ , \\ & w \in C(\mathbb{R}^+, L_q) \cap C(\dot{\mathbb{R}}^+, W^2_{q,\mathcal{B}}) \cap C^1(\dot{\mathbb{R}}^+, L_q) \ . \end{aligned}$$

This solution satisfies

$$t^\eta \, \|p(t)\|_{W^{2\eta}_q} \to 0 \quad \text{and} \quad t^\lambda \, \|m(t)\|_{W^2_q} \to 0 \quad \text{as} \quad t \to 0^+ \tag{8}$$

for all (η, λ) *such that*

$$n/q < 2\eta < 2\ , \quad 2\eta \geq 1\ , \quad (1-\delta) \vee \eta \leq \lambda < 1\ , \tag{9}$$

and it is the only solution satisfying (8) *for some* (η, λ) *as in* (9).

Remarks 1.2. (a) Except for (H_1), which lacks a smoothing effect due to diffusion, the regularity assumptions on the initial values and the restriction on the integrability index q seem to be fairly weak. In particular, we do not impose bounded initial values or assume that $q > n$, and also the Sobolev regularity on m^0 can be arbitrary low.

(b) The no-flux boundary condition on p in (H_5) is correct from a modeling point of view, since neither diffusion nor haptotaxis should change the tumor mass. Notice that it reduces to a Neumann boundary condition $\partial_\nu p(t) = 0$ provided that $\partial_\nu f(t) = 0$. The latter is guaranteed due to the imposed Neumann boundary conditions on f^0 and a. Thus, these assumptions decouple p and f on the boundary.

(c) The solution depends continuously on the initial value in the sense stated in Proposition 3.1.

(d) The local existence and uniqueness statement of the theorem above is also true for space dimensions $n > 3$ as it follows from the proof given below.

We state the following simplified version of the above theorem for the particular case $q > n$.

COROLLARY 1.3. *Let* a, b, d, e, g *be nonnegative constants and suppose* (5), (6). *If* $q > n$, *then problem* (H_1)–(H_6) *has, for any nonnegative initial value*

$$(f^0, m^0, p^0, w^0) \in X := W_q^2 \times W_q^1 \times W_q^1 \times L_q$$

such that $\partial_\nu f^0 = 0$, *a unique global nonnegative classical solution* (f, m, p, w) *in the space* $C(\mathbb{R}^+, X)$.

A proof of Corollary 1.3 could be obtained by applying the general semigroup theory for semilinear parabolic problems. However, we shall point out that Theorem 1.1 actually ensures existence and uniqueness of classical solutions under considerably weaker assumptions on the integrability index q but also on the regularity of the initial values p^0 and m^0. Also note that any classical solution belonging to $C(\mathbb{R}^+, X)$ satisfies (8) for some (η, λ) as in (9). In this sense, the uniqueness (and existence) result stated in Theorem 1.1 is more general than in Corollary 1.3.

The outline of this paper is as follows: In section 2 we collect some auxiliary results which are used in the proof of local existence and uniqueness of solutions in section 3. Section 4 is devoted to positivity of solutions, and in section 5 we prove global existence. In section 6 some numerical examples are given in order to illustrate the role of haptotaxis in spatial movement. In section 7 we summarize our results.

2. Preliminaries. In what follows, we denote for $1 < q < \infty$ by $\Delta := \Delta_q$ the Laplace operator defined on $W^2_{q,\mathcal{B}}$ and observe that it generates a positive, strongly continuous analytic semigroup $\{e^{t\Delta}\,;\, t \geq 0\}$ of contractions on L_q [1, 22]. Moreover, we will use the inequality

$$\|e^{t\Delta}\|_{\mathcal{L}(W^{2\sigma}_{q,\mathcal{B}}, W^{2\tau}_{q,\mathcal{B}})} \leq c(T)\, t^{\sigma-\tau}\ , \quad 0 < t \leq T\ , \tag{10}$$

which is true provided that $0 \leq 2\sigma \leq 2\tau \leq 2$ with $2\sigma, 2\tau \neq 1 + 1/q$, where $c(T)$ depends on the involved parameters. We also use the inequality

$$\|e^{t\Delta}\|_{\mathcal{L}(L_q, L_p)} \leq\ c(T)\, t^{-(1/q-1/p)n/2}\ , \quad t \in (0, T]\ , \tag{11}$$

for $1 < q \leq p \leq \infty$. Given $\xi > 0$ we then put $U_\xi(t) := e^{t\xi\Delta}$. Furthermore, for any measurable function $u : \dot{J} \to L_q$ we set

$$U_\xi \star u(t) := \int_0^t U_\xi(t-s)\, u(s)\ \mathrm{d}s\ , \quad t \in \dot{J}\ ,$$

whenever these integrals exist. If E is a Banach space and $\mu \in \mathbb{R}$, we denote by $BC_\mu(\dot{J}, E)$ the Banach space of all functions $u : \dot{J} \to E$ such that $\big(t \mapsto t^\mu u(t)\big)$ is bounded and continuous from $\dot{J}$ into E, equipped with the norm

$$u \mapsto \|u\|_{C_\mu(\dot{J},E)} := \sup_{t\in\dot{J}} t^\mu \, \|u(t)\|_E\ .$$

We write $C_\mu(\dot{J}, E)$ for the closed linear subspace consisting of all u satisfying $t^\mu u(t) \to 0$ in E as $t \to 0^+$. Note that $C_\nu((0,T], E) \hookrightarrow C_\mu((0,T], E)$ for $\nu \leq \mu$ and $T > 0$.

For later use we state the following auxiliary result on pointwise multiplication.

LEMMA 2.1. *Suppose that $n/q < 2\eta$ with $2\eta \geq 1$ and let $0 < 2r < (s \wedge 2\eta)$. Then pointwise multiplication is a continuous mapping*

(i) $W_q^{2\eta} \times L_q \to L_q$,
(ii) $W_q^{2\eta-1} \times W_q^1 \to L_q$,
(iii) $C^s(\bar{\Omega}) \times W_q^{2\eta} \to W_q^{2r}$.

Proof. (i) follows from the embedding $W_q^{2\eta} \hookrightarrow L_\infty$, while statements (ii) and (iii) are easy consequences of [2, Thm. 4.1]. □

Evidently, given suitable functions $f^0 = f^0(x)$ and $m = m(x,t)$, the solution to (H_1) is

$$F_1(m) := F_1[f^0](m) := \left[t \mapsto \exp\left(-\int_0^t am(s)\mathrm{d}s\right) f^0\right]\ .$$

Note then that the gradient and the Laplacian take the form

$$\nabla F_1(m)(t) = \exp\left(-\int_0^t am(s)\mathrm{d}s\right)\left[\nabla f^0 - \int_0^t \nabla(am)(s)\mathrm{d}s\, f^0\right] \tag{12}$$

and

$$\begin{aligned}\Delta F_1(m)(t) = \exp\left(-\int_0^t am(s)\mathrm{d}s\right)\Big[&\Delta f^0 - \int_0^t \Delta(am)(s)\mathrm{d}s\, f^0 \\ &+ \Big|\int_0^t \nabla(am)(s)\mathrm{d}s\Big|^2 f^0 - 2\int_0^t \nabla(am)(s)\mathrm{d}s\cdot\nabla f^0\Big]\ .\end{aligned} \tag{13}$$

In particular, (12) warrants that $\partial_\nu F_1(m)(t) = 0$ provided $\partial_\nu f^0 = \partial_\nu m(t) = 0$ for all t since $a \in W^2_{\infty,\mathcal{B}}$. Furthermore, F_1 has the following properties.

LEMMA 2.2. *For $0 < T \leq T_0$ put $I := [0,T]$. If $(1 \vee n/2) < q < \infty$ and $f^0 \in W^2_{q,\mathcal{B}}$, there holds*

(i) *$F_1(m) \in C^1(I, W^2_{q,\mathcal{B}})$ for $m \in C(I, W^2_{q,\mathcal{B}})$;*
(ii) *$F_1(m) \in C(I, W^2_{q,\mathcal{B}})$ for $m \in C_\mu(\dot{I}, W^2_{q,\mathcal{B}})$ and $\mu < 1$, and for $R_0 > 0$ there exists a constant $k := k(T_0, R_0) > 0$ such that*

$$\|F_1(m) - F_1(\bar{m})\|_{C(I,W^2_{q,\mathcal{B}})} \leq k\, \|f^0\|_{W_q^2}\, T^{1-\mu}\, \|m - \bar{m}\|_{C_\mu(\dot{I},W^2_{q,\mathcal{B}})}$$

provided $\|m\|_{C_\mu(\dot{I},W^2_{q,\mathcal{B}})}$, $\|\bar{m}\|_{C_\mu(\dot{I},W^2_{q,\mathcal{B}})} \leq R_0$.

Proof. (i) Fix $m \in C(I, W^2_{q,\mathcal{B}}) \hookrightarrow C(I, L_\infty)$ and temporarily set $F_1 := F_1(m)$. Owing to Lemma 2.1, (4), and (13) we deduce $(1+\Delta)F_1 \in C(I, L_q)$, from which it follows that $F_1 \in C(I, W^2_{q,\mathcal{B}})$ since $\partial_\nu F_1 = 0$. Clearly, this implies $F_1 \in C^1(I, W^2_{q,\mathcal{B}})$ owing to $\partial_t F_1 = -amF_1$ and the fact that pointwise multiplication is a continuous mapping from $W^2_{q,\mathcal{B}} \times W^2_{q,\mathcal{B}}$ into $W^2_{q,\mathcal{B}}$.

(ii) Given $m, \bar m \in C_\mu(\dot I, W^2_{q,\mathcal{B}})$ with norm less than $R_0 > 0$, we have

$$\int_0^t \|m(s)\|_\infty \,\mathrm{d}s \,\le\, c \int_0^t \|m(s)\|_{W^2_q} \,\mathrm{d}s \,\le\, c(R_0)\, t^{1-\mu} \,, \quad t \in I \,.$$

This yields for $0 \le t \le T$

$$\begin{aligned} \|F_1(m)(t) - F_1(\bar m)(t)\|_{L_q} \,&\le\, c(T_0, R_0) \int_0^t \|m(s) - \bar m(s)\|_\infty \,\mathrm{d}s\, \|f^0\|_{L_q} \\ &\le\, c(T_0, R_0)\, \|f^0\|_{L_q}\, T^{1-\mu}\, \|m - \bar m\|_{C_\mu(\dot I, W^2_{q,\mathcal{B}})} \,. \end{aligned}$$

Similarly, Lemma 2.1 and (13) entail

$$\|\Delta F_1(m)(t) - \Delta F_1(\bar m)(t)\|_{L_q} \,\le c(T_0, R_0)\, \|f^0\|_{W^2_q}\, T^{1-\mu}\, \|m - \bar m\|_{C_\mu(\dot I, W^2_{q,\mathcal{B}})}$$

for $0 \le t \le T$, and the assertion follows. □

LEMMA 2.3. *Let $1 < q < \infty$, $2\sigma \in (0,2) \setminus \{1 + 1/q\}$ and $T, \xi > 0$. Then*

(i) *$U_\xi u := [t \mapsto U_\xi(t)u] \in C_\sigma\big((0,T], W^{2\sigma}_{q,\mathcal{B}}\big)$ for $u \in L_q$;*

(ii) *$U_\xi u = [t \mapsto U_\xi(t)u] \in C_{1-\sigma}\big((0,T], W^2_{q,\mathcal{B}}\big)$ for $u \in W^{2\sigma}_{q,\mathcal{B}}$.*

Proof. The proof of [4, Prop. 6] is easily adapted to the case (i). In much the same way one shows (ii). □

3. Local existence and uniqueness. In the following we use the abbreviations

$$\begin{aligned} S(m,p) \,&:=\, d\,p \,-\, b\,m \,, \\ Q(f,p,w) \,&:=\, -\nabla \cdot \big(p\,\chi(f)\,\nabla f\big) \,+\, \vartheta(w)\,p \,, \\ R(f,p,w) \,&:=\, -\,e\,w \,-\, \omega(p)\,w \,+\, g\,f \,. \end{aligned}$$

Here and below we denote by $\vartheta(w)$ and $\omega(p)$ the Nemitskii operators of $\vartheta(\cdot, w)$ and $\omega(\cdot, p)$, respectively; that is, we set $\phi(u) := [x \mapsto \phi(x, u(x))]$ for $\phi \in \{\vartheta, \omega\}$ and $u : \Omega \to \mathbb{R}$.

The proof of the existence and uniqueness statement of Theorem 1.1 is based on the next result.

PROPOSITION 3.1. *Let $1 < q < \infty$ and $n/q < 2\eta \le 2\xi \le 2\mu < 2$ with $2\eta \ge 1$. Given $B \ge 1$ there exists $T := T(B) > 0$ such that, for any*

$$u^0 := \big(f^0, m^0, p^0, w^0\big) \in E := W^2_{q,\mathcal{B}} \times W^{2(1-\mu)}_{q,\mathcal{B}} \times L_q \times L_q$$

with $\|u^0\|_E \le B$, the problem

$$\text{(M)} \qquad \begin{cases} f(t) = \exp\big(-\int_0^t am(s)\mathrm{d}s\big) f^0 \,, & t \in I \,, \\ m(t) = U_\alpha(t)\, m^0 \,+\, U_\alpha \star S(m,p)(t) \,, & t \in I \,, \\ p(t) = U_\beta(t)\, p^0 \,+\, U_\beta \star Q(f,p,w)(t) \,, & t \in I \,, \\ w(t) = U_\gamma(t)\, w^0 \,+\, U_\gamma \star R(f,p,w)(t) \,, & t \in I \,, \end{cases}$$

has a unique solution

$$u := (f,m,p,w) \in \mathcal{V}_T := C(I,W^2_{q,\mathcal{B}}) \times C_\mu(\dot I,W^2_{q,\mathcal{B}}) \times C_\xi(\dot I,W^{2\eta}_{q,\mathcal{B}}) \times C(I,L_q)\ ,$$

where $I := [0,T]$. Moreover, the solution depends continuously on the initial value in the sense that if $\bar u \in \mathcal{V}_T$ denotes the solution corresponding to $\bar u^0 \in E$ with $\|\bar u^0\|_E \le B$, then $\bar u \to u$ in $\mathcal{V}_T$ as $\bar u^0 \to u^0$ in E.

Proof. Given $T \in (0,1]$ we put

$$W_T := C\big([0,T],W^2_{q,\mathcal{B}}\big)\ ,\quad X_T := C_\mu\big((0,T],W^2_{q,\mathcal{B}}\big)\ ,$$
$$Y_T := C_\xi\big((0,T],W^{2\eta}_{q,\mathcal{B}}\big)\ ,\quad Z_T := C\big([0,T],L_q\big)\ ,$$

so that $\mathcal{V}_T = W_T \times X_T \times Y_T \times Z_T$. For $u^0 = \big(f^0,m^0,p^0,w^0\big) \in E$ it then follows from Lemma 2.3 that

$$V^0 := \big(f^0,U_\alpha m^0,U_\beta p^0,U_\gamma w^0\big) \in \mathcal{V}_T\ .$$

Defining

$$F_2(m,p) := U_\alpha m^0 + U_\alpha \star S(m,p)\ ,$$
$$F_3(f,p,w) := U_\beta p^0 + U_\beta \star Q(f,p,w)\ ,$$
$$F_4(f,p,w) := U_\gamma w^0 + U_\gamma \star R(f,p,w)\ ,$$

and

$$F(u) := F(f,m,p,w) := \big(F_1(m),F_2(m,p),F_3(f,p,w),F_4(f,p,w)\big)\ ,$$

problem (M) can be recast as a fixed point problem of the form $F(u) = u \in \mathcal{V}_T$. In order to solve this problem, we first recall that Lemma 2.2(ii) implies that there exists for any given $R_0 > 0$ a constant $c(R_0) > 0$ with

$$\|F_1(m) - F_1(\bar m)\|_{W_T} \le c(R_0)\,T^{1-\mu}\,\|m-\bar m\|_{X_T} \tag{14}$$

provided $m\,,\,\bar m \in X_T$ with $\|m\|_{X_T}\,,\,\|\bar m\|_{X_T} \le R_0$ and $\|f^0\|_{W^2_{q,\mathcal{B}}} \le R_0$. We fix r such that $0 < 2r < (s \wedge 2\eta \wedge (1+1/q))$, where $s > 0$ is given in (4). For $m \in X_T$ and $p \in Y_T$ we derive from Lemma 2.1(iii), (10), and (4)

(15)
$$\begin{aligned}\|U_\alpha \star S(m,p)(t)\|_{W^2_q} &\le c\int_0^t \|U_\alpha(t-s)\|_{\mathcal{L}(W^{2r}_{q,\mathcal{B}},W^2_{q,\mathcal{B}})}\big(\|p(s)\|_{W^{2\eta}_q} + \|m(s)\|_{W^2_q}\big)\,\mathrm{d}s\\ &\le c\,t^{r-\xi}\,\mathsf{B}(r,1-\xi)\,\|p\|_{Y_T} + c\,t^{r-\mu}\,\mathsf{B}(r,1-\mu)\,\|m\|_{X_T}\ ,\end{aligned}$$

where B denotes the beta function. Therefore,

$$\|F_2(m,p) - F_2(\bar m,\bar p)\|_{X_T} \le c\,T^r\,(\|p-\bar p\|_{Y_T} + \|m-\bar m\|_{X_T}) \tag{16}$$

for $m,\bar m \in X_T$ and $p,\bar p \in Y_T$. Next observe that

$$\nabla\cdot\big(p\,\chi(f)\,\nabla f\big) = p\,\chi(f)\,\Delta f + \chi(f)\,\nabla p\cdot\nabla f + p\,\chi'(f)\,|\nabla f|^2;$$

hence

$$\|\nabla\cdot\big(p\,\chi(f)\,\nabla f\big)\|_{L_q}\ \le\ c\,\|p\|_{W_q^{2\eta}}\,(1+\|f\|_{W_q^2}^3)\ ,\quad p\in W_{q,\mathcal{B}}^{2\eta}\ ,\quad f\in W_{q,\mathcal{B}}^2\ ,$$

by Lemma 2.1 and (5). Given $f\in W_T$, $p\in Y_T$, and $w\in Z_T$ we thus compute, using (10) and (7),

(17)
$$\begin{aligned}\|U_\beta\star Q(f,p,w)(t)\|_{W_q^{2\eta}}\le c\int_0^t\|U_\beta(t-s)\|_{\mathcal{L}(L_q,W_{q,\mathcal{B}}^{2\eta})}\Big\{&\|p(s)\|_{W_q^{2\eta}}\left(1+\|f(s)\|_{W_q^2}^3\right)\\&+\left(1+\|w(s)\|_{L_q}\right)\|p(s)\|_{W_q^{2\eta}}\Big\}\,\mathrm{d}s\\\le\ c\,t^{1-\eta-\xi}\,\mathsf{B}(1-\eta,1-\xi)\,\|p\|_{Y_T}&\left(1+\|f\|_{W_T}^3+\|w\|_{Z_T}\right)\ .\end{aligned}$$

Similarly, for $f,\bar f\in W_T$, $p,\bar p\in Y_T$, and $w,\bar w\in Z_T$ we obtain

(18)
$$\begin{aligned}\|F_3(f,p,w)\ -\ F_3(\bar f,\bar p,\bar w)\|_{Y_T}\ \le&\,c\,T^{1-\eta}\,\|p\|_{Y_T}\left(1+\|f\|_{W_T}+\|\bar f\|_{W_T}\right)^2\|f-\bar f\|_{W_T}\\&+c\,T^{1-\eta}\left(1+\|\bar f\|_{W_T}^3+\|w\|_{Z_T}\right)\|p-\bar p\|_{Y_T}\\&+c\,T^{1-\eta}\,\|p\|_{Y_T}\,\|w-\bar w\|_{Z_T}\ .\end{aligned}$$

Given $f,\bar f\in W_T$, $p,\bar p\in Y_T$, and $w,\bar w\in Z_T$ analogous computations show that

(19)
$$\begin{aligned}\|F_4(f,p,w)-F_4(\bar f,\bar p,\bar w)\|_{Z_T}\ \le&\,c\,T^{1-\xi}\left(1+\|\bar p\|_{Y_T}\right)\|w-\bar w\|_{Z_T}\\&+\ c\,T^{1-\xi}\,\|w\|_{Z_T}\,\|p-\bar p\|_{Y_T}\\&+\ c\,T\,\|f-\bar f\|_{W_T}\ .\end{aligned}$$

Combining (14), (16), (18), (19), and defining $\lambda:=(1-\mu)\wedge r\wedge(1-\xi)>0$ we find a constant $\kappa(R_0)>0$ such that

$$\|F(u)-F(\bar u)\|_{\mathcal{V}_T}\ \le\ \kappa(R_0)\,T^\lambda\left(1+\|u\|_{\mathcal{V}_T}+\|\bar u\|_{\mathcal{V}_T}\right)\|u-\bar u\|_{\mathcal{V}_T}\ ,\tag{20a}$$

$$\|F(u)-V^0\|_{\mathcal{V}_T}\ \le\ \kappa(R_0)\,T^\lambda\left(1+\|u\|_{\mathcal{V}_T}\right)\|u\|_{\mathcal{V}_T}\ ,\tag{20b}$$

provided $u=(f,m,p,w)$, $\bar u=(\bar f,\bar m,\bar p,\bar w)\in\mathcal{V}_T$ are such that $\|m\|_{X_T},\|\bar m\|_{X_T}\le R_0$ and $\|f^0\|_{W_q^2}\le R_0$, where $R_0>0$ and $T\in(0,1]$ are arbitrary. Put

$$\mathcal{K}:=1+\sup_{0<t\le1}\left(t^\mu\,\|U_\alpha(t)\|_{\mathcal{L}(W_{q,\mathcal{B}}^{2(1-\mu)},W_{q,\mathcal{B}}^2)}\ +\ t^\xi\,\|U_\beta(t)\|_{\mathcal{L}(L_q,W_{q,\mathcal{B}}^{2\eta})}\right),$$

which is a finite constant according to (10), and let $R_0:=(1+\mathcal{K})B$ for $B\ge1$ given. Choose then $T:=T(B)\in(0,1]$ such that

$$\kappa(R_0)\,(1+R_0)\,R_0\,T^\lambda\ \le\ \frac12\quad\text{and}\quad k(1,R_0)\,B\,T^{1-\mu}\ \le\frac14\ ,\tag{21}$$

the constant $k(1,R_0)>0$ stemming from Lemma 2.2(ii). Notice that, in particular, for $u^0=\big(f^0,m^0,p^0,w^0\big)\in E$ with $\|u^0\|_E\le B$, there holds

$$\|V^0\|_{\mathcal{V}_T}\le\mathcal{K}\,B\ ,\quad V^0=\big(f^0,U_\alpha m^0,U_\beta p^0,U_\gamma w^0\big)\ .$$

Denoting by $\mathcal{B}_T$ the closed ball in $\mathcal{V}_T$ with center V^0 and radius B, we hence have

$$\|u\|_{\mathcal{V}_T} \le (1+\mathcal{K})\, B = R_0 , \quad u \in \mathcal{B}_T .$$

Therefore, in view of (20a), (20b), and (21), the mapping $F : \mathcal{B}_T \to \mathcal{B}_T$ is a contraction (with contraction constant less than $1/2$), which implies the existence of a unique solution to problem (M) for any $u^0 = \big(f^0, m^0, p^0, w^0\big) \in E$ with $\|u^0\|_E \le B$. If $\bar{u}^0 = \big(\bar{f}^0, \bar{m}^0, \bar{p}^0, \bar{w}^0\big) \in E$ with $\|\bar{u}^0\|_E \le B$ is another initial value, there exists a corresponding unique solution $\bar{u} = \big(\bar{f}, \bar{m}, \bar{p}, \bar{w}\big) \in \mathcal{V}_T$ satisfying $\|\bar{u}\|_{\mathcal{V}_T} \le R_0$. Defining

$$\tilde{F} := (F_2, F_3, F_4) \quad \text{and} \quad \bar{V}^0 := \big(\bar{f}^0, U_\alpha \bar{m}^0, U_\beta \bar{p}^0, U_\gamma \bar{w}^0\big)$$

we derive from (20a), (20b), and Lemma 2.2(ii) that

$$\begin{aligned}
\|u - \bar{u}\|_{\mathcal{V}_T} &\le \|F_1[f^0](m) - F_1[\bar{f}^0](\bar{m})\|_{W_T} + \|\tilde{F}(u) - \tilde{F}(\bar{u})\|_{X_T \times Y_T \times Z_T} \\
&\quad + \|U_\alpha(m^0 - \bar{m}^0)\|_{X_T} + \|U_\beta(p^0 - \bar{p}^0)\|_{Y_T} + \|U_\gamma(w^0 - \bar{w}^0)\|_{Z_T} \\
&\le k(1, R_0)\|f^0\|_{W^2_{q,\mathcal{B}}}\, T^{1-\mu}\, \|m - \bar{m}\|_{X_T} + k(1, R_0)\, \|\bar{m}\|_{X_T}\, \|f^0 - \bar{f}^0\|_{W^2_{q,\mathcal{B}}} \\
&\quad + \frac{1}{2}\, \|u - \bar{u}\|_{\mathcal{V}_T} + \|V^0 - \bar{V}^0\|_{\mathcal{V}_T} .
\end{aligned}$$

But then, due to (21),

$$\|u - \bar{u}\|_{\mathcal{V}_T} \le c(R_0)\, \|V^0 - \bar{V}^0\|_{\mathcal{V}_T} \le c(R_0)\, \mathcal{K}\, \|u^0 - \bar{u}^0\|_E ,$$

whence $\bar{u} \to u$ in $\mathcal{V}_T$ as $\bar{u}^0 \to u^0$ in E. This proves the proposition. □

We now focus on the existence of a unique maximal solution to (H_1)–(H_6) enjoying the regularity properties stated in Theorem 1.1.

Let $(1 \vee n/2) < q < \infty$ and $2\delta \in (0,2) \setminus \{1 + 1/q\}$ be given. Fix η and λ such that $n/q < 2\eta < 2$ with $2\eta \ge 1$ and $(1-\delta) \vee \eta \le \lambda < 1$ and put $(\xi, \mu) := (\eta, \lambda)$. Then, for $\big(f^0, m^0, p^0, w^0\big) \in W^2_{q,\mathcal{B}} \times W^{2\delta}_{q,\mathcal{B}} \times L_q \times L_q$, Proposition 3.1 ensures the existence of $T > 0$ and a unique solution

$$u = (f, m, p, w) \in C([0,T], W^2_{q,\mathcal{B}}) \times C_\mu((0,T], W^2_{q,\mathcal{B}}) \times C_\xi((0,T], W^{2\eta}_{q,\mathcal{B}}) \times C([0,T], L_q)$$

to problem (M). As in (17),

$$\|U_\beta \star Q(f,p,w)(t)\|_{L_q} \le c(T)\, t^{1-\xi} \to 0 \quad \text{as} \quad t \to 0^+ ,$$

and therefore

$$p = U_\beta\, p^0 + U_\beta \star Q(f,p,w) \in C([0,T], L_q)$$

is a mild L_q-solution to (H_3). From this and the identity

$$m = U_\alpha\, m^0 + U_\alpha \star S(m,p)$$

we obtain that $m \in C([0,T], W^{2\delta}_{q,\mathcal{B}})$ is a mild L_q-solution to (H_2). Clearly, $w \in C([0,T], L_q)$ is a mild L_q-solution to (H_4).

Next we show that these mild solutions are actually classical solutions. First, we fix $\varepsilon \in (0,T]$ and set $I := [0, T-\varepsilon]$. Then

$$f_\varepsilon := f(\cdot + \varepsilon) \in C(I, W^2_{q,\mathcal{B}}) , \qquad m_\varepsilon := m(\cdot + \varepsilon) \in C(I, W^2_{q,\mathcal{B}}) ,$$
$$p_\varepsilon := p(\cdot + \varepsilon) \in C(I, W^{2\eta}_{q,\mathcal{B}}) , \qquad w_\varepsilon := w(\cdot + \varepsilon) \in C(I, L_q) .$$

Furthermore, Lemma 2.2(i) warrants $f_\varepsilon \in C^1(I, W^2_{q,\mathcal{B}})$ and thus, letting $\varepsilon \to 0^+$, we obtain $f \in C^1((0,T], W^2_{q,\mathcal{B}})$. Next, $h_\varepsilon := S(m_\varepsilon, p_\varepsilon) \in C(I, W^{2r}_{q,\mathcal{B}})$ with $r > 0$ sufficiently small by Lemma 2.1(iii). Therefore, observing that $W^{2r}_{q,\mathcal{B}}$ is a (real) interpolation space between L_q and $W^2_{q,\mathcal{B}}$ (cf. [24]) and taking into account that m_ε is a mild L_q-solution to the linear problem

$$\dot{M} - \alpha \Delta M = h_\varepsilon(t) \ , \quad t \in \dot{I} \ , \qquad M(0) = m_\varepsilon(0) \in W^2_{q,\mathcal{B}} \ ,$$

we conclude that $m_\varepsilon \in C^1(\dot{I}, L_q) \cap C(I, W^2_{q,\mathcal{B}})$ since mild solutions to linear problems are unique; see [3, II.Thm. 1.2.2]. Letting $\varepsilon \to 0^+$ we deduce that m is a classical solution to (H_2) possessing the same regularity properties as m_ε on $(0,T]$. Next, define $j_\varepsilon := Q(f_\varepsilon, p_\varepsilon, w_\varepsilon) \in C(I, L_q)$ and notice that p_ε is a mild L_q-solution to the linear problem

$$\dot{P} - \beta \Delta P = j_\varepsilon(t) \ , \quad t \in \dot{I} \ , \qquad P(0) = p_\varepsilon(0) \in W^{2\eta}_{q,\mathcal{B}} \ . \tag{22}$$

Thus, $p_\varepsilon \in C^\varrho(I, W^{2\sigma}_{q,\mathcal{B}})$ with $2\eta > 2\sigma > n/q$ and $2\sigma \ge 1$, where $\varrho := \eta - \sigma > 0$ owing to [3, II.Thm. 5.3.1]. Clearly, due to $k_\varepsilon := R(f_\varepsilon, p_\varepsilon, w_\varepsilon) \in C(I, L_q)$ and (10) we have

$$w_\varepsilon \ = \ U_\gamma \, w_\varepsilon(0) \ + \ U_\gamma \star k_\varepsilon \in C(I, W^{2\nu}_{q,\mathcal{B}}) \ , \quad \nu < 1 \ .$$

Applying again [3, II.Thm. 5.3.1] we obtain $w_\varepsilon \in C^\varrho(I, L_q)$. From (6) and the fact that $f_\varepsilon \in C^1(I, W^2_{q,\mathcal{B}})$ and $p_\varepsilon \in C^\varrho(I, W^{2\sigma}_{q,\mathcal{B}})$ it follows that $k_\varepsilon \in C^\varrho(I, L_q)$; hence, as above, $w_\varepsilon \in C^1(\dot{I}, L_q) \cap C(\dot{I}, W^2_{q,\mathcal{B}})$ by [3, II.Thm. 1.2.2], which ensures that w is a classical solution to (H_4) with the corresponding regularity properties. Moreover, recalling that $2\sigma > n/q$ with $2\sigma \ge 1$ and invoking Lemma 2.1, we deduce $j_\varepsilon \in C^\varrho(I, L_q)$ thanks to (5). Due to [3, II.Thm. 1.2.2] and (22) this implies that p_ε belongs to $C^1(\dot{I}, L_q) \cap C(\dot{I}, W^2_{q,\mathcal{B}})$, whence $p \in C^1((0,T], L_q) \cap C((0,T], W^2_{q,\mathcal{B}})$ is a classical solution to (H_3).

Let us now prove that this solution is unique in the sense stated in Theorem 1.1. Suppose therefore that there exist two solutions $(\tilde{f}, \tilde{m}, \tilde{p}, \tilde{w})$ and $(\bar{f}, \bar{m}, \bar{p}, \bar{w})$ to (H_1)–(H_6) on some interval $[0,T]$ satisfying

$$\tilde{f} \, , \, \bar{f} \in C\big([0,T], W^2_{q,\mathcal{B}}\big) \ , \qquad \tilde{w} \, , \, \bar{w} \in C\big([0,T], L_q\big) \ ,$$

$$\tilde{m} \in C_{\tilde{\lambda}}\big((0,T], W^2_{q,\mathcal{B}}\big) \ , \qquad \bar{m} \in C_{\bar{\lambda}}\big((0,T], W^2_{q,\mathcal{B}}\big) \ ,$$

$$\tilde{p} \in C_{\tilde{\eta}}\big((0,T], W^{2\tilde{\eta}}_{q,\mathcal{B}}\big) \ , \qquad \bar{p} \in C_{\bar{\eta}}\big((0,T], W^{2\bar{\eta}}_{q,\mathcal{B}}\big)$$

for some $n/q < 2\tilde{\eta}$, $2\bar{\eta} < 2$ with $2\tilde{\eta}$, $2\bar{\eta} \ge 1$ and $\tilde{\lambda}$, $\bar{\lambda} < 1$. Defining

$$\eta := \tilde{\eta} \wedge \bar{\eta} \ , \qquad \xi := \tilde{\eta} \vee \bar{\eta} \ , \qquad \mu := \tilde{\lambda} \vee \bar{\lambda} \vee \xi \vee (1 - \delta)$$

we obtain two solutions to (M) such that both $\tilde{m}, \bar{m}$ belong to $C_\mu((0,T], W^2_{q,\mathcal{B}})$ and both $\tilde{p}, \bar{p}$ belong to $C_\xi((0,T], W^{2\eta}_{q,\mathcal{B}})$, where $n/q < 2\eta \le 2\xi \le 2\mu < 2$ with $2\eta \ge 1$. Making T smaller if necessary, Proposition 3.1 guarantees that $(\tilde{f}, \tilde{m}, \tilde{p}, \tilde{w})$ coincides with $(\bar{f}, \bar{m}, \bar{p}, \bar{w})$ on $[0,T]$.

Evidently, local uniqueness warrants that we may extend the solution (f,m,p,w) constructed above to a maximal solution on an interval $J := [0,t^+)$. Since, according to Proposition 3.1, the local existence time $T > 0$ can be chosen uniformly with respect to initial values that are bounded in $W^2_{q,\mathcal{B}} \times W^{2\delta}_{q,\mathcal{B}} \times L_q \times L_q$, we surely have

$$\limsup_{t \nearrow t^+} \big\| \big(f(t),m(t),p(t),w(t)\big) \big\|_{W^2_q \times W^{2\delta}_q \times L_q \times L_q} \;=\; \infty \tag{23}$$

in the case that $t^+ < \infty$.

Summing up, we have shown thus far that problem (H_1)–(H_6) admits a maximal solution being unique and possessing the regularity properties in the sense stated in Theorem 1.1. Moreover, this solution satisfies (8) and, if $t^+ < \infty$, then (23) also holds.

Remark 3.2. Given $p^0 \in W^{2\eta}_{q,\mathcal{B}}$ with $n/q < 2\eta < 2$ and $2\eta \geq 1$ there holds $p \in C(J, W^{2\eta}_{q,\mathcal{B}})$. In particular, one may choose $2\eta = 1$ if $q > n$; see Corollary 1.3. This readily follows by taking $C([0,T], W^{2\eta}_{q,\mathcal{B}})$ as state space for p instead of the weighted space $C_\xi((0,T], W^{2\eta}_{q,\mathcal{B}})$ in the above proof.

4. Positivity. Using ideas as in [18] we now show positivity of the solution corresponding to positive initial values. Given

$$\big(f^0,m^0,p^0,w^0\big) \in W^2_{q,\mathcal{B}} \times W^{2\delta}_{q,\mathcal{B}} \times L_q \times L_q$$

such that $f^0 \geq 0$, $m^0 \geq 0$, $p^0 \geq 0$, and $w^0 \geq 0$ (a.e. on Ω) let (f,m,p,w) denote the maximal solution on J constructed in the previous section. Then obviously $f(t) \geq 0$ on Ω for $t \in J$.

First suppose that $q > (n \vee 2)$ and $p^0, w^0 \in W^2_{q,\mathcal{B}}$. Fix $T \in \dot{J}$ and $n/q < 2\sigma < 1$. Then choose $\eta \in \big(1/2, 1-\sigma\big)$ and observe that $p \in C_\eta(\dot{J}, W^{2\eta}_{q,\mathcal{B}})$ in view of (8). Analogously to (17) it follows from Lemma 2.1 that

$$\|U_\beta \star Q(f,p,w)(t)\|_{W^{2\sigma}_q} \;\leq\; c(T)\, t^{1-\sigma-\eta} \;\longrightarrow 0 \quad \text{as} \quad t \to 0^+ \;,$$

and consequently

$$p \;=\; U_\beta\, p^0 \;+\; U_\beta \star Q(f,p,w) \in C([0,T], W^{2\sigma}_{q,\mathcal{B}}) \hookrightarrow C([0,T]\times\bar{\Omega}) \;.$$

Similarly, there holds

$$w \;=\; U_\gamma\, w^0 \;+\; U_\gamma \star R(f,p,w) \in C([0,T], W^{2\sigma}_{q,\mathcal{B}}) \hookrightarrow C([0,T]\times\bar{\Omega}) \;,$$

and thus, in particular,

$$\vartheta(w) \in C([0,T]\times\bar{\Omega}) \;. \tag{24}$$

According to [18, p. 451] there exist a function $H \in C^2(\mathbb{R})$ and a constant $c_0 > 0$ such that $H(z) = 0$ for $z \geq 0$ and $H(z) > 0$ for $z < 0$ and such that

$$0 \;\leq\; H''(z)\, z^2 \;\leq\; c_0\, H(z) \;, \quad z \in \mathbb{R} \;,$$

and

$$0 \;\leq\; H'(z)\, z \;\leq\; c_0\, H(z) \;, \quad z \in \mathbb{R} \;.$$

Define $M \in C^1((0,T]) \cap C([0,T])$ as

$$M(t) := \int_\Omega H\big(p(t,x)\big)\, \mathrm{d}x\ , \quad t \in [0,T]\ .$$

Owing to $\partial_\nu p(t) = \partial_\nu f(t) = 0$ we deduce from (H_3) that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} M(t) \ &= \ \int_\Omega H'(p) \left(\beta \Delta p - \nabla \cdot \big(p\, \chi(f)\, \nabla f \big) + \vartheta(w)\, p \right) \mathrm{d}x \\ &= \ -\beta \int_\Omega H''(p)\, |\nabla p|^2 \ \mathrm{d}x \ + \ \int_\Omega H''(p)\, p\, \chi(f)\, \nabla p \cdot \nabla f \ \mathrm{d}x \\ &\quad + \ \int_\Omega H'(p)\, \vartheta(w)\, p \ \mathrm{d}x. \end{aligned}$$

Therefore, since

$$|p\, \chi(f)\, \nabla p \cdot \nabla f| \ \le \ \frac{\beta}{2}\, |\nabla p|^2 \ + \ \frac{1}{2\beta}\, p^2\, \chi(f)^2\, |\nabla f|^2$$

we infer from (7) and the fact that both f and ∇f belong to $C([0,T] \times \bar{\Omega})$, from (24), and the properties of the function H that

$$\frac{\mathrm{d}}{\mathrm{d}t} M(t) \le c(T) M(t)\ , \quad t \in (0,T]\ .$$

Thus $M(0) = 0$ ensures $M(t) = 0$ for $t \in [0,T]$, that is, $p(t) \ge 0$ on Ω for $t \in [0,T]$. It is then straightforward to prove that $m(t) \ge 0$ and $w(t) \ge 0$ on Ω for $t \in [0,T]$. But $T > 0$ was arbitrary, so the desired positivity follows.

Finally, to show positivity in the general case $q > (1 \vee n/2)$ we approximate $p^0, w^0 \in L_q$ by nonnegative smooth functions and use the continuous dependence of the solution on the initial value provided by Proposition 3.1.

5. Global existence. It remains only to prove global existence. We denote by (f, m, p, w) the maximal nonnegative solution on $J = [0, t^+)$ corresponding to the nonnegative initial value

$$\big(f^0, m^0, p^0, w^0\big) \in W^2_{q,\mathcal{B}} \times W^{2\delta}_{q,\mathcal{B}} \times L_q \times L_q\ .$$

We first claim that it suffices to prove

$$\sup_{t \in J \cap [0,T]} \|p(t)\|_{L_q} \ < \ \infty\ , \quad T > 0\ , \tag{25}$$

in order to conclude that $t^+ = \infty$. Indeed, suppose that (25) holds for any $T > 0$ and set $J_T := J \cap [0,T]$. Replacing the solution by the shifted solution $(f_\varepsilon, m_\varepsilon, p_\varepsilon, w_\varepsilon)$ introduced in the existence proof in section 3, we may assume without loss of generality that all m, p, w belong to $C(J, W^2_{q,\mathcal{B}}) \cap C^1(J, L_q)$, in particular that $m^0 \in W^2_{q,\mathcal{B}}$. Observe then that $w \in L_\infty(J_T, L_\infty)$ as it follows from (H_4) since $w(t) \ge 0$ and $\|f(t)\|_\infty \le \|f^0\|_\infty$. Next, since $b \in L_\infty$, we may choose $\lambda > 0$ sufficiently large such that $-(\lambda + b - \alpha\Delta)$ has bounded imaginary powers with angle strictly less than $\pi/2$ (for instance, see [3, III.Ex. 4.7.3(d), III.Thm. 4.8.7]). Therefore, defining $n(t) := e^{-\lambda t} m(t)$ and noticing that

$$\dot{n} \ + \ (\lambda + b - \alpha\Delta)\, n \ = \ d\, e^{-\lambda t}\, p(t) \ =: \ z(t)\ , \quad n(0) \ = \ m^0 \in W^2_{q,\mathcal{B}}\ ,$$

it follows from [3, III.Thm. 4.10.7] that $n \in L_q(J_T, W^2_{q,\mathcal{B}})$ since $z \in L_\infty(J_T, L_q)$ by (25) and $n(0) \in W^2_{q,\mathcal{B}}$. But then

$$\int_0^t \|m(s)\|_{W^2_q} \, \mathrm{d}s \, \le \, c(T) \, , \quad t \in J_T \, ,$$

and we deduce from (13) that $f \in L_\infty(J_T, W^2_{q,\mathcal{B}})$. Finally, owing to $p \in L_\infty(J_T, L_q)$, (10), and Gronwall's inequality we conclude from (H_2) that $m \in L_\infty(J_T, W^{2\delta}_{q,\mathcal{B}})$. Consequently, combining all the estimates on f, m, p, and w we see that the blowup criterion (23) implies $t^+ = \infty$ since $T > 0$ was arbitrary. Therefore, (25) is indeed sufficient to conclude global existence.

To derive the desired L_q-bound on p we employ a change of variable of the form $p \mapsto \frac{p}{\phi(f)}$, where ϕ solves

$$\phi'(z) \, = \, \frac{\chi(z)}{\beta} \, \phi(z) \, , \quad z > 0 \, , \qquad \phi(0) = 1 \, .$$

This device has been used in [11, 12, 14] for equations of the form (1), (2), (3) and leads in our case to the equation in divergence form

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{p}{\phi(f)} \, = \, \frac{\beta}{\phi(f)} \, \nabla \cdot \left(\phi(f) \, \nabla \frac{p}{\phi(f)} \right) \, + \, \vartheta(w) \, \frac{p}{\phi(f)} \, + \, \frac{a}{\beta} \, \chi(f) \, f \, m \, \frac{p}{\phi(f)} \, . \tag{26}$$

Global existence is then an easy consequence of the following proposition, where the basic idea of its proof is adapted from [11]. We point out here again that in our case, the coupling of (H_1) and (H_3) via (H_2) allows us to derive the a priori estimate for p (which does not seem to be possible without a smallness condition on the initial value in the case of (1), (2), (3) with $\sigma = -1$; see [11]).

PROPOSITION 5.1. *Suppose that* $\|p(t)\|_{L_\rho} \le c(T)$, $t \in J_T := J \cap [0, T]$, *for some* $\rho \in [1, q)$ *and suppose there exists* $\varrho \in (\rho, 2\rho \wedge q]$ *such that*

$$\varrho \left(\frac{n}{\rho} - 2 \right) \, < \, 2 \left(\rho - 1 + \frac{2\rho}{n} \right) \, . \tag{27}$$

Then $\|p(t)\|_{L_\varrho} \le c(T)$ *for* $t \in J_T$.

Proof. We first observe that (27) allows to fix $r > 1$ such that

$$\frac{n\varrho}{n\varrho + 2\rho} \, < \, \frac{1}{r} \, < \, 1 + \frac{2}{n} - \frac{1}{\rho} \, . \tag{28}$$

If $\varrho \ge 2$, we set $\mu := 0$; otherwise we fix $\mu \in (0, 1)$. Then we put $p_\mu := p + \mu \ge \mu$ and note that

$$\nabla \left(\frac{p_\mu}{\phi(f)} \right)^{\varrho/2} \, = \, \frac{\varrho}{2} \, \left(\frac{p_\mu}{\phi(f)} \right)^{\varrho/2-1} \, \nabla \frac{p_\mu}{\phi(f)}$$

by the chain rule. Hence $\left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2} \in W^1_2$ since $W^1_q \hookrightarrow L_2$ due to $q > n/2$. Moreover, $\partial_\nu \frac{p_\mu}{\phi(f)} = 0$ owing to $\partial_\nu p = \partial_\nu f = 0$. Thus, given any $\Lambda \in C^2\big((0, \infty)\big)$ we derive

from (26)

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi(f)\,\Lambda\left(\frac{p_\mu}{\phi(f)}\right)\,\mathrm{d}x\\
&\quad=\beta\int_\Omega \Lambda'\left(\frac{p_\mu}{\phi(f)}\right)\,\nabla\cdot\left(\phi(f)\,\nabla\frac{p_\mu}{\phi(f)}\right)\,\mathrm{d}x+\int_\Omega \Lambda'\left(\frac{p_\mu}{\phi(f)}\right)\,\vartheta(w)\,p\,\mathrm{d}x\\
&\qquad+\frac{1}{\beta}\int_\Omega \Lambda'\left(\frac{p_\mu}{\phi(f)}\right)\,a\,m\,\chi(f)\,f\,p\,\mathrm{d}x-\frac{1}{\beta}\int_\Omega \Lambda\left(\frac{p_\mu}{\phi(f)}\right)\,a\,m\,\chi(f)\,f\,\phi(f)\,\mathrm{d}x\\
&\qquad+\mu\int_\Omega \Lambda'\left(\frac{p_\mu}{\phi(f)}\right)\,\nabla\cdot\left(\chi(f)\nabla f\right)\,\mathrm{d}x+\frac{\mu}{\beta}\int_\Omega \Lambda'\left(\frac{p_\mu}{\phi(f)}\right)\,a\,m\,\chi(f)\,f\,\mathrm{d}x\\
&\quad=-\beta\int_\Omega \Lambda''\left(\frac{p_\mu}{\phi(f)}\right)\,\phi(f)\,\left|\nabla\frac{p_\mu}{\phi(f)}\right|^2\,\mathrm{d}x\\
&\qquad+\frac{1}{\beta}\int_\Omega a\,m\,\chi(f)\,f\left[p\,\Lambda'\left(\frac{p_\mu}{\phi(f)}\right)-\phi(f)\,\Lambda\left(\frac{p_\mu}{\phi(f)}\right)\right]\,\mathrm{d}x\\
&\qquad+\int_\Omega \Lambda'\left(\frac{p_\mu}{\phi(f)}\right)\,\vartheta(w)\,p\,\mathrm{d}x-\mu\int_\Omega \Lambda''\left(\frac{p_\mu}{\phi(f)}\right)\,\chi(f)\,\nabla\frac{p_\mu}{\phi(f)}\cdot\nabla f\,\mathrm{d}x\\
&\qquad+\frac{\mu}{\beta}\int_\Omega \Lambda'\left(\frac{p_\mu}{\phi(f)}\right)\,a\,m\,\chi(f)\,f\,\mathrm{d}x
\end{aligned}
$$

for $t\in J$. In particular, taking $\Lambda(z)=z^\varrho$ we have

(29)

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi(f)\,\left(\frac{p_\mu}{\phi(f)}\right)^\varrho\,\mathrm{d}x\;\le\;&-4\,\beta\,\frac{\varrho-1}{\varrho}\int_\Omega \phi(f)\,\left|\nabla\left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2}\right|^2\,\mathrm{d}x\\
&+S_0\int_\Omega m\,\left(\frac{p_\mu}{\phi(f)}\right)^\varrho\,\mathrm{d}x\\
&+\varrho\,\|\vartheta(w)\|_\infty\int_\Omega \phi(f)\,\left(\frac{p_\mu}{\phi(f)}\right)^\varrho\,\mathrm{d}x\\
&+\mu\,\varrho\,(\varrho-1)\|\chi(f)\|_\infty\int_\Omega \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho-2}\,\left|\nabla\frac{p_\mu}{\phi(f)}\cdot\nabla f\right|\,\mathrm{d}x\\
&+\mu\,S_0\int_\Omega m\,\left(\frac{p_\mu}{\phi(f)}\right)^{\varrho-1}\,\mathrm{d}x
\end{aligned}
$$

for $t\in J$, where

$$
S_0\;:=\;\frac{\varrho-1}{\beta}\,\|a\|_\infty\sup_{0<z<\|f^0\|_\infty}\left(z\,\chi(z)\,\phi(z)\right)\;<\;\infty
$$

since $\|f(t)\|_\infty\le\|f^0\|_\infty$ and $\|\vartheta(w)\|_\infty<\infty$ on J_T due to $w\in L_\infty(J_T,L_\infty)$. Next, we use the second inequality of (28), (11), the given L_ρ-bound on p, and Gronwall's inequality to derive from (H_2) that

$$
\|m(t)\|_{L_{r'}}\;\le\;c(T)\;,\quad t\in J_T\;,
$$

where r' denotes the dual exponent of r. Hence, taking into account that the first inequality of (28) warrants the following version of the Gagliardo–Nirenberg inequality (see [15, p. 37])

$$\|\cdot\|_{L_{2r}}^{2r} \le c_0 \|\cdot\|_{L_{2\rho/\varrho}}^{2(r-1)} \|\cdot\|_{W_2^1}^2 ,$$

applying Young's inequality, and using once again the given L_ρ-bound on p, it follows for $\varepsilon > 0$ that

$$\begin{aligned} S_0 \int_\Omega m \left(\frac{p_\mu}{\phi(f)}\right)^\varrho \,\mathrm{d}x &\le c(\varepsilon) \int_\Omega m^{r'} \,\mathrm{d}x + \varepsilon \int_\Omega \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho r} \,\mathrm{d}x \\ &\le c(T,\varepsilon) + \varepsilon \left\| \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2} \right\|_{L_{2r}}^{2r} \\ &\le c(T,\varepsilon) + \varepsilon\, c_0 \left\| \frac{p_\mu}{\phi(f)} \right\|_{L_\rho}^{\varrho(r-1)} \left\| \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2} \right\|_{W_2^1}^2 \\ &\le c(T,\varepsilon) + c(T,\varepsilon) \int_\Omega \left(\frac{p_\mu}{\phi(f)}\right)^\varrho \,\mathrm{d}x \\ &\quad + \varepsilon\, c(T) \int_\Omega \left| \nabla \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2} \right|^2 \,\mathrm{d}x . \end{aligned}$$

We can estimate the last term in (29) similarly, since

$$\begin{aligned} \mu\, S_0 \int_\Omega m \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho-1} \,\mathrm{d}x &\le \mu\, c(T,\varepsilon) + \mu\, c(T,\varepsilon) \int_\Omega \left(\frac{p_\mu}{\phi(f)}\right)^\varrho \,\mathrm{d}x \\ &\quad + \mu\, \varepsilon\, c(T) \int_\Omega \left| \nabla \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2} \right|^2 \,\mathrm{d}x . \end{aligned}$$

In the case that $\varrho < 2$ we have by Young's inequality for $\delta > 0$

$$\begin{aligned} \mu\, \varrho\, (\varrho-1) \|\chi(f)\|_\infty &\int_\Omega \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho-2} \left| \nabla \frac{p_\mu}{\phi(f)} \cdot \nabla f \right| \,\mathrm{d}x \\ &\le \delta\, \mu\, \frac{\varrho^2}{4} \int_\Omega \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho-2} \left| \nabla \frac{p_\mu}{\phi(f)} \right|^2 \,\mathrm{d}x \\ &\quad + \mu\, c(\delta) \int_\Omega \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho-2} |\nabla f|^2 \,\mathrm{d}x \\ &\le \mu\, \delta \int_\Omega \left| \nabla \left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2} \right|^2 \,\mathrm{d}x + \mu^{\varrho-1}\, c(\delta) \int_\Omega \phi(f)^{\varrho-2} |\nabla f|^2 \,\mathrm{d}x . \end{aligned}$$

Therefore, due to $\phi(f) \ge 1$ and $\mu < 1$, we infer from (29) by combining the above

estimates that for all $t \in J_T$

(30)
$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi(f)\,\left(\frac{p_\mu}{\phi(f)}\right)^\varrho\,\mathrm{d}x \le c(T,\varepsilon) + c(T,\varepsilon)\int_\Omega \phi(f)\,\left(\frac{p_\mu}{\phi(f)}\right)^\varrho\,\mathrm{d}x + \left(\varepsilon\,c(T) + \delta - 4\,\beta\,\frac{\varrho-1}{\varrho}\right)\int_\Omega \left|\nabla\left(\frac{p_\mu}{\phi(f)}\right)^{\varrho/2}\right|^2\,\mathrm{d}x + \mu^{\varrho-1}\,c(\delta)\int_\Omega \phi(f)^{2-\varrho}\,|\nabla f|^2\,\mathrm{d}x\ .$$

We then choose $\varepsilon > 0$ and $\delta > 0$ sufficiently small so that the term involving the gradient of $\frac{p_\mu}{\phi(f)}$ becomes negative. Recalling that $\|\phi(f)\|_\infty \le c\big(\|f^0\|_\infty\big)$ on J_T, that $\nabla f(t) \in L_2$, and that $p \in C^1(J, L_q)$ we may then let $\mu \to 0^+$ and use Lebesgue's theorem to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \phi(f)\,\left(\frac{p}{\phi(f)}\right)^\varrho\,\mathrm{d}x \le c(T) + c(T)\int_\Omega \phi(f)\,\left(\frac{p}{\phi(f)}\right)^\varrho\,\mathrm{d}x$$

for all $t \in J_T$ since $1 < \varrho \le q$. Thus, we conclude $\|p(t)\|_{L_\varrho} \le c(T)$ for $t \in J_T$. □

We are now in a position to prove that indeed $J = \mathbb{R}^+$. Since p is nonnegative, $\|\vartheta(w(t))\|_\infty \le c(T)$, and $\partial_\nu p(t) = \partial_\nu f(t) = 0$ for $t \in J_T$ it follows that $\|p(t)\|_{L_1} \le c(T)$, $t \in J_T$, by integrating (H_3). Therefore, we may apply Proposition 5.1 successively to derive $\|p(t)\|_{L_q} \le c(T)$ for $t \in J_T$; hence $J = \mathbb{R}^+$ according to (25). Consequently, the proof of Theorem 1.1 is complete.

6. Numerical examples. We illustrate the theoretical results above with numerical examples (a numerical treatment of a more general model is given in [7]). The parameters for the example are chosen for illustrative purposes.

The region Ω is $[0,6] \times [0,6] \subset \mathbb{R}^2$, the parameters are $a(x) \equiv 5.0$, $\alpha = .01$, $d(x) \equiv 1.0$, $b(x) \equiv 1.0$, $\beta = .01$, $\chi(f) \equiv 0.0$, or $\chi(f) \equiv 0.4$, $\theta(x,w) \equiv 0.1$, $\varrho(x,w) = 2.0\,w/(1.0+w)$, $\gamma = 0.1$, $e(x) \equiv 1.0$, $\omega(x,p) = 2.0\,p/(1.0+p)$, $g(x) \equiv 5.0$, and the initial conditions are

$$f^0(x) = 0.05\,\cos\big((10.0\,\pi/36.0)\,x_1^2\big)\,\sin\big((13.0\,\pi/72.0)\,x_2^2\big) + 0.3\ ,$$
$$p^0(x) = 5.0\,\max\big\{0.3 - (x_1-3.0)^2 - (x_2-3.0)^2\,,\,0.0\big\}\,,$$

$m^0(x) = p^0(x)$, and $w^0(x) = 4.0\,f^0(x)$, where $x = (x_1, x_2)$. The normalized tumor density is initially distributed symmetrically in a circle. The normalized extracellular matrix density is immobile and heterogeneous above a uniform background value. The haptotactic parameter χ is an indicator of the relative strength of cell-matrix adhesion, and the value of χ increases through successive mutations of the tumor cell lines, as tumor cells gain greater capacity to invade the surrounding bound substrate [5]. We provide two choices for the haptotaxis parameter χ to demonstrate this increase in χ. In Figures 1, 2, 3 the value of χ is 0.0, so that all movement of cells is due only to cell motility. In Figures 4, 5, 6 the value of χ is 0.4, so that movement of cells is due to both cell motility and haptotactic directed attraction. The simulations demonstrate that haptotaxis produces a profound distinction in the spatial behavior in the two cases. Without haptotaxis the tumor expands slowly and symmetrically (Figures 1 and 2) as the total population declines (Figure 3). With haptotaxis the

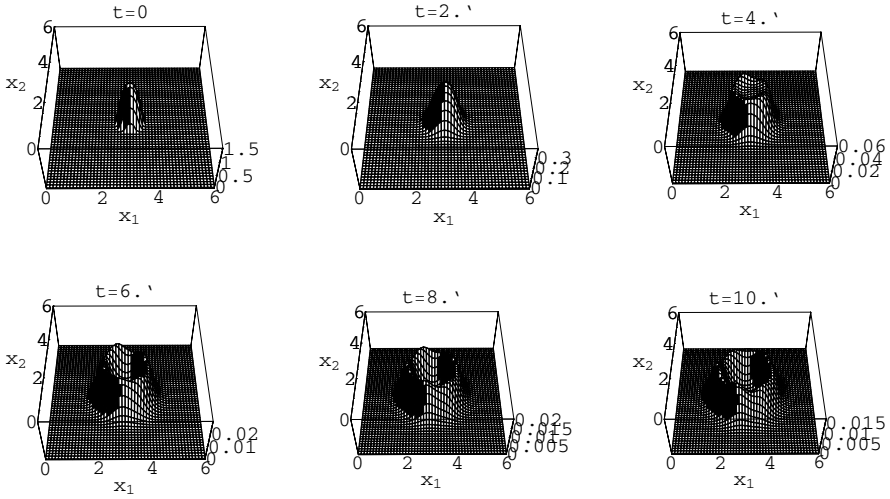


FIG. 1. *The normalized tumor cell density for various times in the case without haptotaxis ($\chi = 0.0$). The tumor slowly expands nearly symmetrically as it decreases in total mass. The interior of the tumor becomes necrotic as tumor cells consume and exhaust the supply of oxygen furnished by the extracellular matrix.*

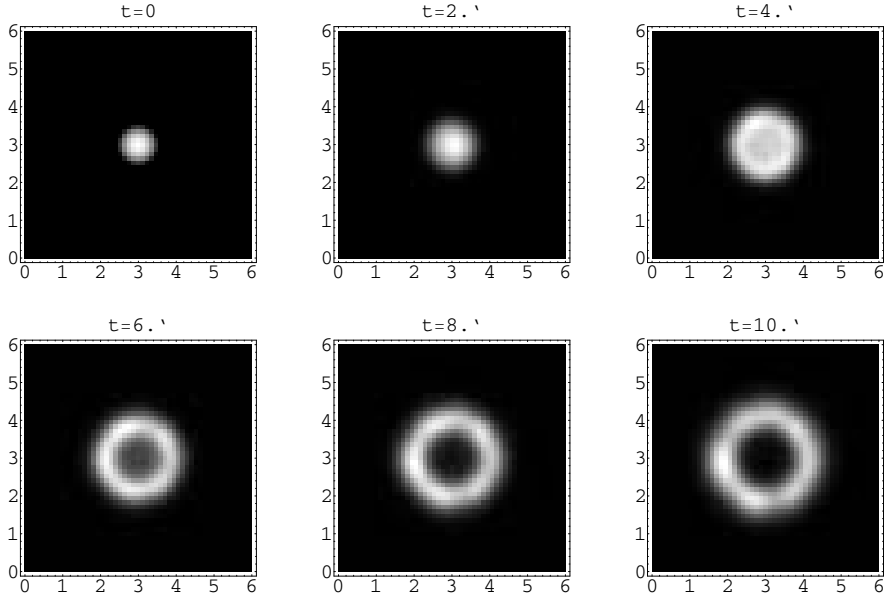


FIG. 2. *The density plots in the (x_1, x_2)-coordinate system of the tumor cell distributions in Figure* 1 *($\chi = 0.0$).*

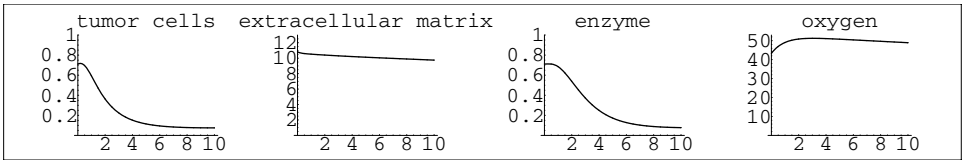


FIG. 3. *The total populations in the case without haptotaxis ($\chi = 0.0$): $\int_\Omega p(x,t)\mathrm{d}x$, $\int_\Omega f(x,t)\mathrm{d}x$, $\int_\Omega m(x,t)\mathrm{d}x$, $\int_\Omega w(x,t)\mathrm{d}x$ as functions of time. The total tumor mass eventually shrinks to a very low value.*

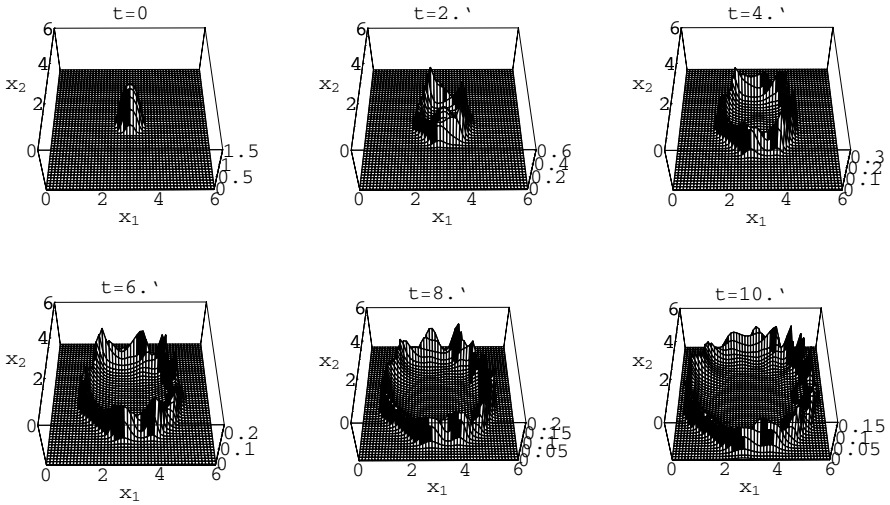


FIG. 4. *The normalized tumor cell density for various times in the case with haptotaxis ($\chi = 0.4$). The tumor expands more rapidly and asymmetrically as it increases in total mass.*

tumor spreads much more rapidly and asymmetrically (Figures 4 and 5) as the total tumor cell population increases (Figure 6) for a time. The distinction of the two cases demonstrates the importance of haptotaxis in the ability of tumors to invade surrounding tissue.

7. Summary. In Theorem 1.1 we have proven the existence of unique classical global solutions to the model of tumor growth (H_1)–(H_6). The model describes the spatial invasion of a tumor mass into its surrounding extracellular matrix. A key feature of the model is that the migration of tumor cells is due primarily to haptotaxis-directed movement. The interpretation of haptotaxis in tumor growth is that cell movement is controlled by the differential strengths of cell-cell adhesion gradients. Haptotaxis differs from chemotaxis in that the directed migration of the tumor cells toward concentrations of the extracellular macromolecules is mediated by a diffusive enzyme produced by the tumor cells. This enzyme degrades the matrix macromolecules, which produce the oxygen essential for tumor growth, and thus alters patterns of tumor movement and proliferation. The haptotaxis process in the model produces technical complications, but also yields the regularity of solutions essential in the analysis. We have demonstrated the role of haptotaxis in two numerical examples. In the first example, without haptotaxis, the only spatial movement of tumor

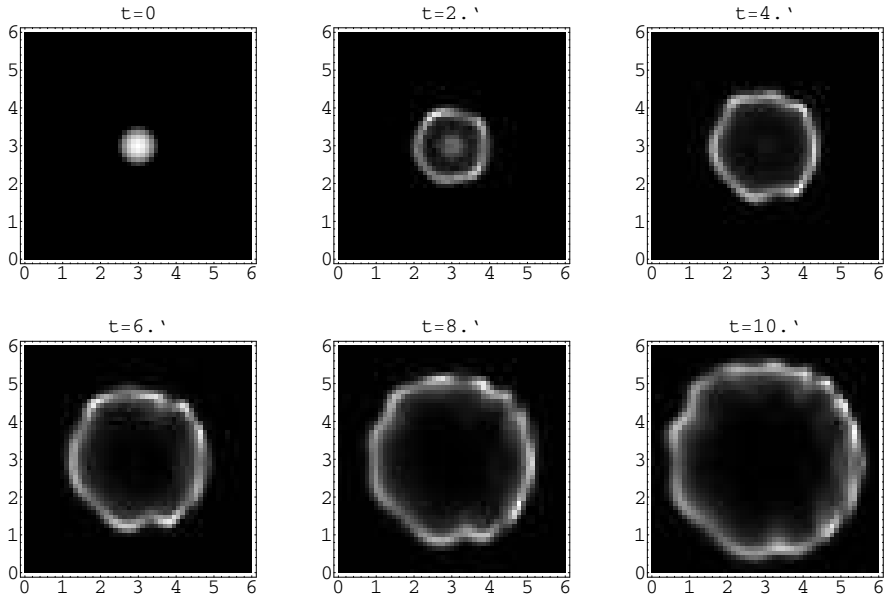


FIG. 5. *The density plots in the* (x_1, x_2)*-coordinate system of the tumor cell distributions in Figure* 4 *(*$\chi = 0.4$*).*

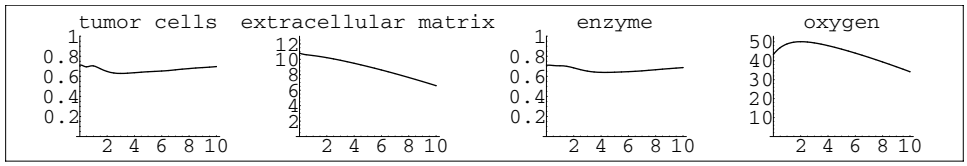


FIG. 6. *The total populations in the case with haptotaxis (*$\chi = 0.4$*):* $\int_\Omega p(x,t)\mathrm{d}x$*,* $\int_\Omega f(x,t)\mathrm{d}x$*,* $\int_\Omega m(x,t)\mathrm{d}x$*,* $\int_\Omega w(x,t)\mathrm{d}x$ *as functions of time. The total tumor mass grows for an interval of time.*

cells is due to cell motility modeled by diffusion. In this example the tumor invades slowly and decreases in total tumor mass. In the second example, with all parameters the same as in the first example, but with the addition of haptotaxis, the tumor invades more rapidly and with increasing total tumor mass. Both examples show the characteristic interior necrosis of tumor cells due to exhaustion of the oxygen supply, but the effect is much more pronounced with haptotaxis. The utilization of oxygen by the tumor cell population is critical in understanding the distinction of the two examples. If the oxygen concentration is constant in time, then the evolution of the total tumor mass is independent of haptotaxis. If the oxygen concentration evolves in time due to tumor consumption and degradation of its source, then haptotaxis-directed spatial migration enables a more efficient utilization of the environmental resources and results in a more aggressive invasion of the tumor into the surrounding tissue.

REFERENCES

[1] H. AMANN, *Dual semigroups and second order linear elliptic boundary value problems*, Israel J. Math., 45 (1983), pp. 225–254.

[2] H. Amann, *Multiplication in Sobolev and Besov spaces*, in Nonlinear Analysis. A Tribute in Honour of Giovanni Prodi, Quaderni, Scuola Norm. Sup., 1991, pp. 27–57.
[3] H. Amann, *Linear and Quasilinear Parabolic Problems, Volume* I: *Abstract Linear Theory*, Birkhäuser, Boston, 1995.
[4] H. Amann and Ch. Walker, *Local and global strong solutions to continuous coagulation-fragmentation equations with diffusion*, J. Differential Equations, 218 (2004), pp. 1–28.
[5] A. R. A. Anderson, *A hybrid mathematical model of solid tumour invasion: The importance of cell adhesion*, Math. Med. Biol. IMA J., 22 (2005), pp. 163–186.
[6] A. R. A. Anderson and M. Chaplain, *Continuous and discrete mathematical models of tumor-induced angiogenesis*, Bull. Math. Biol., 60 (1998), pp. 857–900.
[7] B. Ayati, A. R. A. Anderson, and G. F. Webb, *Computational methods and results for structured multiscale models of tumor invasion*, Multiscale Model. Simul., 5 (2006), pp. 1–20.
[8] S. B. Carter, *Principles of cell motility: The direction of cell movement and cancer invasion*, Nature, 208 (1965), pp. 1183–1187.
[9] S. B. Carter, *Haptotaxis and the mechanism of cell motility*, Nature, 213 (1967), pp. 256–260.
[10] L. Corrias, B. Perthame, and H. Zaag, *A chemotaxis model motivated by angiogenesis*, C. R. Math. Acad. Sci. Paris, 336 (2003), pp. 114–146.
[11] L. Corrias, B. Perthame, and H. Zaag, *Global solutions of some chemotaxis and angiogenesis systems in high space dimensions*, Milan J. Math., 72 (2004), pp. 159–186.
[12] M. A. Fontelos, A. Friedman, and B. Hu, *Mathematical analysis of a model for the initiation of angiogenesis*, SIAM J. Math. Anal., 33 (2002), pp. 1330–1355.
[13] A. Friedman and J. I. Tello, *Stability of solutions of chemotaxis equations in reinforced random walks*, J. Math. Anal. Appl., 272 (2002), pp. 138–163.
[14] F. R. Guarguaglini and R. Natalini, *Global existence of solutions to a nonlinear model of sulphation phenomena in calcium carbonate stones*, Nonlinear Anal. Real World Appl., 6 (2005), pp. 477–494.
[15] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer, Berlin, New York, 1981.
[16] D. Horstmann, *From* 1970 *until present: The Keller-Segel model in chemotaxis and its consequences* I, Jahresber. Deutsch. Math.-Verein., 105 (2003), pp. 103–165.
[17] H. A. Levine and B. D. Sleeman, *A system of reaction diffusion equations arising in the theory of reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 683–730.
[18] K. Osaki and A. Yagi, *Finite dimensional attractor for one-dimensional Keller-Segel equations*, Funkcial. Ekvac., 44 (2001), pp. 441–469.
[19] H. G. Othmer and A. Stevens, *Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks*, SIAM J. Appl. Math., 57 (1997), pp. 1044–1081.
[20] M. Rascle, *On a system of nonlinear strongly coupled partial differential equations arising in biology*, in Ordinary and Partial Differential Equations, W. N. Everitt and B. D. Sleeman, eds., Lecture Notes in Math. 846, Springer-Verlag, New York, 1980, pp. 290–298.
[21] M. Rascle and C. Ziti, *Finite time blow-up in some models of chemotaxis*, J. Math. Biol., 33 (1995), pp. 388–414.
[22] F. Rothe, *Global Solutions of Reaction-Diffusion Systems*, Lecture Notes in Math. 1072, Springer, Berlin, Heidelberg, New York, Tokyo, 1984.
[23] A. Stevens, *The derivation of chemotaxis equations as limit dynamics of moderately interacting stochastic many-particle systems*, SIAM J. Appl. Math., 61 (2000), pp. 183–212.
[24] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., Johann Ambrosius Barth, Heidelberg, Leipzig, 1995.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1715–1732

THE DISTRIBUTION OF SURFACE SUPERCONDUCTIVITY ALONG THE BOUNDARY: ON A CONJECTURE OF X. B. PAN*

YANIV ALMOG† AND BERNARD HELFFER‡

Abstract. We consider the Ginzburg–Landau model of superconductivity in two dimensions in the large κ limit. For applied magnetic fields weaker than the onset field H_{C_3} but greater than H_{C_2} it is well known that the superconductivity order parameter decays exponentially fast away from the boundary. It has been conjectured by X. B. Pan that this surface superconductivity solution converges pointwise to a constant along the boundary. For applied fields that are in some sense between H_{C_2} and H_{C_3}, we prove that the solution indeed converges to a constant but in a much weaker sense.

Key words. superconductivity, surface, Ginzburg–Landau

AMS subject classifications. 82D55

DOI. 10.1137/050636796

1. Introduction.

The Ginzburg–Landau energy functional of superconductivity is given in the form

$$\mathcal{J}(\Psi, A) = \int_\Omega \left(-|\Psi|^2 + \frac{|\Psi|^4}{2} + |h - h_{ex}|^2 + \left| \left(\frac{i}{\kappa}\nabla + A \right) \Psi \right|^2 \right) dx_1 dx_2, \tag{1.1}$$

in which $\Omega \subset\subset \mathbb{R}^2$ is smooth, and Ψ is the (complex) superconducting order parameter, such that $|\Psi|$ varies from $|\Psi| = 0$ (when the material is at a normal state) to $|\Psi| = 1$ (for the purely superconducting state). The magnetic vector potential is denoted by A (the magnetic field is, then, given by $h = \nabla \times A$), h_{ex} is the constant applied magnetic field, and κ is the Ginzburg–Landau parameter which is a property of the material. The functional $\mathcal{J}$ is invariant under the gauge transformation

$$\Psi \to e^{i\kappa\eta}\Psi,\ A \to A + \nabla\eta, \tag{1.2}$$

where η is a smooth function. We focus here on the properties, for a given h_{ex}, of the global minimizers[1] (Ψ_κ, A_κ) of $\mathcal{J}$ in $H^1(\Omega, \mathbb{C}) \times H^1(\Omega, \mathbb{R}^2)$ for type II superconductors (for which $\kappa > 1/\sqrt{2}$). Note that every global minimizer actually represents an orbit of minimizers associated to the group of transformations (1.2).

1.1. The onset of superconductivity.

It is known both from experiments [18] and rigorous analysis [13] that for a sufficiently strong magnetic field the normal state ($\Psi \equiv 0$, $h = h_{ex}$) would prevail. If the field is then decreased, there is a critical field, depending on the sample's geometry, where the material would enter the superconducting state. For samples with boundaries, this field is known as the onset

*Received by the editors July 25, 2005; accepted for publication (in revised form) July 3, 2006; published electronically February 15, 2007.
http://www.siam.org/journals/sima/38-6/63679.html
†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (almog@math.lsu.edu).
‡Laboratoire de Mathématiques, Université de Paris XI (Paris-Sud), 91405 Orsay, France (bernard.helffer@math.u-psud.fr).

[1]We could have actually written more precisely $(\Psi_{\kappa,h_{ex}}, A_{\kappa,h_{ex}})$ but will omit the reference to h_{ex} in order to avoid cumbersome notation.

critical field (or nucleation field) and is called H_{C_3}. This leads to the definition (cf. [19, 15, 10], for instance)

$$H_{C_3}(\kappa) = \inf\{h_{ex} > 0 : (0, \hat{A}) \text{ is the unique global minimizer of } \mathcal{J}\}, \tag{1.3}$$

where $\hat{A} : \Omega \to \mathbb{R}^2$ satisfies $\nabla \times \hat{A} = h_{ex}$. The minimizer $(0, \hat{A})$ is unique in the sense that any other minimizer is gauge equivalent to it, i.e., it should be in the form $(0, \hat{A} + \nabla\eta)$. We note that for our choice of scaling in (1.1) we have $H_{C_3} \sim \frac{\kappa}{\beta_0}$ as $\kappa \to \infty$ for smooth Ω [17], where β_0 is approximately 0.59.

The simplest case in which the bifurcation from the normal state $(0, \hat{A})$ to the superconducting one was described is the case of a half-plane [20]. The analysis in this case is one dimensional: the linearized Ginzburg–Landau equations were solved on $\mathbb{R}_+$. A similar situation occurs in two dimensions: it was proved in [17] and [7] that the bifurcating mode in $\mathbb{R}^2_+$ is one dimensional and that the value of H_{C_3} is exactly the same as in the one-dimensional case.

In addition, Saint-James and de Gennes [20] found that superconductivity appears first near the boundary for a half-plane, i.e., the order parameter Ψ_κ decays exponentially fast away from the boundary. This phenomenon, which appears only in the presence of boundaries, is therefore called surface superconductivity. It was later proved for general two-dimensional domains with smooth boundaries [17, 7], that as the domain's scale tends to infinity the onset field tends to de Gennes' value, and that Ψ_κ decays exponentially fast away from the boundary.

Another related problem that has been considered in the literature is the distribution of $|\Psi_\kappa|$ along the boundary near the critical field. In [4], this distribution was formally obtained. This led to the conjecture that $|\Psi_\kappa|$ should be maximal at the point of maximal curvature along the boundary. This was indeed proved a few years later [15, 16, 11, 12]. Furthermore, it was shown that Ψ_κ decays exponentially fast away from the points of maximal curvature along the boundary.

1.2. Weakly nonlinear analysis. Suppose now that h_{ex} is further decreased below H_{C_3}. While the minimizer Ψ_κ still decays exponentially fast away from the boundary much after the nucleation in the highly nonlinear regime when $\kappa < h_{ex} < H_{C_3}$ [1, 19, 2], the exponential decay along the boundary disappears quite rapidly as h_{ex} decreases. More precisely, if we introduce the distance to the nucleation field ρ by

$$\rho(\kappa) = H_{C_3}(\kappa) - h_{ex},$$

then exponential rate of decay along the boundary (far from the points of maximal curvature) is guaranteed only when

$$\rho(\kappa) \xrightarrow[\kappa\to\infty]{} 0.$$

Furthermore, it was proved in [10] that if ρ satisfies

$$\lim_{\kappa\to\infty} \rho(\kappa) = \infty; \quad \lim_{\kappa\to\infty} \frac{\rho(\kappa)}{\kappa^{1/2}} = 0,$$

then there exists $u \in \mathcal{S}(\overline{\mathbb{R}^+})$ such that

$$\int_\Omega \left| |\Psi_\kappa(x)|^2 - \frac{\rho}{\kappa} u\left(\frac{\kappa}{\sqrt{\lambda}}\, t(x)\right)\right|^2 dx = o(\rho\kappa^{-3}), \tag{1.4}$$

where $t = d(x, \partial\Omega)$, $\lambda = \kappa/h_{ex}$.

This leaves open the situation when $\rho(\kappa)/\kappa^{\frac{1}{2}}$ does not tend to 0 as $\kappa \to \infty$ and in particular becomes of the order of $\kappa^{\frac{1}{2}}$. It is this last case which will be considered in this article.

1.3. Highly nonlinear analysis: Pan's conjecture. Given some $\lambda \in]\beta_0, 1[$, let $(\kappa_n, h_{ex}^n)_{n\in\mathbb{N}}$ denote a pair of sequences satisfying

$$\lim_{n\to+\infty} \kappa_n = \infty; \quad \lim_{n\to+\infty} \frac{\kappa_n}{h_{ex}^n} = \lambda.$$

In the above $\beta_0 = \lim_{\kappa\to\infty} \kappa/H_{C_3}(\kappa)$ (we provide a better definition of β_0 in the next section). In [19, Conjecture 1], Pan conjectures the existence of a function $]\beta_0, 1[\ni \lambda \mapsto C(\lambda) \in \mathbb{R}^+$ such that, for any sequence as above,

$$|\Psi_{\kappa_n}(x)| \to C(\lambda)\ \forall x \in \partial\Omega. \tag{1.5}$$

While the conjecture appears to be correct in its essence—any minimizer, as the results in [10] and in the present contribution suggest, does tend in some weak sense to a constant along the boundary—we believe that either the convergence assumed in (1.5) cannot be uniform, or else that the global minimizer must be discontinuous in h_{ex} and κ. Let us sketch the heuristic arguments supporting this belief. We first write the Euler–Lagrange equations associated with (1.1) (or the Ginzburg–Landau equations):

$$\left(\frac{i}{\kappa}\nabla + A_\kappa\right)^2 \Psi_\kappa = \Psi_\kappa \left(1 - |\Psi_\kappa|^2\right), \tag{1.6a}$$

$$-\nabla \times \nabla \times A_\kappa = \frac{i}{2\kappa}\left(\Psi_\kappa^* \nabla\Psi_\kappa - \Psi_\kappa \nabla\Psi_\kappa^*\right) + |\Psi_\kappa|^2 A_\kappa. \tag{1.6b}$$

If $|\Psi_\kappa| > 0$ for all $x \in \partial\Omega$ (and this is indeed the case if we assume uniform convergence in (1.5)), then we can divide (1.6b) by $|\Psi_\kappa|^2$ and integrate over $\partial\Omega$ (the measure on $\partial\Omega$ being denoted by ds), to obtain the existence of an integer $N(\Psi_\kappa)$ such that

$$\int_{\partial\Omega} \frac{\nabla \times (h_\kappa - h_{ex})}{|\Psi|^2} ds + \int_\Omega h_\kappa\, dx = \frac{2\pi}{\kappa} N(\Psi_\kappa),$$

where $h_\kappa = \nabla \times A_\kappa$ is the induced magnetic field. The integer $N(\Psi_\kappa) \in \mathbb{Z}$ is the winding number (or the degree) of Ψ_κ around $\partial\Omega$, which is invariant under the transformation (1.2) since η must be smooth.

In [9] it is proved that Ψ_κ vanishes at isolated points or curves which should end on $\partial\Omega$. If $|\Psi_\kappa|$ does not vanish on the boundary, as implied by (1.5), then it is clear that Ψ_κ can vanish only at isolated points. Thus, we can conclude that $N(\Psi_\kappa)$ is the number of vortices of Ψ_κ, including multiplicities, in Ω.

In [16], it is proved (see [19] for an extension to the case which is considered here) that, for any $\epsilon_0 > 0$, there exist $C > 0$ and κ_0, such that, if $\kappa/h_{ex} \in [\beta_0 + \epsilon_0, 1 - \epsilon_0]$ and $\kappa \geq \kappa_0$, then

$$\|h - h_{ex}\|_{L^\infty(\Omega)} + \|\nabla(h_\kappa - h_{ext})\|_{L^\infty(\Omega)} \leq C.$$

Hence there exists a constant C such that

$$\left| N(\Psi_\kappa) - \kappa h_{ex} \frac{|\Omega|}{2\pi} \right| \leq C\kappa. \tag{1.7}$$

Suppose now that the minimizer is unique when varying h_{ex} and κ as above. It is in this case reasonable to think that (Ψ_κ, A_κ) varies continuously. If there exist κ_0, $\epsilon_1 > 0$ and $C > 0$ such that

$$|\Psi_\kappa| \geq \frac{1}{C} \text{ on } \partial\Omega \; \forall \kappa \geq \kappa_0 \text{ s. t. } \frac{\kappa}{h_{ex}} \in [\lambda - \epsilon_1, \lambda + \epsilon_1]$$

(which would be a consequence of a uniform version of (1.5)), then $N(\Psi_\kappa)$ must be fixed, by continuity, for all $\kappa \geq \kappa_0$ such that $\kappa/h_{ex} \in [\lambda - \epsilon_1, \lambda + \epsilon_1]$, and this is in contradiction with (1.7).

The above argument works not only for (ψ_κ, A_κ) but for any solution of (1.6). If indeed critical points of (1.1) are continuous functions of κ and h_{ex} in this regime, then (1.7) would contradict another conjecture of Pan (Conjecture 2 in [19]), implying that any solution of (1.6) converges to a constant along the boundary when $\kappa \to \infty$ and $\kappa/h_{ex} \in (\beta_0, 1)$. While the existence of continuous branches of critical points appears to be reasonable, two counterexamples come to mind while discussing the continuity of the global minimizer.

1. Serfaty [21] proves, for much lower external fields, that a large number of local minimizers of (1.1) in a disc, characterized by different winding numbers, exist for sufficiently large κ whenever $\frac{1}{C}\kappa^{-1} \leq h_{ex} \leq C\kappa^{\alpha-1}$ for some $0 < \alpha < 1$. In this regime of applied magnetic field values the magnetic field is nonuniform inside the domain, and hence the vortices are kept near the disc's center, which minimizes the magnetic field term in (1.1).
   While in the present case h_{ex} and κ have the same order, if we allow for an $O(1)$ change in the applied magnetic field we might still encounter a global minimizer which turns into a local minimizer (or a critical point) and vice versa. Thus, this result suggests that the contradiction between (1.7) and the convergence to a uniform constant along the boundary might be explained by arguing that the global minimizer is discontinuous. However, unlike the case discussed in [21], no equivalent mechanism which keeps the vortices away from the boundary is presently known as the magnetic field uniformly converges to h_{ex} in Ω [2].
2. Bauman, Phillips, and Tang [3] found radially symmetric solutions of the linearized version of (1.6). These solutions are characterized by a "fat" vortex at the disc's center. The degree of the vortex is determined, to leading order in the large κ limit, by the magnetic flux through the disc. Thus, there is a sequence of critical flux values where the bifurcating mode changes its winding number. It is shown in [3] that the bifurcating mode is locally stable near the bifurcation for κ large enough.
   Based on the results in [3] one can argue that the minimizer undergoes an abrupt change when the flux varies around one of the above critical values (and when κ is appropriately tuned to guarantee that weakly nonlinear analysis still holds). However, this result seems to follow from the special geometry, and, in general, for different geometries or away from the linear regime, nothing would hold the vortices in the center.

1.4. Statement of the main result. In the present contribution we focus on the case

$$\lim_{\kappa\to\infty} \frac{h_{ex}(\kappa)}{\kappa} = \frac{1}{\lambda},$$

with λ close to β_0. We prove the following theorem.

THEOREM 1.1. *Let $\delta > 0$ be sufficiently small, so that $t = d(x, \partial\Omega)$ is a smooth function of x for $0 \le t \le \delta$, and let*

$$\Omega_\delta = \left\{ x \in \bar{\Omega} \,:\, d(x, \partial\Omega) \le \delta \right\}.$$

Then there exist $\epsilon > 0$, a function

$$[0, +\infty[\times]\beta_0, \beta_0 + \epsilon[\ni (\tau, \lambda) \mapsto U(\tau, \lambda) \in \mathbb{R}^+,$$

a constant $C > 0$, and κ_0, such that, for $\kappa \ge \kappa_0$ and $h_{ex} = \frac{\kappa}{\lambda}$ with $\lambda \in]\beta_0, \beta_0 + \epsilon[$,

$$\int_{\Omega_\delta} \left[|\Psi_\kappa(x)|^2 - U\left(\frac{\kappa}{\sqrt{\lambda}} t(x), \lambda \right) \right]^2 dx_1 dx_2 \le \frac{C}{\kappa^2}, \tag{1.8a}$$

$$\int_{\partial\Omega} \left[\left| \Psi_\kappa \big/_{\partial\Omega} \right|^2 - U(0, \lambda) \right]^2 ds \le \frac{C}{\kappa^{1/2}}. \tag{1.8b}$$

The function $U(\tau, \lambda)$ is defined for $\tau \in \overline{\mathbb{R}^+}$ by

$$U(\tau, \lambda) = |f_{\zeta(\lambda)}(\tau; \lambda)|^2,$$

where $f_z(\tau; \lambda)$ and $\zeta(\lambda)$ are associated to minimizers of a family of one-dimensional problems, which will be analyzed in section 2. The second statement in the theorem gives the $L^2(\partial\Omega)$ convergence of $|\Psi_\kappa|^2$ to a constant and is consequently a weak form of Pan's conjecture.

The rest of the contribution is arranged as follows.

In section 2 we consider a one-dimensional differential operator and prove that it is positive for $\beta_0 < \lambda < \beta_0 + \epsilon$. In section 3 we use the results of section 2 to analyze a simplified two-dimensional minimization problem, which was proved in [19] to be a good approximation of the full Ginzburg–Landau model for $\beta_0 < \lambda < 1$. The last section gives the proof of Theorem 1.1.

2. A one-dimensional problem. Let

$$\beta(z) = \inf_{\phi \in H^1_{mag}(]0,\infty[) \setminus \{0\}} \frac{\int_0^\infty |\phi'(\tau)|^2 + (\tau + z)^2 |\phi(\tau)|^2 \, d\tau}{\int_0^\infty |\phi(\tau)|^2 \, d\tau}. \tag{2.1}$$

Here

$$H^1_{mag}(]0, +\infty[) = \{ u \in L^2(]0, +\infty[) \,,\; u' \in L^2(]0, +\infty[) \text{ and } \tau u \in L^2(]0, +\infty[) \}.$$

It is well known (see [5]) that $\beta(z)$ has a unique local minimum at $z_0 < 0$, where

$$\beta(z_0) = \beta_0 = z_0^2.$$

Furthermore, $\beta(z) \xrightarrow[z \to \infty]{} \infty$ and $\beta(z) \xrightarrow[z \to -\infty]{} 1$. Clearly, for $\beta_0 < \lambda < 1$ there exist $z_1(\lambda) < z_0 < z_2(\lambda)$, such that

$$]z_1(\lambda), z_2(\lambda)[= \beta^{-1}([\beta_0, \lambda[).$$

It is also easy to show [6] that

$$\beta''(z_0) = -2 z_0 \phi^2(0) > 0, \tag{2.2}$$

where ϕ is the minimizer of (2.1) whose $L^2(\mathbb{R}_+)$ norm is unity.

Let $f_z(\tau;\lambda)$ denote the minimizer of

$$\mathcal{E}_{z,\lambda}(\phi) = \int_0^\infty |\phi'(\tau)|^2 + (\tau+z)^2|\phi(\tau)|^2 + \frac{\lambda}{2}|\phi(\tau)|^4 - \lambda|\phi(\tau)|^2\, d\tau \tag{2.3}$$

in $H^1_{mag}(]0,\infty[)$. The Euler–Lagrange equation associated with (2.3) is

$$-f_z''(\tau;\lambda) + (\tau+z)^2 f_z(\tau;\lambda) = \lambda f_z(\tau;\lambda)(1-f_z(\tau;\lambda)^2). \tag{2.4}$$

It has been proved in [19, Theorems 3.1 and 3.3] that whenever $z_1(\lambda) < z < z_2(\lambda)$, there exists a unique positive global minimizer to (2.3). Furthermore, let

$$b(z,\lambda) = \inf_{\varphi\in H^1_{mag}(]0,\infty[)} \mathcal{E}_{z,\lambda}(\varphi). \tag{2.5}$$

Then there exists $\zeta(\lambda)\in\,]z_1(\lambda), z_2(\lambda)[$, where $z\mapsto b(z,\lambda)$ attains its minimum over $\mathbb{R}$,

$$b(\zeta(\lambda),\lambda) = \inf_z b(z,\lambda).$$

Moreover,

$$\int_0^\infty (\tau+\zeta(\lambda))|f_{\zeta(\lambda)}(\tau;\lambda)|^2\, d\tau = 0. \tag{2.6}$$

Remark 2.1. Note that when $z\notin\,]z_1(\lambda), z_2(\lambda)[$, then $b(z,\lambda)=0$, and the minimizer of $\mathcal{E}_{z,\lambda}$ is the 0-function. In particular,

$$b(\zeta(\lambda),\lambda) < 0 \ \text{ if } \lambda > \beta_0.$$

The following lemma will play a crucial role in the analysis of the two-dimensional problem in section 3.

LEMMA 2.2. *Let*

$$\gamma(\alpha,\lambda) = \inf_{\phi\in H^1_{mag}(]0,\infty[)} \frac{\int_0^\infty |\phi'(\tau)|^2 + (\tau+\zeta+\alpha)^2|\phi(\tau)|^2 - \lambda(1-f_\zeta(\tau;\lambda)^2)|\phi(\tau)|^2\, d\tau}{\int_0^\infty |\phi(\tau)|^2\, d\tau}, \tag{2.7}$$

with $\zeta=\zeta(\lambda)$.

Then there exists $\epsilon>0$ *such that, for* $\lambda\in[\beta_0,\beta_0+\epsilon[$,

$$\min_{\alpha\in\mathbb{R}} \gamma(\alpha,\lambda) = 0. \tag{2.8}$$

Proof. We divide the proof into three steps.

Step 1. $\gamma(0,\lambda) = \gamma_\alpha(0,\lambda) = 0$.

Let $\mathbb{R}_+ \ni \tau\mapsto u(\tau;\alpha,\lambda)$ denote the positive minimizer of (2.7), whose $L^2(\mathbb{R}_+)$ norm is one. Then u satisfies

$$-u''(\tau;\alpha,\lambda) + (\tau+\alpha+\zeta)^2 u(\tau;\alpha,\lambda) - \lambda(1-f_\zeta(\tau;\lambda)^2)u(\tau;\alpha,\lambda) = \gamma(\alpha,\lambda)u(\tau;\alpha,\lambda), \tag{2.9a}$$

$$u'(0;\alpha,\lambda) = 0. \tag{2.9b}$$

For $\alpha = 0$, we multiply (2.9a) by f_ζ and integrate over $\mathbb{R}_+$ to obtain

$$\gamma(0,\lambda)\int_0^\infty f_\zeta(\tau;\lambda)u(\tau;\alpha,\lambda)\,d\tau = 0.$$

Since both u and f_ζ are positive, we have using (2.4),

$$\gamma(0,\lambda) = 0, \quad u(\tau,0,\lambda) = \frac{f_\zeta(\tau;\lambda)}{\|f_\zeta\|_2}, \tag{2.10}$$

where, for $p \in [1,+\infty]$, $\|\cdot\|_p = \|\cdot\|_{L^p(\mathbb{R}_+)}$. Next, we differentiate (2.9) with respect to α to obtain, having in mind (2.4),

$$-u''_\alpha + (\tau+\alpha+\zeta)^2 u_\alpha - \lambda(1-f_\zeta^2)u_\alpha = \gamma u_\alpha + \gamma_\alpha u - 2(\tau+\alpha+\zeta)u, \tag{2.11a}$$

$$u'_\alpha(0) = 0, \tag{2.11b}$$

where $u_\alpha(\tau;\alpha,\lambda) = (\frac{\partial u}{\partial\alpha})(\tau;\alpha,\lambda)$ and $\gamma_\alpha(\alpha,\lambda) = \frac{\partial\gamma}{\partial\alpha}(\alpha,\lambda)$. Multiplying (2.11a) by u and integrating by parts, we obtain

$$\gamma_\alpha(\alpha,\lambda) = 2\int_0^\infty (\tau+\alpha+\zeta(\lambda))|u(\tau;\alpha,\lambda)|^2\,d\tau. \tag{2.12}$$

In view of (2.6) and (2.10), we thus have

$$\gamma_\alpha(0,\lambda) = 0. \tag{2.13}$$

Step 2.

$$\exists\epsilon_1 > 0 :\ \lambda < \beta_0 + \epsilon_1 \Rightarrow \gamma_{\alpha\alpha}(0,\lambda) > \frac{1}{2}\beta''(z_0) > 0. \tag{2.14}$$

To prove the above statement we notice that $z_1(\lambda) \uparrow z_0$ and $z_2(\lambda) \downarrow z_0$ as $\lambda \to \beta_0$. Hence, since $z_1(\lambda) < \zeta(\lambda) < z_2(\lambda)$, we have

$$\zeta(\lambda) \xrightarrow[\lambda\to\beta_0]{} z_0. \tag{2.15}$$

Moreover, one gets from the fact that f_z is a minimizer the property that

$$\mathcal{E}_{z,\lambda}(f_z) \le 0.$$

From this inequality and (2.1) we easily obtain

$$\frac{1}{2}||f_z||_4^4 \le \frac{(\lambda-\beta_0)}{\lambda}||f_z||_2^2,$$

$$||(\tau+z)f_z||_2^2 \le \lambda||f_z||_2^2, \tag{2.16}$$

and

$$||f_z||_{H^1}^2 \le (\lambda+1)||f_z||_2^2.$$

Let $z = \zeta(\lambda)$. Since $|\zeta(\lambda)|$ is bounded in some right semineighborhood of β_0, it follows immediately from (2.16) that for R large enough we get

$$||f_\zeta||_2^2 \le 2||f_\zeta||_{L^2(]0,R[)}^2.$$

We now observe that

$$||f_\zeta||_4^4 \leq C(\lambda-\beta_0)||f_\zeta||^2_{L^2(]0,R[)} \leq C(\lambda-\beta_0)R^{\frac{1}{2}}||f_\zeta||_4^2.$$

This first gives that

$$||f_\zeta||_4 \leq \tilde{C}(\lambda-\beta_0)^{\frac{1}{2}},$$

and hence that

$$||f_\zeta||_2 \leq \hat{C}(\lambda-\beta_0)^{\frac{1}{2}}.$$

By interpolation, we obtain

$$||f_\zeta||_\infty \leq C||f_\zeta||_2^{\frac{1}{2}}\ ||f'_\zeta||_2^{\frac{1}{2}} \leq C'(\lambda-\beta_0)^{\frac{1}{4}}, \tag{2.17}$$

which implies that

$$\lim_{\lambda\to\beta_0} \|f_{\zeta(\lambda)}(\,\cdot\,;\lambda)\|_\infty = 0.$$

Substituting the above and (2.15) into (2.7) yields

$$\gamma(\alpha,\lambda) \xrightarrow[\lambda\to\beta_0]{} \beta(\alpha+z_0),$$

where the convergence is uniform on every compact set in $\mathbb{R}$. Since γ is holomorphic in α, its derivatives must uniformly converge as well, and hence

$$\gamma_{\alpha\alpha}(\alpha,\lambda) \xrightarrow[\lambda\to\beta_0]{} \beta''(\alpha+z_0), \tag{2.18}$$

from which (2.14) easily follows. We note that a tedious calculation shows that

$$\gamma_{\alpha\alpha}(0,\lambda) = -2\zeta\frac{f_\zeta^2(0;\lambda)}{\|f_\zeta\|_2^2} + \frac{6\lambda^2}{\|f_\zeta\|_2^2}\int_0^\infty f_\zeta^6(\tau;\lambda)\,d\tau\ - \frac{2\lambda}{3}\int_0^\infty f_\zeta^4(\tau;\lambda)[\lambda-(\tau+\zeta)^2]d\tau,$$

with $\zeta=\zeta(\lambda)$, from which one can easily prove (2.14) as well.

From (2.14) we obtain that

$$\exists\alpha_0>0:\ \lambda<\beta_0+\epsilon_1 \Rightarrow \gamma(\alpha,\lambda)\geq 0\ \forall|\alpha|\leq\alpha_0.$$

The last step would thus be to prove the above statement for $|\alpha|>\alpha_0$.

Step 3. Proof of (2.8).

From the definition of γ (2.7), it follows that

$$\gamma(\alpha,\lambda)\geq\beta(\zeta+\alpha)-\lambda.$$

Clearly, for any $\alpha_1>0$, there exists $\epsilon_2>0$, such that, if $\lambda\leq\beta_0+\epsilon_2$, then $[z_1(\lambda),z_2(\lambda)]\subset[z_0-\alpha_1,z_0+\alpha_1]$. We now take $\alpha_1=\alpha_0$. This gives that $\beta(\zeta+\alpha)\geq\lambda$ for all $|\alpha|\geq\alpha_0$, and (2.8) follows. □

3. On two-dimensional models on half cylinders. We can now prove the following theorem.

THEOREM 3.1. *For $\omega \in]0,+\infty[$ and $\lambda \in [\beta_0,+\infty[$, let us consider the functional*

$$\mathcal{H}_\omega \ni \psi \mapsto E_\omega(\psi,\lambda) = \int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty \left[|(i\nabla + \xi_1 \hat{i}_2)\psi|^2 + \frac{1}{2}\lambda|\psi|^4 - \lambda|\psi|^2\right] d\xi_1 d\xi_2, \tag{3.1}$$

where

$$|(i\nabla + \xi_1\hat{i}_2)\psi|^2 = |i\partial_{\xi_1}\psi|^2 + |(i\partial_{\xi_2} + \xi_1)\psi|^2,$$

and

$$\mathcal{H}_\omega = \Big\{\psi \in H^1_{mag}\big(\mathbb{R}_+\times]-L,L[,\mathbb{C}\big)\ \forall L>0 \ \Big|\ \exists z\in\mathbb{R}:\ \psi(\xi_1,\xi_2+2\pi/\omega) = e^{-iz\frac{2\pi}{\omega}}\psi(\xi_1,\xi_2)\Big\}.$$

Let ψ_λ be the function

$$(\mathbb{R}_+\times\mathbb{R}) \ni (\xi_1,\xi_2) \mapsto \psi_\lambda(\xi_1,\xi_2) := e^{-i\zeta(\lambda)\xi_2} f_{\zeta(\lambda)}(\xi_1;\lambda). \tag{3.2}$$

Then there exists $\epsilon > 0$ such that

$$E_\omega(\psi,\lambda) \geq E_\omega(\psi_\lambda,\lambda)\ \forall\lambda\in]\beta_0,\beta_0+\epsilon[,\ \forall\omega>0, \quad \text{and } \forall\psi\in\mathcal{H}_\omega. \tag{3.3}$$

Remark 3.2. Clearly ψ_λ is in $\mathcal{H}_\omega$ (take $z=\zeta(\lambda)$). Hence, the theorem states that ψ_λ is the global minimizer of E_ω in $\mathcal{H}_\omega$.

Proof. Consider first functions in $\mathcal{H}_\omega$ which are given in the form

$$(\xi_1,\xi_2) \mapsto \psi(\xi_1,\xi_2) := f_\zeta(\xi_1;\lambda)e^{-i\zeta\xi_2}v, \tag{3.4}$$

with v periodic,

$$v(\xi_1,\xi_2) = v(\xi_1,\xi_2+2\pi/\omega), \tag{3.5}$$

and

$$\zeta = \zeta(\lambda).$$

Then

$$E_\omega(\psi,\lambda) = \int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty \left[|(i\nabla + (\xi_1+\zeta)\hat{i}_2)f_\zeta v|^2 + \frac{1}{2}\lambda|f_\zeta v|^4 - \lambda|f_\zeta v|^2\right] d\xi_1 d\xi_2.$$

Clearly,

$$\begin{aligned}\int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty |(i\nabla + (\xi_1+\zeta)\hat{i}_2)f_\zeta v|^2\, d\xi_1 d\xi_2 \\ = \int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty \Big[|v|^2\big[|f_\zeta'|^2 + (\xi_1+\zeta)^2|f_\zeta|^2\big] + f_\zeta^2|\nabla v|^2 + \frac{1}{2}(f_\zeta^2)'\frac{\partial}{\partial\xi_1}(|v|^2) \\ + i(\xi_1+\zeta)f_\zeta^2\left(\bar{v}\frac{\partial v}{\partial\xi_2} - v\frac{\partial\bar{v}}{\partial\xi_2}\right)\Big] d\xi_1 d\xi_2.\end{aligned}$$

Furthermore, integration by parts and (2.4) yield

$$\int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty \left[|v|^2\left[|f_\zeta'|^2+(\xi_1+\zeta)^2|f_\zeta|^2\right]+\frac{1}{2}(f_\zeta^2)'\frac{\partial}{\partial\xi_1}(|v|^2)\right]d\xi_1 d\xi_2 = \lambda\int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty |v|^2 f_\zeta^2(1-f_\zeta^2)\,d\xi_1 d\xi_2. \tag{3.6}$$

Hence,

$$\begin{aligned}\Delta E_\omega &= E_\omega(\psi,\lambda)-E_\omega(f_\zeta e^{-i\zeta\xi_2},\lambda)\\ &= \int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty f_\zeta^2\left[|\nabla v|^2+i(\xi_1+\zeta)\left(\bar v\frac{\partial v}{\partial\xi_2}-v\frac{\partial\bar v}{\partial\xi_2}\right)\right]d\xi_1 d\xi_2\\ &\quad+\frac{1}{2}\int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty f_\zeta(\xi_1;\lambda)^4(1-|v(\xi_1,\xi_2)|^2)^2\,d\xi_1 d\xi_2.\end{aligned}$$

Using (3.5), we can write

$$v(\xi_1,\xi_2)=\sum_{n=-\infty}^{\infty} v_n(\xi_1)e^{in\omega\xi_2}.$$

Then

$$\Delta E_\omega = \sum_{n=-\infty}^{\infty}\int_0^\infty f_\zeta(\xi_1;\lambda)^2\left[|v_n'(\xi_1)|^2+(n^2\omega^2+2n\omega\xi_1)|v_n(\xi_1)|^2\right]d\xi_1 + \frac{1}{2}\int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty f_\zeta(\xi_1;\lambda)^4(1-|v(\xi_1,\xi_2)|^2)^2\,d\xi_1 d\xi_2. \tag{3.7}$$

Consider now the functional

$$u\mapsto \mathcal{J}(u,\alpha)=\int_0^\infty |f_\zeta(\xi_1;\lambda)|^2\left[|u'(\xi_1)|^2+[\alpha^2+2\alpha(\xi_1+\zeta)]|u(\xi_1)|^2\right]d\xi_1.$$

Substituting $w(\xi_1)=f_\zeta(\xi_1;\lambda)\,u(\xi_1)$ and utilizing (2.4), we obtain

$$\begin{aligned}\int_0^\infty |f_\zeta(\xi_1)|^2|u'(\xi_1)|^2\,d\xi_1 &= \int_0^\infty\left[-\left(w^2\frac{f_\zeta'}{f_\zeta}\right)'+w^2\frac{f_\zeta''}{f_\zeta}+|w'|^2\right]d\xi_1\\ &= \int_0^\infty\left[|w'|^2+[(\xi_1+\zeta)^2-\lambda(1-f_\zeta^2)]|w|^2\right]d\xi_1.\end{aligned}$$

Consequently,

$$\mathcal{J}(w/f_\zeta,\alpha)=\int_0^\infty\left(|w'|^2+[(\xi_1+\zeta+\alpha)^2-\lambda(1-f_\zeta^2)]|w|^2\right)d\xi_1\ge\gamma(\alpha,\lambda)\int_0^\infty|w|^2\,d\xi_1.$$

Combining the above with (3.7), we obtain

$$\Delta E_\omega\ge\sum_{n=-\infty}^{\infty}\gamma(n\omega,\lambda)\int_0^\infty f_\zeta^2|v_n|^2\,d\xi_1+\frac{1}{2}\int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty f_\zeta^4(1-|v|^2)^2\,d\xi_1 d\xi_2\ge 0,$$

which proves, using (2.8), inequality (3.3) for every function in $\mathcal{H}_\omega$ satisfying (3.4).

Note for later use that this implies

(3.8)
$$|E_\omega(\psi,\lambda) - E_\omega(f_\zeta\, e^{-i\zeta\xi_2},\lambda)| \geq \frac{1}{2}\int_{-\pi/\omega}^{\pi/\omega}\int_0^\infty f_\zeta(\xi_1;\lambda)^4(1-|v(\xi_1,\xi_2)|^2)^2\, d\xi_1 d\xi_2.$$

To prove (3.3) for all $\psi \in \mathcal{H}_\omega$, we consider now functions of the form

(3.9) $$(\xi_1,\xi_2) \mapsto \psi_0(\xi_1,\xi_2) = f_\zeta(\xi_1;\lambda)e^{-iz\xi_2}v, \text{ with } v(\xi_1,\xi_2) = v(\xi_1,\xi_2+2\pi/\omega).$$

Consider first the case when $\omega \in \mathbb{R}_+$ satisfies

(3.10) $$\frac{\zeta - z}{\omega} = \frac{p}{q} \text{ for some pair } (p,q) \in \mathbb{Z}\times\mathbb{N}.$$

Clearly, if ψ_0 satisfies (3.9) for some $\omega \in \mathbb{R}_+$, then it also satisfies (3.9) for $\omega/\hat{q}$, for every $\hat{q} \in (\mathbb{N}\setminus\{0\})$. Moreover, it is easy to show that

(3.11) $$E_{\omega/\hat{q}}(\psi_0) = \hat{q}E_\omega(\psi_0),\ E_{\omega/\hat{q}}(\psi_\lambda) = \hat{q}E_\omega(\psi_\lambda).$$

We now choose $\hat{q} = q$, and observe that, according to (3.10), $\hat{\omega} = \omega/q$ satisfies

(3.12) $$\frac{\zeta - z}{\hat{\omega}} \in \mathbb{Z}.$$

But in this case, ψ_0 admits the representation (3.4), and hence

$$E_{\hat{\omega}}(\psi_0) \geq E_{\hat{\omega}}(\psi_\lambda).$$

Coming back to ω and using (3.11), we have the proof of (3.3) when ω satisfies (3.10) (with the additional condition that z is fixed).

The proof of (3.3) in the general case follows immediately from the density of the rational numbers in $\mathbb{R}$. □

4. Surface superconductivity. Let $\mathcal{J}$ be given by (1.1). Let (Ψ_κ, A_κ) denote a minimizer of $\mathcal{J}$ in $H^1(\Omega,\mathbb{C})\times H^1(\Omega,\mathbb{R}^2)$. We prove in this section that $|\Psi_\kappa|^2$ is nearly a constant along the boundary, in $L^2(\partial\Omega)$ sense, as $\kappa\to\infty$, and for

$$\beta_0 < \lambda = \frac{\kappa}{h_{ex}} < \beta_0 + \epsilon,$$

where ϵ is defined in (2.8).

To this end we need to adapt the results in [19]. Then let

(4.1) $$x = \mathcal{F}(t,s)$$

denote a diffeomorphism from

$$D(\delta) = \{(s,t)\ :\ |s| \leq |\partial\Omega|/2,\ 0 \leq t \leq \delta\}$$

to

$$\Omega_\delta = \left\{x \in \bar{\Omega}\ :\ d(x,\partial\Omega) \leq \delta\right\}.$$

In the previous equation $t = d(x,\partial\Omega)$ and s denotes the arclength along $\partial\Omega$.

In order to formulate and prove the results of this section it is necessary to fix a specific gauge for (ψ_κ, A_κ). To this end we first define the magnetic potential P_κ to be the solution of

$$\begin{cases} \nabla \times P_\kappa = \nabla \times A_\kappa - h_{ex} & \text{in } \Omega, \\ \nabla \cdot P_\kappa = 0 & \text{in } \Omega, \\ P_\kappa \cdot \hat{\nu} = 0 & \text{in } \partial\Omega \end{cases} \tag{4.2}$$

(see, for example, [8] for the proof of existence of a unique solution for (4.2)). Moreover, the map associating the solution P_κ of (4.2) to the field $(h_\kappa - h_{ex})$ is linear continuous from $L^p(\Omega)$ into $W^{1,p}(\Omega)$ for any $p \in [1, +\infty[$, and, using the Sobolev injection theorem, one can show that

$$||P_\kappa||_{L^\infty(\Omega)} \leq C_\Omega \, ||\nabla \times A_\kappa - h_{ex}||_{L^\infty(\Omega)}. \tag{4.3}$$

Then let $e_1 = -\hat{\nu}$ denote an inward unit normal vector on $\partial\Omega$ and let e_2 denote a unit tangential vector. Further, let

$$g = \mathrm{Det}(D\mathcal{F}) = 1 - t\kappa_r(s),$$

where κ_r denotes the local curvature on $\partial\Omega$. Let (see [19]) F be any vector potential such that $\nabla \times F = h_{ex}$ and let a be defined by

$$a = a_1 e_1 + a_2 e_2 = [F \cdot e_1]e_1 + [gF \cdot e_2]e_2. \tag{4.4}$$

By [14], or the appendix in [12], there exists $\hat{A}_\kappa$ such that if we substitute $F = \hat{A}_\kappa$ in (4.4), we obtain

$$a_1(s,t) = 0; \quad a_2(s,t) = h_{ex}[c_2 + t - t^2\kappa_r(s)],$$

where

$$c_2 = \frac{|\Omega|}{|\partial\Omega|}.$$

The gauge in (1.2) is now fixed by the condition that the $\hat{A}_\kappa$ has in the new coordinates the normal form given above and that A_κ satisfies

$$A_\kappa - \hat{A}_\kappa = P_\kappa.$$

We now introduce the change of variables

$$(\xi_1, \xi_2) = \left(\frac{\kappa}{\sqrt{\lambda}} t, \, \frac{\kappa}{\sqrt{\lambda}} s \right) \tag{4.5}$$

and prove the following lemma.

LEMMA 4.1. *Let*

$$\tilde{\Psi}_\kappa(\xi_1, \xi_2) = \begin{cases} \Psi_\kappa(\frac{\sqrt{\lambda}}{\kappa}\xi_1, \frac{\sqrt{\lambda}}{\kappa}\xi_2) e^{-ic_2\xi_2} & \text{for } 0 \leq \xi_1 \leq \frac{\kappa}{\sqrt{\lambda}}\delta, \\ \Psi_\kappa(\kappa\delta/\sqrt{\lambda}, \xi_2) e^{-ic_2\xi_2} e^{-(\xi_1 - \kappa\delta/\sqrt{\lambda})} & \text{for } \xi_1 \geq \frac{\kappa}{\sqrt{\lambda}}\delta, \end{cases}$$

where $c_2 = c_2(\Omega)$, and let

$$\omega_\kappa = \frac{2\pi\sqrt{\lambda}}{\kappa|\partial\Omega|}.$$

Then, as κ tends to $+\infty$,

$$\mathcal{J}(\Psi_\kappa, A_\kappa) \geq \frac{1}{\kappa^2} E_{\omega_\kappa}(\tilde{\Psi}_\kappa, \lambda) + \mathcal{O}(1/\kappa^2). \tag{4.6}$$

We will later prove (see (4.18)), that $\left|E_{\omega_\kappa}(\tilde{\Psi}_\kappa, \lambda)\right| \geq C\kappa$, and hence the correction term on the right-hand side of (4.6) is much smaller than the first term as $\kappa \to \infty$.

Proof. In [2] (see also [16]) it was proved that for $\lambda < 1$, there exists $\mu > 0$ such that

$$|\nabla(\nabla \times A_\kappa)| \leq Ce^{-\mu\kappa d(x,\partial\Omega)}. \tag{4.7}$$

Consequently, for $x \in \Omega_\delta$, we have

$$|\nabla \times A_\kappa - h_{ex}|(x) \leq \int_0^{d(x,\partial\Omega)} |\nabla(\nabla \times A_\kappa)(t, s(x))| dt \leq C \int_0^\infty e^{-\mu\kappa t} dt.$$

Hence, there exists $C_1 > 0$ such that

$$\|\nabla \times A_\kappa - h_{ex}\|_{L^\infty(\Omega_\delta)} \leq \frac{C_1}{\kappa}.$$

In view of (4.7) we can state the above inequality for the $L^\infty(\Omega)$ norm of $\nabla \times A_\kappa - h_{ex}$, and thus (4.3) gives that, for some $C_2 > 0$,

$$\|A_\kappa - \hat{A}_\kappa\|_{L^\infty(\Omega)} \leq \frac{C_2}{\kappa}.$$

Hence, for some $C_3 > 0$,

$$\int_\Omega |\nabla \times A_\kappa - h_{ex}|^2 \, dx_1 dx_2 \leq \frac{C_3}{\kappa^2}, \tag{4.8}$$

and

$$\begin{aligned}
&\int_\Omega \left| \left(\frac{i}{\kappa}\nabla + A_\kappa \right) \Psi_\kappa \right|^2 dx = \int_\Omega \left| \left(\frac{i}{\kappa}\nabla + \hat{A}_\kappa \right) \Psi_\kappa \right|^2 dx \\
&+ \int_\Omega |\Psi_\kappa|^2 |A_\kappa - \hat{A}_\kappa|^2 \, dx \; + \; \int_\Omega (A_\kappa - \hat{A}_\kappa) \cdot \left[\frac{i}{\kappa} (\overline{\Psi_\kappa} \nabla \Psi_\kappa - \Psi_\kappa \nabla \overline{\Psi_\kappa}) + 2\hat{A}_\kappa \right] dx \\
&\geq \int_\Omega \left| \left(\frac{i}{\kappa}\nabla + \hat{A}_\kappa \right) \Psi_\kappa \right|^2 dx \; - 2||A_\kappa - \hat{A}_\kappa||_{L^\infty(\Omega)} \, ||\Psi_\kappa||_{L^\infty} \int_\Omega \left| \left(\frac{i}{\kappa}\nabla + \hat{A}_\kappa \right) \Psi_\kappa \right| dx.
\end{aligned}$$

In [2] it is shown that

$$|\Psi_\kappa| + \left| \left(\frac{i}{\kappa}\nabla + \hat{A}_\kappa \right) \Psi_\kappa \right| \leq Ce^{-\mu\kappa d(x,\partial\Omega)} \tag{4.9}$$

for some $\mu > 0$ when $\lambda < 1$. Hence,

$$\int_\Omega \left| \left(\frac{i}{\kappa}\nabla + A_\kappa \right) \Psi_\kappa \right|^2 dx \geq \int_\Omega \left| \left(\frac{i}{\kappa}\nabla + \hat{A}_\kappa \right) \Psi_\kappa \right|^2 dx \; - \frac{C}{\kappa^2}. \tag{4.10}$$

Combining (4.8) and (4.10) we obtain

$$\mathcal{J}(\Psi_\kappa, A_\kappa) \geq \int_\Omega \left(\left| \left(\frac{i}{\kappa}\nabla + \hat{A}_\kappa \right) \Psi_\kappa \right|^2 + \frac{1}{2}|\Psi_\kappa|^4 - |\Psi_\kappa|^2 \right) dx \; - \frac{C}{\kappa^2}. \tag{4.11}$$

Using the coordinates (4.1) we obtain

$$\int_{\Omega_\delta} \left(\left| \left(\frac{i}{\kappa}\nabla + \hat{A}_\kappa \right) \Psi_\kappa \right|^2 + \frac{1}{2}|\Psi_\kappa|^4 - |\Psi_\kappa|^2 \right) dx_1 dx_2$$

$$= \int_{D(\delta)} \left\{ \frac{1}{\kappa^2} \left| \frac{\partial \Psi_\kappa}{\partial t} \right|^2 + \frac{1}{g^2} \left| \left(\frac{i}{\kappa}\frac{\partial}{\partial s} + a_2 \right) \Psi_\kappa \right|^2 + \frac{1}{2}|\Psi_\kappa|^4 - |\Psi_\kappa|^2 \right\} g \, ds dt.$$

Applying the transformation (4.5), we obtain

$$\int_0^{\frac{\kappa\delta}{\sqrt{\lambda}}} d\xi_1 \int_{-\frac{\kappa|\partial\Omega|}{2\sqrt{\lambda}}}^{\frac{\kappa}{2\sqrt{\lambda}|\partial\Omega|}} d\xi_2 \frac{\tilde{g}}{\kappa^2}$$

$$\times \left\{ \left| \frac{\partial \tilde{\Psi}_\kappa}{\partial \xi_1} \right|^2 + \frac{1}{\tilde{g}^2} \left| \left(i\frac{\partial}{\partial \xi_2} + \xi_1 - \kappa_r(s)\frac{\sqrt{\lambda}}{\kappa}\xi_1^2 \right) \tilde{\Psi}_\kappa \right|^2 + \frac{1}{2}\lambda|\tilde{\Psi}_\kappa|^4 - \lambda|\tilde{\Psi}_\kappa|^2 \right\},$$

where $\tilde{g}$ is defined by

$$\tilde{g}(\xi_1, \xi_2) = 1 - \frac{\sqrt{\lambda}}{\kappa}\xi_1 \kappa_r(\sqrt{\lambda}\xi_2/\kappa).$$

Since by (4.9), there exist $\tilde{\mu} > 0$ and $\tilde{C}$ such that on $[\, 0, \frac{\kappa\delta}{\sqrt{\lambda}} \,[\times] - \frac{\kappa|\partial\Omega|}{2\sqrt{\lambda}}, +\frac{\kappa|\partial\Omega|}{2\sqrt{\lambda}} \,[$,

$$\left| \left(i\frac{\partial}{\partial \xi_2} + \xi_1 - \kappa_r(\sqrt{\lambda}\xi_2/\kappa)\frac{\sqrt{\lambda}}{\kappa}\xi_1^2 \right) \tilde{\Psi}_\kappa \right|^2 + |\xi_1|^2 |\tilde{\Psi}_\kappa|^2 \leq \tilde{C}\, e^{-\tilde{\mu}\xi_1},$$

there exist $\mu > 0$ and C such that

$$\left| \left(i\frac{\partial}{\partial \xi_2} + \xi_1 \right) \tilde{\Psi}_\kappa \right|^2 \leq C e^{-\mu\xi_1}.$$

We thus obtain

$$\int_0^{\frac{\kappa}{\sqrt{\lambda}}\delta} d\xi_1 \int_{-\frac{\kappa}{2\sqrt{\lambda}}|\partial\Omega|}^{\frac{\kappa}{2\sqrt{\lambda}}|\partial\Omega|} d\xi_2 \frac{1}{\tilde{g}} \left| \left(i\frac{\partial}{\partial \xi_2} + \xi_1 - \kappa_r(\sqrt{\lambda}\xi_2/\kappa)\frac{\sqrt{\lambda}}{\kappa}\xi_1^2 \right) \tilde{\Psi}_\kappa \right|^2$$

$$= \int_0^\infty d\xi_1 \int_{-\frac{\kappa}{2\sqrt{\lambda}}|\partial\Omega|}^{\frac{\kappa}{2\sqrt{\lambda}}|\partial\Omega|} d\xi_2 \left| \left(i\frac{\partial}{\partial \xi_2} + \xi_1 \right) \tilde{\Psi}_\kappa \right|^2 + \mathcal{O}(1).$$

Using the above arguments for the remaining terms yields

$$\mathcal{J}(\Psi_\kappa, \hat{A}_\kappa) = \frac{1}{\kappa^2} \int_0^{\frac{\kappa}{\sqrt{\lambda}}\delta} d\xi_1 \int_{-\frac{\kappa}{2\sqrt{\lambda}}|\partial\Omega|}^{\frac{\kappa}{2\sqrt{\lambda}}|\partial\Omega|} d\xi_2$$

$$\times \left(\left| \frac{\partial \tilde{\Psi}_\kappa}{\partial \xi_1} \right|^2 + \left| \left(i\frac{\partial}{\partial \xi_2} + \xi_1 \right) \tilde{\Psi}_\kappa \right|^2 + \frac{1}{2}\lambda|\Psi_\kappa|^4 - \lambda|\Psi_\kappa|^2 \right) + \mathcal{O}(\kappa^{-2}),$$

so

$$\mathcal{J}(\Psi_\kappa, \hat{A}_\kappa) = \frac{1}{\kappa^2} E_{\omega_\kappa}(\tilde{\Psi}_\kappa, \lambda) + \mathcal{O}(\kappa^{-2}). \tag{4.12}$$

Combining (4.12) with (4.11) yields (4.6). □

We can now prove the main result of this work.

Proof of Theorem 1.1. Let $\hat{\psi}_{\lambda,\kappa} : \Omega_\delta \to \mathbb{C}$ be given by

$$\hat{\psi}_{\lambda,\kappa}(x) = \psi_\lambda\left(\frac{\kappa}{\sqrt{\lambda}} t(x)\right) \exp\left\{-ic_2\left(\frac{\kappa}{\sqrt{\lambda}} s(x)\right)\right\}.$$

Further, let $\chi : \mathbb{R}_+ \to [0,1]$ denote a smooth cutoff function satisfying

$$\chi(t) = \begin{cases} 1 & t \le \frac{1}{2}, \\ 0 & t \ge 1. \end{cases}$$

Then $\chi\big(t(x)/\delta\big)\hat{\psi}_{\lambda,\kappa}(x)$ is in $H^1(\Omega, \mathbb{C})$, and it is not difficult to show that

$$\mathcal{J}(\Psi_\kappa, A_\kappa) \le \mathcal{J}(\hat{\psi}_{\lambda,\kappa}, \hat{A}_\kappa) = -\frac{C_\lambda |\partial\Omega|}{\kappa\sqrt{\lambda}} + \mathcal{O}(\kappa^{-2}), \tag{4.13a}$$

where

$$C_\lambda = -\frac{\omega}{2\pi} E_\omega(\psi_\lambda, \lambda). \tag{4.13b}$$

By Theorem 3.1 there exists $\epsilon > 0$ such that, for $\beta_0 < \lambda < \beta_0 + \epsilon$, we have, for all ω,

$$C_\lambda = -\frac{\omega}{2\pi} \inf_{\psi \in \mathcal{H}_\omega} E_\omega(\psi, \lambda). \tag{4.14}$$

Note that this implies, in particular,

$$C_\lambda = -\lim_{\omega \to 0} \frac{\omega}{2\pi} \inf_{\psi \in \mathcal{H}_\omega} E_\omega(\psi, \lambda). \tag{4.15}$$

Combining (4.6) and (4.13) we obtain

$$\mathcal{J}(\Psi_\kappa, A_\kappa) = -\frac{C_\lambda |\partial\Omega|}{\kappa\sqrt{\lambda}} + \mathcal{O}(\kappa^{-2}), \tag{4.16}$$

In [19, Lemma 7.3], Pan proves (4.16), for any fixed $\beta_0 < \lambda < 1$, by using as a test functions the unknown minimizer of E_{ω_κ} in $\mathcal{H}_{\omega_\kappa}$ instead of ψ_λ, and (4.15) as the definition of C_λ. He also proves (4.16) when $\lambda(\kappa) \to \lambda$ (with $\lambda(\kappa) = \frac{\kappa}{h_{ex}(\kappa)}$) but with an additional $\mathcal{O}([\lambda(\kappa) - \lambda]/\kappa)$ error. Note that when $\lambda = \beta_0$, this result is no more useful since $C_{\beta_0} = 0$, and hence the leading order term of $\mathcal{J}$ is unknown in this case (see [10] for results in this case).

By (2.5) and (3.3), we have

$$C_\lambda = -b(\zeta(\lambda), \lambda),$$

which shows that $C_\lambda > 0$ for every $\beta_0 < \lambda < 1$. Consequently, we have by (4.6) and (4.16),

$$E_{\omega_\kappa}(\tilde{\Psi}_\kappa, \lambda) \le E_{\omega_\kappa}(\psi_\lambda, \lambda) + C. \tag{4.17}$$

Thus, by (4.13b)

$$E_{\omega_\kappa}(\tilde{\Psi}_\kappa,\lambda) \leq -C_\lambda \frac{|\partial\Omega|}{\sqrt{\lambda}}\kappa + C,$$

which proves that indeed

$$\left|E_{\omega_\kappa}(\tilde{\Psi}_\kappa,\lambda)\right| \geq C\kappa, \tag{4.18}$$

and that the correction term on the right-hand side of (4.6) is much smaller than the leading order term.

Let w_κ be defined by

$$\tilde{\Psi}_\kappa(\xi_1,\xi_2) = f_\zeta(\xi_1;\lambda) w_\kappa(\xi_1,\xi_2)\, e^{-ic_2\xi_2}.$$

Clearly, w_κ is periodic in ξ_2. Thus, by (3.8), we get

$$\left|E_{\omega_\kappa}(\tilde{\Psi}_\kappa,\lambda) - E_{\omega_\kappa}(\psi_\lambda,\lambda)\right| \geq \frac{1}{2}\int_{-\pi/\omega_\kappa}^{\pi/\omega_\kappa} d\xi_2 \int_0^\infty d\xi_1\, |f_\zeta|^4 (1-|w_\kappa|^2)^2.$$

Consequently, there exists $C_0 > 0$ such that

$$\int_{-\pi/\omega_\kappa}^{\pi/\omega_\kappa} d\xi_2 \int_0^\infty d\xi_1 |f_\zeta|^4 (1-|w_\kappa|^2)^2 \leq C_0,$$

and hence, for suitable constants C_1 and C_2,

$$\int_{\Omega_\delta} [|\Psi_\kappa|^2 - |\hat{\psi}_{\lambda,\kappa}|^2]^2\, dx \leq \frac{C_1}{\kappa^2}\int_{-\pi/\omega_\kappa}^{\pi/\omega_\kappa} d\xi_2 \int_0^\infty d\xi_1\, |f_\zeta|^4 (1-|w_\kappa|^2)^2 \leq \frac{C_2}{\kappa^2},$$

which proves (1.8a).

To prove (1.8b), we first notice that it is proved in [2] that

$$|\Psi_\kappa| + \frac{1}{\kappa}|\nabla\Psi_\kappa| \leq C,$$

and using the explicit form of $\hat{\psi}_{\lambda,\kappa}$, we obtain

$$\left|\nabla(|\Psi_\kappa|^2 - |\hat{\psi}_{\lambda,\kappa}|^2)\right| \leq C\kappa. \tag{4.19}$$

Evidently, as a consequence of the mean value formula, there exist $C > 0$ and $\delta_0 > 0$, such that, for every $0 < \delta' \leq \delta_0$, there exists $0 \leq \delta'' \leq \delta'$ such that

$$\int_{t=\delta''} \left[\, |\Psi_\kappa|^2 - |\hat{\psi}_{\lambda,\kappa}|^2 \,\right]^2 ds \leq \frac{C}{\delta'}\int_{\Omega_{\delta'}} \left[\, |\Psi_\kappa|^2 - |\hat{\psi}_{\lambda,\kappa}|^2 \,\right]^2 dx_1 dx_2.$$

Furthermore, by (4.19), we have

$$\int_{\partial\Omega} \left[\, |\Psi_\kappa|^2 - |\hat{\psi}_{\lambda,\kappa}|^2 \,\right]^2 ds \leq C\int_{t=\delta''} [|\Psi_\kappa|^2 - |\hat{\psi}_{\lambda,\kappa}|^2]^2\, ds \;+\; C\kappa\delta''.$$

Consequently, there exists $C > 0$ such that

$$\int_{\partial\Omega} \left[\, |\Psi_\kappa|^2 - |\hat{\psi}_{\lambda,\kappa}|^2 \,\right]^2 ds \leq \frac{C}{\delta'}\frac{1}{\kappa^2} + C\kappa\delta'.$$

Choosing $\delta' = \kappa^{-3/2}$ proves (1.8b). ∎

Finally, we compare Theorem 1.1 with the results in [10, Remark 1.5]. As was already stated in the introduction, when $\rho(\kappa) = o(\kappa^{1/2})$ and tends to ∞ as $\kappa \to +\infty$, (1.4) holds. The function u in (1.4) is given by

$$u(\tau) = \beta_0 \frac{|u_0(\tau)|^2}{\|u_0\|_4^4},$$

u_0 denoting the minimizer of (2.1).

We first note that, since as $\lambda \to \beta_0$, we have

$$f_{\zeta_\lambda}^2(\tau) \sim \frac{\lambda - \beta_0}{\beta_0^2} u(\tau),$$

and since

$$\frac{\lambda - \beta_0}{\beta_0^2} \sim \frac{\rho}{\kappa} \quad \text{as } \kappa \to \infty,$$

(1.4) and (1.8a) match. The error in (1.4) is substantially smaller than in (1.8a). The difference is explained by the fact that ψ_κ itself is small on $\partial\Omega$ when $\lambda \to \beta_0$. Thus, if we extrapolate the error term in (1.4) to external fields for which $\rho/\kappa \approx 1$, it becomes $\mathcal{O}(\kappa^{-2})$ exactly as in (1.8a).

Acknowledgments. The authors wish to thank the European program HPRN-CT-2002-00274 (Front Singularities) and their Scientists in Charge in France, Haim Brezis and Danielle Hilhorst, for providing support for the visit of the first author at the University of Paris XI in February 2005.

REFERENCES

[1] Y. Almog, *Non-linear surface superconductivity for type* II *superconductors in the large domain limit*, Arch. Ration. Mech. Anal., 165 (2002), pp. 271–293.

[2] Y. Almog, *Nonlinear surface superconductivity in the large κ limit*, Rev. Math. Phys., 16 (2004), pp. 961–976.

[3] P. Bauman, D. Phillips, and Q. Tang, *Stable nucleation for the Ginzburg-Landau system with an applied magnetic field*, Arch. Ration. Mech. Anal., 142 (1998), pp. 1–43.

[4] A. Bernoff and P. Sternberg, *Onset of superconductivity in decreasing fields for general domains*, J. Math. Phys., 39 (1998), pp. 1272–1284.

[5] C. Bolley and B. Helffer, *Rigorous results on Ginzburg-Landau models in a film submitted to an exterior parallel magnetic field*, I, Nonlinear Stud., 3 (1996), pp. 1–29.

[6] M. Dauge and B. Helffer, *Eigenvalue variations.* I. *Neumann problem for Sturm-Liouville operators*, J. Differential Equations, 104 (1993), pp. 243–262.

[7] M. del Pino, P. L. Felmer, and P. Sternberg, *Boundary concentration for eigenvalue problems related to the onset of superconductivity*, Comm. Math. Phys., 210 (2000), pp. 413–446.

[8] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg-Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[9] C. M. Eliott, H. Matano, and Q. Tang, *Zeros of complex Ginzburg-Landau order parameters with applications to superconductivity*, European J. Appl. Math., 5 (1994), pp. 431–448.

[10] S. Fournais and B. Helffer, *Energy asymptotics for type* II *superconductors*, Calc. Var. Partial Differential Equations, 24 (2005), pp. 341–376.

[11] S. Fournais and B. Helffer, *Accurate spectral estimates for the Neumann magnetic Laplacian*, Ann. Inst. Fourier, 56 (2006), pp. 1–67.

[12] S. Fournais and B. Helffer, *On the third critical field in Ginzburg-Landau theory*, Comm. Math. Phys., 266 (2006), pp. 153–196.

[13] T. Giorgi and D. Phillips, *The breakdown of superconductivity due to strong fields for the Ginzburg-Landau model*, SIAM J. Math. Anal., 30 (1999), pp. 341–359.

[14] B. Helffer, *Erratum: Magnetic Bottles in Connection with Superconductivity*, http://mahery.math.u-psud.fr/∼helffer/erratum164.pdf (18 April 2005).
[15] B. Helffer and A. Morame, *Magnetic bottles in connection with superconductivity*, J. Funct. Anal., 185 (2001), pp. 604–680.
[16] B. Helffer and X. B. Pan, *Upper critical field and location of nucleation of surface superconductivity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 145–181.
[17] K. Lu and X. B. Pan, *Gauge invariant eigenvalue problems in* $\mathbb{R}^2$ *and* $\mathbb{R}^2_+$, Trans. Amer. Math. Soc., 352 (2000), pp. 1247–1276.
[18] W. Meissner and R. Ochsenfeld, Naturwissenschaften, 21 (1933), p. 787.
[19] X. B. Pan, *Surface superconductivity in applied magnetic fields above* H_{c_2}, Comm. Math. Phys., 228 (2002), pp. 327–370.
[20] D. Saint-James and P. G. de Gennes, *Onset of superconductivity in decreasing fields*, Phys. Lett., 7 (1963), pp. 306–308.
[21] S. Serfaty, *Stable configurations in superconductivity: Uniqueness, multiplicity, and vortex-nucleation*, Arch. Ration. Mech. Anal., 149 (1999), pp. 329–365.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1733–1759

QUASI-LINEAR THERMOELASTICITY SYSTEM ARISING IN SHAPE MEMORY MATERIALS*

SHUJI YOSHIKAWA†, IRENA PAWŁOW‡, AND WOJCIECH M. ZAJĄCZKOWSKI§

Abstract. In this paper we establish the global existence and uniqueness of a solution for the three-dimensional and two-dimensional forms of the quasi-linear thermoelasticity system which arises as a mathematical model of shape memory alloys. The system represents a multidimensional version with viscosity and capillarity of the well-known Falk model for one-dimensional martensitic phase transitions. In the setup considered by Pawłow and Zajączkowski [*Math. Methods Appl. Sci.,* 28 (2005), pp. 407–442; 551–592], some conditions have been required for the nonlinear term. In the present paper we improve the result by imposing less restrictive assumptions.

Key words. quasi-linear parabolic equations, maximal regularity, shape memory, thermoelastic

AMS subject classifications. 35Q72, 35K50, 35K60, 74B20

DOI. 10.1137/060653159

1. Introduction.

We consider the following initial-boundary value problem in quasi-linear thermoelasticity:

$$(\mathrm{TE})_d \begin{cases} u_{tt} + \kappa QQu - \nu Qu_t = \nabla \cdot F_{,\varepsilon}(\varepsilon,\theta), \\ [c_v - F_{,\theta\theta}(\varepsilon,\theta)\theta]\theta_t - k\Delta\theta = \theta F_{,\theta\varepsilon}(\varepsilon,\theta) : \varepsilon_t + \nu(A\varepsilon_t) : \varepsilon_t & \text{in } \Omega_T, \\ u = Qu = \nabla\theta \cdot n = 0 & \text{on } S_T, \\ u(0,x) = u_0(x),\ u_t(0,x) = u_1(x),\ \theta(0,x) = \theta_0(x) \geq 0 & \text{in } \Omega, \end{cases}$$

where $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a bounded domain with a smooth boundary $\partial\Omega$, $\Omega_T = (0,T] \times \Omega$, and $S_T = [0,T] \times \partial\Omega$. Let $u = (u_i) \in \mathbb{R}^d$ denote the displacement vector, $\varepsilon = (\varepsilon_{ij})$ with $\varepsilon_{ij}(u) = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})$ the linearized strain tensor, θ the absolute temperature, and $F \in \mathbb{R}$ the elastic energy density. The capillarity term QQu with constant coefficient $\kappa > 0$ corresponds to interaction effects on phase interfaces. The coefficients ν, c_v, and k are positive constants corresponding to the viscosity coefficient, caloric specific heat, and the heat conductivity, respectively.

We use the notation $F_{,\varepsilon} = (\frac{\partial F}{\partial \varepsilon_{ij}})$, $F_{,\theta} = \frac{\partial F}{\partial \theta}$, and $\tilde{\varepsilon} : \varepsilon = \sum_{i,j=1}^{d} \tilde{\varepsilon}_{ij}\varepsilon_{ij}$. We define the linearized elasticity operator Q by the following second order differential operator:

$$Qu = \mu\Delta u + (\lambda + \mu)\nabla(\nabla \cdot u),$$

where λ and μ are the Lamé constants such that

$$\mu > 0 \quad \text{and} \quad d\lambda + 2\mu > 0. \tag{1.1}$$

*Received by the editors February 28, 2006; accepted for publication (in revised form) September 13, 2006; published electronically February 15, 2007.
http://www.siam.org/journals/sima/38-6/65315.html

†Department of Business Administration, Ube National College of Technology, Tokiwadai, 755-8555 Ube, Japan (shoe@ube-k.ac.jp). The work of this author was partly supported by the Research Fellowships of the Japan Society of Promotion of Science (JSPS) for Young Scientists.

‡Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland (pawlow@ibspan.waw.pl) and Institute of Mathematics and Cryptology, Cybernetics Faculty, Military University of Technology, S. Kaliskiego 2, 00-908 Warsaw, Poland.

§Institute of Mathematics, Polish Academy of Sciences, Śniadeckich 8, 00-956 Warsaw, Poland (wz@impan.gov.pl) and Institute of Mathematics and Cryptology, Cybernetics Faculty, Military University of Technology, S. Kaliskiego 2, 00-908 Warsaw, Poland.

The fourth order tensor A represents linear isotropic Hooke's law, defined by

$$A_{ijkl} := \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}).$$

We note that the tensor has the following symmetry properties:

$$A_{ijkl} = A_{klij}, \quad A_{ijkl} = A_{jikl}, \quad A_{ijkl} = A_{ijlk},$$

and the relation $Qu = \nabla\cdot\varepsilon(u)A$ holds. Assumption (1.1) ensures the strong ellipticity of the operator Q and the following inequality:

$$a_*|\varepsilon|^2 \le (A\varepsilon) : \varepsilon \le a^*|\varepsilon|^2,$$

where $a_* = \min\{d\lambda + 2\mu, 2\mu\}$ and $a^* = \max\{d\lambda + 2\mu, 2\mu\}$. In this article, we consider the following structure of the elastic energy density:

(A) $F(\varepsilon,\theta) = G(\theta)H(\varepsilon) + \overline{H}(\varepsilon)$ such that

(i) $G \in \mathrm{C}^3(\mathbb{R},\mathbb{R})$ is as follows:

$$G(\theta) = \begin{cases} C_1\theta & \text{if } \theta \in [0,\theta_1], \\ \varphi(\theta) & \text{if } \theta \in [\theta_1,\theta_2], \\ C_2\theta^r & \text{if } \theta \in [\theta_2,\infty), \end{cases}$$

where $\varphi \in \mathrm{C}^3(\mathbb{R},\mathbb{R})$, $\varphi'' \le 0$, and C_1 and C_2 are positive constants for some fixed θ_1, θ_2 satisfying $0 < \theta_1 < \theta_2 < \infty$. We extend G defined on $\mathbb{R}$ as an odd function.

(ii) $H \in \mathrm{C}^3(\mathbb{S}^2,\mathbb{R})$ satisfies the condition $H(\varepsilon) \ge 0$, where $\mathbb{S}^2$ denotes the set of symmetric second order tensors in $\mathbb{R}^d$.

(iii) $\overline{H} \in \mathrm{C}^3(\mathbb{S}^2,\mathbb{R})$ satisfies $\overline{H}(\varepsilon) \ge -C_3$, where C_3 is some real number.

(iv) $H(\varepsilon)$ and $\overline{H}(\varepsilon)$ satisfy the following growth conditions:

$$\begin{aligned} &|H_{,\varepsilon}(\varepsilon)| \le C|\varepsilon|^{K_1-1}, \quad |H_{,\varepsilon\varepsilon}(\varepsilon)| \le C|\varepsilon|^{K_1-2}, \quad |H_{,\varepsilon\varepsilon\varepsilon}(\varepsilon)| \le C|\varepsilon|^{K_1-3}, \\ &|\overline{H}_{,\varepsilon}(\varepsilon)| \le C|\varepsilon|^{K_2-1}, \quad |\overline{H}_{,\varepsilon\varepsilon}(\varepsilon)| \le C|\varepsilon|^{K_2-2}, \quad |\overline{H}_{,\varepsilon\varepsilon\varepsilon}(\varepsilon)| \le C|\varepsilon|^{K_2-3} \end{aligned}$$

for large $|\varepsilon|$.

Here we note that the regularity assumption for $H(\varepsilon)$ and $\overline{H}(\varepsilon)$ ensures that there exists a positive constant M such that

$$|H_{,\varepsilon}(\varepsilon)| + |H_{,\varepsilon\varepsilon}(\varepsilon)| + |H_{,\varepsilon\varepsilon\varepsilon}(\varepsilon)| + |\overline{H}_{,\varepsilon}(\varepsilon)| + |\overline{H}_{,\varepsilon\varepsilon}(\varepsilon)| + |\overline{H}_{,\varepsilon\varepsilon\varepsilon}(\varepsilon)| \le M$$

for small $|\varepsilon|$. Under the above structure of nonlinearity the system $(\mathrm{TE})_d$ can be rewritten as follows:

$$u_{tt} + \kappa QQu - \nu Qu_t = \nabla\cdot[G(\theta)H_{,\varepsilon}(\varepsilon) + \overline{H}_{,\varepsilon}(\varepsilon)], \tag{1.2}$$

$$c_v\theta_t - k\Delta\theta = \theta G''(\theta)\theta_t H(\varepsilon) + \theta G'(\theta)\partial_t H(\varepsilon) + \nu(A\varepsilon_t) : \varepsilon_t \quad \text{in } \Omega_T, \tag{1.3}$$

$$u = Qu = \nabla\theta\cdot n = 0 \quad \text{on } S_T, \tag{1.4}$$

$$u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x), \quad \theta(0,x) = \theta_0(x) \ge 0 \quad \text{in } \Omega. \tag{1.5}$$

In this paper we show the unique global existence of a solution for (1.2)–(1.5) under the following power of nonlinearity:

$$0 \le r < \frac{5}{6}, \qquad 0 \le K_1, K_2 < 6, \qquad 6r + K_1 < 6 \tag{1.6}$$

in the 3-D case, and

$$0 \le r < 1, \qquad 0 \le K_1, K_2 < \infty \tag{1.7}$$

in the 2-D case.

Before discussing the result of this paper more precisely we shall explain the related results and the physical background of this model. In [10], Falk presents the Landau–Ginzburg type theory using the shear strain $\varepsilon := \partial_x u$ as an order parameter to describe the martensitic-austenitic phase transitions occurring in 1-D shape memory allows (SMAs). There are many papers related to 1-D SMAs (e.g., [2], [3], [6], [12], [16], [17], and [24]). The system $(\mathrm{TE})_d$ is a generalization of the 1-D Falk model with internal viscosity to the 3-D case. The Helmholtz free energy density takes the following form:

$$\widetilde{F}(\varepsilon, \nabla\varepsilon, \theta) = F_0(\theta) + F(\varepsilon, \theta) + |Qu|^2,$$
$$F_0(\theta) = -c_v\theta\log(\theta/\theta_3) + c_v\theta + \tilde{c},$$

and the stress tensor is given by

$$\sigma = \frac{\delta\widetilde{F}}{\delta\varepsilon} + \nu A\varepsilon_t,$$

where $\tilde{c}$ and θ_3 denote the positive physical constants. System $(\mathrm{TE})_d$ can be derived by an argument similar to that in the 1-D case (see [5]). For more details on the derivation of this system, we refer to [19]. In [11], Falk and Konopka give the form of the elastic energy density F as follows:

$$F(\varepsilon, \theta) = \sum_{i=1}^{3} \alpha_i^2(\theta - \theta_c)J_i^2(\varepsilon) + \sum_{i=1}^{5} \alpha_i^4(\theta - \theta_c)J_i^4(\varepsilon) + \sum_{i=1}^{2} \alpha_i^6 J_i^6(\varepsilon), \tag{1.8}$$

where α_i^k, θ_c are constants and J_i^k denote certain kth order monomials with respect to (ε_{ij}). Here we remark that in the 1-D case the elastic energy density takes the following form:

$$F_{1D}(\varepsilon, \theta) = \alpha_1\varepsilon^2(\theta - \theta_c) - \alpha_2\varepsilon^4 + \alpha_3\varepsilon^6, \tag{1.9}$$

where $\varepsilon := \partial_x u$ and α_i, θ_c are positive constants. Comparing (1.8) with the 1-D form (1.9), we see that in the 3-D case, $H(\varepsilon)$ must be fourth order with respect to ε. This causes some difficulties in the mathematical treatment of system (1.2)–(1.5). Moreover, the difficulties arise also from the fact that the useful embedding $H^1 \hookrightarrow L^\infty$ does not hold in the multidimensional case. There had been no papers on the solvability of this system with the Falk–Konopka elastic energy density (1.8), $r = 1$, $K_1 = 4$, and $K_2 = 6$. Then Pawłow and Żochowski [20] studied the energy density F under several stronger assumptions than (1.8), namely, lower order powers of nonlinearity. Moreover, for the simplification of treatments, they first considered the semilinearized equations of the quasi-linear system $(\mathrm{TE})_d$:

$$(\mathrm{SLTE})_d \begin{cases} u_{tt} + \kappa QQu - \nu Qu_t = \nabla \cdot F_{,\varepsilon}(\varepsilon, \theta), & \\ c_v\theta_t - k\Delta\theta = \theta F_{,\theta\varepsilon}(\varepsilon, \theta) : \varepsilon_t + \nu(A\varepsilon_t) : \varepsilon_t & \text{in } \Omega_T, \\ u = Qu = \nabla\theta \cdot n = 0 & \text{on } S_T, \\ u(0, x) = u_0(x),\ u_t(0, x) = u_1(x),\ \theta(0, x) = \theta_0(x) \ge 0 & \text{in } \Omega, \end{cases}$$

which is the model $(\mathrm{TE})_d$ with removed quasi-linear term $\theta G''(\theta)\theta_t H(\varepsilon)$. They showed the unique global existence of a sufficiently smooth solution for $(\mathrm{SLTE})_d$ under the following assumptions on the nonlinearity:

$$0 \le r < \frac{1}{2}, \qquad 0 \le K_1 \le \left(\frac{1}{2} - r\right) K_2 + 1, \qquad 0 \le K_2 \le \frac{7}{2} \tag{1.10}$$

in the 3-D case, and

$$0 \le r < \frac{1}{2}, \qquad 0 \le K_1 \le \left(\frac{1}{2} - r\right) K_2 + 1, \qquad 0 \le K_2 < \infty \tag{1.11}$$

in the 2-D case. In addition, due to the applied parabolic decomposition of the elasticity system, they assumed the condition $0 < 2\sqrt{\kappa} \le \nu$ between viscosity and capillarity. Such an assumption, however, does not seem realistic for SMA viscosity effects which are negligibly small. In [25] the unique global existence of the solution to $(\mathrm{SLTE})_3$ in a larger class is proved by using the contraction mapping principle. The result is proved without conditions between κ and ν, and the class of nonlinearities is generalized to $K_2 < 6$. The first two assumptions in (1.10) are present due to the semilinearization which causes the absence of energy conservation law (Lemma 4.1 below). Recently, Pawłow and Zajączkowski [21] have proved the unique global existence for the quasi-linear system (1.2)–(1.5) under the assumptions

$$\begin{gathered} 0 < r < \frac{2}{3}, \quad 0 < K_1 < \frac{15}{4} \ \text{ and } \ 15r + 4K_1 = 15 \ \text{ if } \ K_1 > 1, \\ 0 < K_2 \le \frac{9}{2}, \quad 0 < 2\sqrt{\kappa} \le \nu. \end{gathered} \tag{1.12}$$

The latter, restrictive condition between viscosity and capillarity has been removed by the above mentioned authors in [23]. The aim of the present paper is to prove the unique global existence of a solution to system (1.2)–(1.5) under weaker assumptions than (1.12). More precisely, we admit the nonlinearity specified in (1.6), (1.7) and arbitrary positive coefficients of capillarity $\kappa > 0$ and viscosity $\nu > 0$. Unfortunately, our result still does not cover the physically realistic case (1.8).

Here we add some remarks on the 2-D case. The results of [20] include the 2-D case of the semilinearized problem $(\mathrm{SLTE})_2$. The unique global existence for the 2-D quasilinear system $(\mathrm{TE})_2$ is established in [22] under the assumption

$$0 \le r < \frac{7}{8}, \qquad 0 \le K_1 < \infty, \qquad 0 \le K_2 < \infty. \tag{1.13}$$

In [26] the unique global existence for $r = 1$ is proved under other strong assumptions. Roughly speaking, the restrictions in [26] are such that $K_1 = 0$ and that the energy of initial data $\|u_0\|_{H^2} + \|u_1\|_{L^2} + \|\theta_0\|_{L^1}$ is sufficiently small. We note that if we take $r = 1$, then the quasi-linear term $\theta G''(\theta)H(\varepsilon)\theta_t$ of (1.3) does not appear. We also describe the result for the 2-D case in section 5 of this paper. We show that the system $(\mathrm{TE})_2$ has a unique global solution under assumptions (1.7). Comparing these assumptions with (1.13), we see that the restriction for r is weaker; nevertheless we cannot admit $r = 1$.

We now introduce some notation and function spaces. Throughout this paper C and Λ are positive constants independent of time T and depending on time T, respectively. In particular, we may use Λ instead of $\Lambda(\|(u_0, u_1, \theta_0)\|_X)$ for some X if there is no danger of confusion.

- $L^p(\Omega_T) = L^p_T L^p = L^p(0,T;L^p(\Omega))$ is the standard Lebesgue space. We often use the notation $L^p(\Omega_I) = L^p_I L^p$ for some interval I.
- $W^{2l,l}_p(\Omega_T)$ is the Sobolev space equipped with the norm

$$\|u\|_{W^{2l,l}_p(\Omega_T)} := \sum_{j=0}^{2l} \sum_{2r+|\alpha|=j} \|D_t^r D_x^\alpha u\|_{L^p(\Omega_T)},$$

  where $D_t := i\frac{\partial}{\partial t}$, $D_x^\alpha = \prod_{\alpha=\alpha_1+\alpha_2+\alpha_3} D_k^{\alpha_k}$, and $D_k := i\frac{\partial}{\partial x_k}$ for multi-index $\alpha = (\alpha_i)_{i=1}^n$.

- $H^j(\Omega) := W^j_2(\Omega)$, where W^j_p is the Sobolev space equipped with the norm $\|u\|_{W^j_p(\Omega)} := \sum_{|\alpha|\le j} \|D_x^\alpha u\|_{L^p(\Omega)}$.
- $B^s_{p,q} = B^s_{p,q}(\Omega)$ is the Besov space. Namely, $B^s_{p,q} := [L^p(\Omega), W^j_p(\Omega)]_{s/j,q}$, where $[X,Y]_{s/j,q}$ is the real interpolation space. For more details we refer to [1] by Adams and Fournier.
- $\mathrm{C}^{\alpha,\alpha/2}(\Omega_T)$ is the Hölder space: the set of all continuous functions in Ω_T satisfying the Hölder condition in x with exponent α, and in t with exponent $\alpha/2$.

We now state the main result of this paper.

THEOREM 1.1. *Let the positive physical constants κ, ν, c_v, and k be fixed arbitrarily. Assume that* $\min_\Omega \theta_0 \ge 0$ *and* (1.6) *holds. Then, given* $5 < p \le q < \infty$, *for any* $T > 0$ *and* $(u_0, u_1, \theta_0) \in B^{4-2/p}_{p,p} \times B^{2-2/p}_{p,p} \times B^{2-2/q}_{q,q} =: U(p,q)$, *there exists at least one solution* (u,θ) *to* (1.2)–(1.5) *satisfying*

$$(u,\theta) \in W^{4,2}_p(\Omega_T) \times W^{2,1}_q(\Omega_T) =: V_T(p,q).$$

Moreover, if we assume $\min_\Omega \theta_0 = \theta_* > 0$, *then there exists a positive constant* ω *such that*

$$\theta \ge \theta_* \exp(-\omega t) \qquad \text{in } \Omega_T.$$

For completeness we recall also the uniqueness result which follows by repeating the arguments of the corresponding result in [22, section 6].

THEOREM 1.2. *In addition to the assumptions of Theorem* 1.1, *suppose that* $F(\varepsilon,\theta) \in \mathrm{C}^4(\mathbb{S}^2 \times \mathbb{R}^+, \mathbb{R})$. *Then the solution* $(u,\theta) \in V_T(p,q)$ *to* (1.2)–(1.5) *constructed above is unique.*

We prove Theorem 1.1 by using the Leray–Schauder fixed point principle. The key estimates are the maximal regularity estimate for (1.2), and the classical energy estimate and the parabolic De Giorgi method for (1.3). In general, the derivative of a solution is less regular than the right-hand side of the corresponding equation. However, for parabolic equations such a loss of regularity does not occur, as in the case of elliptic equations. The estimate ensuring this regularity is called the maximal regularity. For more precise information on the maximal regularity we refer to [4], and for more recent topics of the maximal L^p-regularity we refer to [9]. Since the maximal regularity theory is limited to linear parabolic equations, we cannot use it directly for the quasi-linear equation (1.3). To obtain the higher order a priori estimates we apply the classical energy methods and the parabolic De Giorgi method (see [14], [15]). Using these methods we can show the Hölder continuity of θ. By virtue of such regularity, we arrive at the estimate in a higher Sobolev norm.

In section 2 we list several preliminary results which are used in the paper. In section 3 we prove the unique global existence of the solution for certain truncated

version of problem (1.2)–(1.5). To this purpose we use the Leray–Schauder fixed point principle. In section 4 we show that the solution of $(\mathrm{TE})_3$ coincides with the solution of the truncated problem constructed in section 3 for a sufficiently large truncation level L. In section 5 we consider the 2-D system $(\mathrm{TE})_2$.

2. Preliminaries. In this section, we present some auxiliary results which will be used in the subsequent sections.

LEMMA 2.1 (maximal regularity).

(i) *Let $p \in (1,\infty)$. Denote by u the solution of the linear problem*

$$\begin{cases} u_{tt} + \kappa QQu - \nu Qu_t = \nabla \cdot f & \text{in } \Omega_T, \\ u = Qu = 0 & \text{on } S_T, \\ u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x) & \text{in } \Omega. \end{cases}$$

Then the following estimates hold:

$$\|u\|_{W_p^{4,2}(\Omega_T)} \le C \left(\|u_0\|_{B_{p,p}^{4-\frac{2}{p}}} + \|u_1\|_{B_{p,p}^{2-\frac{2}{p}}} + \|\nabla \cdot f\|_{L^p(\Omega_T)} \right) \tag{2.1}$$

for any $(u_0,u_1) \in B_{p,p}^{4-2/p} \times B_{p,p}^{2-2/p}$ and $\nabla \cdot f \in L^p(\Omega_T)$, and

$$\|\nabla u\|_{W_p^{2,1}(\Omega_T)} \le C \left(\|u_0\|_{B_{p,p}^{3-\frac{2}{p}}} + \|u_1\|_{B_{p,p}^{1-\frac{2}{p}}} + \|f\|_{L^p(\Omega_T)} \right) \tag{2.2}$$

for any $(u_0,u_1) \in B_{p,p}^{3-2/p} \times B_{p,p}^{1-2/p}$ and $f \in L^p(\Omega_T)$.

(ii) *Let $q \in (1,\infty)$. Assume that $\rho(x)$ is Hölder continuous in Ω such that $\inf_\Omega \rho > 0$. Denote by θ the solution of the linear problem*

$$\begin{cases} \theta_t - \rho \Delta \theta = g & \text{in } \Omega_T, \\ n \cdot \nabla \theta = 0 & \text{on } S_T, \\ \theta(0,x) = \theta_0(x) & \text{in } \Omega. \end{cases}$$

Then the following estimate holds:

$$\|\theta\|_{W_q^{2,1}(\Omega_T)} \le C \left(\|\theta_0\|_{B_{q,q}^{2-\frac{2}{q}}} + \|g\|_{L^q(\Omega)} \right) \tag{2.3}$$

for any $\theta_0 \in B_{q,q}^{2-2/q}$, where C depends on $\inf_\Omega \rho$.

For the proof of (i) we refer to [25, Lemma 2.1, Proposition 2.4], and (ii) is the particular case of [13, Examples 3.2, A), 2)]. Next, we recall the useful space-time embedding lemma.

LEMMA 2.2 (embedding [14, Lemma II.3.3]). *Let $f \in W_p^{2l,l}(\Omega_T)$. Then, for $l \in \mathbb{Z}^+$ and multi-index α, it follows that*

$$\|D_t^r D_x^\alpha f\|_{L^q(\Omega_T)} \le C\delta^{l-\psi} \|f\|_{W_p^{2l,l}(\Omega_T)} + C\delta^{-\psi} \|f\|_{L^p(\Omega_T)}, \tag{2.4}$$

provided $q \ge p$ and $\psi := r + \frac{|\alpha|}{2} + \frac{d+2}{2}(\frac{1}{p} - \frac{1}{q}) \le l$. If $\varphi := r + \frac{|\alpha|}{2} + \frac{d+2}{2p} < l$, then

$$\|D_t^r D_x^\alpha f\|_{L^\infty(\Omega_T)} \le C\delta^{l-\varphi} \|f\|_{W_p^{2l,l}(\Omega_T)} + C\delta^{-\varphi} \|f\|_{L^p(\Omega_T)}; \tag{2.5}$$

moreover, $D_t^r D_x^\alpha f$ is Hölder continuous. Here, $\delta \in (0, \min(T, \zeta^2)]$, ζ is the altitude of the cone in the statement of the cone condition satisfied by Ω.

LEMMA 2.3. *Let φ be given as in* (A)(i). *Then the function $\varphi(s)$ satisfies*

$$\varphi(s) - s\varphi'(s) \geq 0 \tag{2.6}$$

for any $s \in [\theta_1, \theta_2]$.

Proof. Putting $f(s) = \varphi(s) - s\varphi'(s)$, we have $f'(s) = -s\varphi''(s) \geq 0$ and $f(\theta_1) = 0$. Then $f(s) = \varphi(s) - s\varphi'(s) \geq 0$ in $[\theta_1, \theta_2]$. □

To show Theorem 1.1 we apply the Leray–Schauder fixed point principle. We recall it here in one of its equivalent formulations for the reader's convenience .

THEOREM 2.4 (Leray–Schauder fixed point principle [8]). *Let X be a Banach space. Assume that $\Phi : [0,1] \times X \to X$ is a map with the following properties:*

(L1) *For any fixed $\tau \in [0,1]$ the map $\Phi(\tau, \cdot) : X \to X$ is compact.*

(L2) *For every bounded subset $\mathcal{B}$ of X, the family of maps $\Phi(\cdot, \xi) : [0,1] \to X$, $\xi \in \mathcal{B}$, is uniformly equicontinuous.*

(L3) *$\Phi(0, \cdot)$ has precisely one fixed point in X.*

(L4) *There is a bounded subset $\mathcal{B}$ of X such that any fixed point in X of $\Phi(\tau, \cdot)$ is contained in $\mathcal{B}$ for every $0 \leq \tau \leq 1$.*

Then $\Phi(1, \cdot)$ has at least one fixed point in X.

3. Truncated problem. To prove the existence theorem we first consider the following truncated problem $(\mathrm{TE})_3^L$:

$$u_{tt} + \kappa QQu - \nu Qu_t = \Gamma_L\Big(\nabla \cdot [G(\theta)H_{,\varepsilon}(\varepsilon) + \overline{H}_{,\varepsilon}(\varepsilon)]\Big), \tag{3.1}$$

$$\begin{aligned} &c_v\theta_t - k\Delta\theta = \theta G''(\theta)\theta_t H(\varepsilon) + \theta G'(\theta)\partial_t H(\varepsilon) + \nu(A\varepsilon_t) : \varepsilon_t && \text{in } \Omega_T, \\ &u = Qu = \nabla\theta \cdot n = 0 && \text{on } S_T, \\ &u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x), \quad \theta(0,x) = \theta_0(x) \geq 0 && \text{in } \Omega, \end{aligned} \tag{3.2}$$

where

$$\Gamma_L(x) = \begin{cases} x & \text{if } |x| \leq L, \\ L\frac{x}{|x|} & \text{if } |x| \geq L. \end{cases}$$

THEOREM 3.1. *Fix L and $5 < p \leq q < \infty$. Assume that $\theta_0 \geq 0$,* (1.6) *holds, and $F(\varepsilon, \theta) \in \mathrm{C}^4(\mathbb{S}^2 \times \mathbb{R}^+, \mathbb{R})$. Then for any $T > 0$ and $(u_0, u_1, \theta_0) \in U(p,q)$, there exists a unique solution (u_L, θ_L) to $(\mathrm{TE})_3^L$ satisfying $(u_L, \theta_L) \in V_T(p,q)$.*

Proof of Theorem 3.1. We apply Theorem 2.4 to the map Φ_τ^L from $V_T(p,q)$ into $V_T(p,q)$,

$$\Phi_\tau^L : (\bar{u}, \bar{\theta}) \mapsto (u, \theta), \qquad \tau \in [0,1],$$

defined by means of the following initial-boundary value problems:

$$\begin{aligned} &u_{tt} + \kappa QQu - \nu Qu_t = \tau\Gamma_L\Big(\nabla \cdot [G(\bar{\theta})H_{,\varepsilon}(\bar{\varepsilon}) + \overline{H}_{,\varepsilon}(\bar{\varepsilon})]\Big), \\ &c_v\theta_t - k\Delta\theta = \tau\left\{\bar{\theta}G''(\bar{\theta})\theta_t H(\varepsilon) + \bar{\theta}G'(\bar{\theta})\partial_t H(\varepsilon) + \nu(A\varepsilon_t) : \varepsilon_t\right\} && \text{in } \Omega_T, \\ &u = Qu = \nabla\theta \cdot n = 0 && \text{on } S_T, \\ &u(0,x) = \tau u_0(x), \quad u_t(0,x) = \tau u_1(x), \quad \theta(0,x) = \tau\theta_0(x) && \text{in } \Omega, \end{aligned}$$

where $\bar{\varepsilon} = \varepsilon(\bar{u})$. A fixed point of $\Phi_\tau^L(1, \cdot)$ in $V_T(p,q)$ is the desired solution of the system $(\mathrm{TE})_3^L$. Therefore to prove the existence statement it is sufficient to check

that the map Φ_τ^L satisfies assumptions (L1)–(L4) of Theorem 2.4. Noting that Γ_L is Lipschitz continuous, we can check assumptions (L1), (L2), and (L3) in the same way as that in [21, section 3]. Then it is sufficient to check assumption (L4), namely, to derive a priori bounds for a fixed point of the solution map Φ_τ^L. Without loss of generality we may set $\tau = 1$. Hence, from now on our purpose is to obtain a priori bounds for $(\mathrm{TE})_3^L$. To this end we prepare several lemmas. If there is no danger of confusion, we write for simplicity (u, θ) instead of (u_L, θ_L).

LEMMA 3.2 (maximum principle). *Let* $(u_0, u_1, \theta_0) \in B^{4-2/p}_{p,p} \times B^{2-2/p}_{p,p} \times L^2$ *for* $p > 5$. *Assume that* $\min_\Omega \theta_0 \geq 0$. *Then the solution* θ *to* $(\mathrm{TE})_3^L$ *is nonnegative almost everywhere in* Ω_T.

Proof. It follows from the maximal regularity (2.1) for (3.1) that

$$
\begin{aligned}
&\|u\|_{W_p^{4,2}(\Omega_T)} \\
&\quad \leq C\left(\|u_0\|_{B^{4-2/p}_{p,p}} + \|u_1\|_{B^{2-2/p}_{p,p}} + \|\Gamma_L\left\{\nabla\cdot[G(\bar\theta)H_{,\varepsilon}(\bar\varepsilon) + \overline{H}_{,\varepsilon}(\bar\varepsilon)]\right\}\|_{L^p(\Omega_T)}\right) \\
&\quad \leq C(\|u_0\|_{B^{4-2/p}_{p,p}} + \|u_1\|_{B^{2-2/p}_{p,p}} + L|\Omega_T|^{\frac{1}{p}}) \\
&\quad \leq \Lambda(L).
\end{aligned} \tag{3.3}
$$

Then taking $p > 5$, by Lemma 2.2 we have

$$
\|\varepsilon\|_{L^\infty(\Omega_T)} + \|\varepsilon_t\|_{L^\infty(\Omega_T)} \leq \Lambda(L) < \infty. \tag{3.4}
$$

Therefore it holds that

$$
\|\partial_t H(\varepsilon)\|_{L^\infty(\Omega_T)} \leq \|\varepsilon_t\|_{L^\infty(\Omega_T)}\|\varepsilon\|^{K_1-1}_{L^\infty(\Omega_T)} \leq \Lambda(L)
$$

for $K_1 > 1$. Since $\sup_{\varepsilon\in\mathbb{S}} |H_{,\varepsilon}(\varepsilon)| \leq M$ for $K_1 \leq 1$, we conclude that

$$
\|\partial_t H(\varepsilon)\|_{L^\infty(\Omega_T)} \leq \Lambda(L) \tag{3.5}
$$

for every $K_1 \geq 0$. From now on, throughout this section we shall write $\Lambda = \Lambda(L)$.

Multiplying (3.2) by $\theta_- := \min\{\theta, 0\}$ and integrating over Ω, we have

$$
\begin{aligned}
&\frac{c_v}{2}\frac{d}{dt}\int_\Omega \theta_-^2 dx + k\int_\Omega |\nabla\theta_-|^2 dx \\
&\qquad = \int_\Omega [\theta_-\theta G''(\theta)\theta_t H(\varepsilon) + \theta_-\theta G'(\theta)\partial_t H(\varepsilon) + \nu\theta_- A\varepsilon_t : \varepsilon_t]\, dx \\
&\qquad = \frac{d}{dt}\int_\Omega H(\varepsilon)G_2(\theta_-)dx + \int_\Omega \overline{G}_2(\theta_-)\partial_t H(\varepsilon)dx + \int_\Omega \nu\theta_- A\varepsilon_t : \varepsilon_t dx,
\end{aligned}
$$

where $G_2(\theta) = \theta^2 G'(\theta) - \overline{G}_2(\theta)$ and $\overline{G}_2(\theta) = 2\int_0^\theta sG'(s)ds$. We have $G_2(0) = 0$ and $G_2'(y) = y^2 G''(y) \geq 0$ for $y \leq 0$, because G'' is the odd function such that $G''(y) \leq 0$ for $y \geq 0$. Then $G_2(y) \leq 0$ for $y \geq 0$. Hence, we have

$$
-\int_\Omega H(\varepsilon)G_2(\theta_-)dx \geq 0.
$$

It follows from (1.1) that

$$
\int_\Omega \nu\theta_- A\varepsilon_t : \varepsilon_t dx \leq \nu a_* \int_\Omega \theta_- |\varepsilon_t|^2 dx \leq 0.
$$

Noting that $\overline{G}_2(\theta) = \frac{1}{2}C_1\theta^2$ for $\theta \in [-\theta_1, \theta_1]$, we have $\sup_{s\in\mathbb{R}} \frac{|\overline{G}_2(s)|}{s^2} \le C$. Therefore we conclude that

$$\begin{aligned}\int_\Omega \overline{G}_2(\theta_-)\partial_t H(\varepsilon)dx &\le \int_\Omega |\theta_-|^2 \frac{|\overline{G}_2(\theta_-)|}{|\theta_-|^2}|\partial_t H(\varepsilon)|dx \\ &\le \Lambda\|\theta_-\|_{L^2}^2.\end{aligned}$$

Consequently, we have

$$\frac{d}{dt}\left(c_v\|\theta_-(t)\|_{L^2}^2 - \int_\Omega H(\varepsilon)G_2(\theta_-)dx\right) \le \Lambda\left(c_v\|\theta_-(t)\|_{L^2}^2 - \int_\Omega H(\varepsilon)G_2(\theta_-)dx\right).$$

Using the Gronwall inequality we obtain

$$\begin{aligned}\|\theta_-(t)\|_{L^2}^2 &\le \|\theta_-(t)\|_{L^2}^2 - \int_\Omega H(\varepsilon)G_2(\theta_-)dx \\ &\le \Lambda e^{\Lambda t}\left(\|\theta_-(0)\|_{L^2}^2 - \int_\Omega H(\varepsilon(0))G_2(\theta_-(0))dx\right) \\ &= 0,\end{aligned}$$

which completes the proof. □

LEMMA 3.3. *Let $l > 2$ be an arbitrary integer. Assume that $r \le 1$. Then for any $(u_0, u_1, \theta_0) \in B^{4-2/p}_{p,p} \times B^{2-2/p}_{p,p} \times L^l =: U_1(l)$, the solution (u, θ) to $(\mathrm{TE})^L_3$ satisfies*

$$\|\theta\|_{L^\infty_T L^l} \le \Lambda,$$

where $\Lambda = \Lambda(T, \|(u_1, u_2, \theta_0)\|_{U_1(l)})$. Moreover, if $(u_0, u_1, \theta_0) \in U_1(\infty)$, then we have

$$\|\theta\|_{L^\infty(\Omega_T)} \le \Lambda,$$

where $\Lambda = \Lambda(T, \|(u_1, u_2, \theta_0)\|_{U_1(\infty)})$.

Proof. Multiplying (3.2) by θ^{l-1} and integrating over Ω, we have

$$\begin{aligned}&\frac{c_v}{l}\frac{d}{dt}\|\theta\|_{L^l}^l + k(l-1)\int_\Omega \theta^{l-2}|\nabla\theta|^2 dx \\ &\quad= \int_\Omega \left(\theta^l G''(\theta)\theta_t H(\varepsilon) + \theta^l G'(\theta)\partial_t H(\varepsilon) + \nu\theta^{l-1}A\varepsilon_t : \varepsilon_t\right)dx \\ &\quad= \frac{d}{dt}\int_\Omega G_l(\theta)H(\varepsilon)dx + \int_\Omega \overline{G}_l(\theta)\partial_t H(\varepsilon)dx + \nu\int_\Omega \theta^{l-1}A\varepsilon_t : \varepsilon_t dx,\end{aligned} \tag{3.6}$$

where $G_l(\theta) = \theta^l G'(\theta) - \overline{G}_l(\theta)$ and $\overline{G}_l(\theta) = l\int_0^\theta s^{l-1}G'(s)ds$. Since

$$\theta^l G''(\theta) = \begin{cases} C_2 r(r-1)\theta^{l+r-2} \le 0 & \text{for } \theta \ge \theta_2, \\ \theta^l\varphi''(\theta) \le 0 & \text{for } \theta_1 \le \theta \le \theta_2, \\ 0 & \text{for } \theta \le \theta_1, \end{cases} \tag{3.7}$$

we have $G_l'(\theta) = \theta^l G''(\theta) \le 0$ for $\theta \ge 0$ and $G_l'(0) = 0$. Thereby, we obtain

$$G_l(\theta) \le 0 \quad \text{for} \quad \theta \ge 0. \tag{3.8}$$

We put

$$\hat{\theta} = \theta \left(1 - \frac{lG_l(\theta)H(\varepsilon)}{c_v \theta^l}\right)^{1/l}.$$

We note that $\hat{\theta} \geq \theta$ due to (3.8). Since $\sup_{s\in[0,\infty)} |G'(s)| =: M < \infty$, we have

$$|\overline{G}_l(\theta)| = \left| l \int_0^\theta s^{l-1} G'(s) ds \right| \leq C\theta^l$$

and

$$|G_l(\theta)| \leq M\theta^l + |\overline{G}_l(\theta)| \leq C\theta^l.$$

In view of (3.4) and (3.5) we obtain

$$\left| \int_\Omega \overline{G}_l(\theta) \partial_t H(\varepsilon) dx \right| \leq C \|\theta^l\|_{L^1(\Omega)} \|\partial_t H(\varepsilon)\|_{L^\infty(\Omega)} \leq \Lambda \|\theta\|^l_{L^l(\Omega)}$$

and

$$\int_\Omega \theta^{l-1} A\varepsilon_t : \varepsilon_t \leq C \|\varepsilon_t\|^2_{L^\infty(\Omega)} \|\theta\|^{l-1}_{L^{l-1}(\Omega)} \leq \Lambda \|\theta\|^{l-1}_{L^l(\Omega)}.$$

Since $\frac{1}{l}\partial_t \|\hat{\theta}\|^l_{L^l} = \|\hat{\theta}\|^{l-1}_{L^l} \partial_t \|\hat{\theta}\|_{L^l}$, it follows from (3.6) that

$$\begin{aligned} \frac{d}{dt}\|\hat{\theta}\|_{L^l(\Omega)} &\leq \Lambda \|\theta\|_{L^l(\Omega)} + \Lambda \\ &\leq \Lambda \|\hat{\theta}\|_{L^l(\Omega)} + \Lambda. \end{aligned}$$

Thus by the Gronwall inequality we have

$$\|\hat{\theta}\|_{L^\infty_T L^l} \leq \Lambda \|\hat{\theta}_0\|_{L^l} + \Lambda. \tag{3.9}$$

Since

$$\begin{aligned} \hat{\theta}_0 &= \theta_0 \left(1 - \frac{lG_l(\theta_0)H(\varepsilon_0)}{c_v \theta_0^l}\right)^{1/l} \\ &\leq \theta_0 \left(1 + \frac{lM\Lambda}{c_v}\right)^{1/l}, \end{aligned}$$

we can obtain the first assertion. Here we note that the constant Λ in (3.9) is independent of l. Therefore taking a limit as $l \to \infty$ we can obtain the second assertion. This completes the proof. □

LEMMA 3.4. *Let T be arbitrarily fixed. Assume that $r \leq 1$. Then for any $(u_0, u_1, \theta_0) \in B^{4-2/p}_{p,p} \times B^{2-2/p}_{p,p} \times H^1 =: U_2$, the solution (u, θ) to $(\mathrm{TE})^L_3$ satisfies*

$$\|\theta\|_{W^{2,1}_2(\Omega_T)} \leq \Lambda,$$

where Λ depends on T and $\|(u_0, u_1, \theta_0)\|_{U_2}$.

Proof. By using Lemma 3.3, thanks to $\theta_0 \in H^1 \hookrightarrow L^2$, we have

$$\|\theta\|_{L^\infty_T L^2} \leq \Lambda. \tag{3.10}$$

Since $\theta G''(\theta) \le 0$ from (3.7) for $l = 1$, the following estimate holds true:

$$\iint_{\Omega_T} \theta_t^2 \theta G''(\theta) H(\varepsilon) dxdt \le 0. \tag{3.11}$$

Multiplying (3.2) by θ_t and integrating over Ω_T, we have

$$\begin{aligned}
&c_v \|\theta_t\|^2_{L^2(\Omega_T)} + \frac{k}{2}\|\nabla\theta\|^2_{L^\infty_T L^2} \\
&\quad \le \frac{k}{2}\|\theta_0\|^2_{H^1} + \iint_{\Omega_T} \theta_t^2 \theta G''(\theta) H(\varepsilon) dxdt \\
&\qquad + \iint_{\Omega_T} \theta_t \theta G'(\theta) \partial_t H(\varepsilon) dxdt + \iint_{\Omega_T} \nu \theta_t A\varepsilon_t : \varepsilon_t dxdt \\
&\quad \le \frac{k}{2}\|\theta_0\|^2_{H^1} + \Lambda \|\theta_t\|_{L^2(\Omega_T)} \|\theta\|^r_{L^\infty_T L^2} \|\partial_t H(\varepsilon)\|_{L^\infty(\Omega_T)} \\
&\qquad + \Lambda \|\theta_t\|_{L^\infty_T L^2} \|\varepsilon_t\|^2_{L^\infty(\Omega_T)} \\
&\quad \le \frac{k}{2}\|\theta_0\|^2_{H^1} + \frac{c_v}{2}\|\theta_t\|^2_{L^2(\Omega_T)} + \Lambda,
\end{aligned}$$

thanks to (3.4), (3.5), (3.10), and (3.11). Therefore we arrive at

$$\|\theta_t\|_{L^2(\Omega_T)} + \|\nabla\theta\|_{L^\infty_T L^2} \le \Lambda(\|(u_0, u_1, \theta_0)\|_{U_2}).$$

Next, multiplying (3.2) by $\frac{-\Delta\theta}{c_v - \theta G''(\theta) H(\varepsilon)}$ and integrating over Ω, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nabla\theta(t)\|^2_{L^2} + \int_\Omega \frac{k(\Delta\theta)^2}{c_v - \theta G''(\theta) H(\varepsilon)} dx \\
&\quad = -\int_\Omega \frac{\Delta\theta}{c_v - \theta G''(\theta) H(\varepsilon)} \left(\theta G'(\theta) \partial_t H(\varepsilon) + \nu A\varepsilon_t : \varepsilon_t\right) dx.
\end{aligned}$$

Here we remark that

$$c_v \le c_v - \theta G''(\theta) H(\varepsilon) \le c_v + M\Lambda,$$

where $0 \le \sup_{\theta \ge 0}(-\theta G''(\theta)) =: M < \infty$. Then integrating over $[0, t]$ for $t \le T$, we conclude with the estimate

$$\begin{aligned}
&\|\nabla\theta(t)\|^2_{L^2} + \frac{2k}{c_v + \Lambda M}\|\Delta\theta\|^2_{L^2(\Omega_T)} \\
&\quad \le \|\nabla\theta_0\|^2_{L^2} + 2\|\Delta\theta\|_{L^2(\Omega_T)} \left\|\theta G'(\theta)\partial_t H(\varepsilon) + \nu A\varepsilon_t : \varepsilon_t\right\|_{L^2(\Omega_T)} \\
&\quad \le \|\nabla\theta_0\|^2_{L^2} + \frac{k}{(c_v + \Lambda M)}\|\Delta\theta\|^2_{L^2(\Omega_T)} \\
&\qquad + \frac{c_v + \Lambda M}{k}\Big(\Lambda\|\theta\|_{L^\infty_T L^2}\|\partial_t H(\varepsilon)\|_{L^\infty(\Omega_T)} + \Lambda\|\varepsilon_t\|_{L^\infty(\Omega_T)}\Big)^2 \\
&\quad \le \frac{k}{(c_v + \Lambda M)}\|\Delta\theta\|^2_{L^2(\Omega_T)} + \Lambda.
\end{aligned}$$

Consequently, we arrive at the desired result. □

The same procedure as in [21, Lemma 6.1] allows us to conclude that $\theta \in \mathrm{C}^{\alpha,\alpha/2}(\Omega_T)$ for some Hölder exponent $0 < \alpha < 1$ depending on T, $\sup_\Omega \theta_0$, and $\|\theta\|_{L^\infty(\Omega_T)}$. The proof relies on the classical parabolic De Giorgi method. For more precise information on this method we refer to [14, Chapter II, section 7] and [15, Chapter VI, section 12]. Here we note that ε is Hölder continuous due to Lemma 2.2.

LEMMA 3.5 (see [21, Lemma 6.1]). *Assume that $k = \sup_\Omega \theta_0 < \infty$. Suppose that*

$$\|\varepsilon\|_{W_s^{2,1}(\Omega_T)} + \|\theta\|_{W_2^{2,1}(\Omega_T)} + \|\theta\|_{L^\infty(\Omega_T)} \le \Lambda \tag{3.12}$$

holds for any $s \in (1,\infty)$. Then $\theta \in \mathrm{C}^{\alpha,\alpha/2}(\Omega_T)$ with Hölder exponent $\alpha \in (0,1)$ depending on Λ and k.

LEMMA 3.6. *Assume that* (3.12) *holds. Then for any $(u_0,u_1,\theta_0) \in U(p,q)$ and $5 < p,q < \infty$ we have*

$$\|(u,\theta)\|_{V_T(p,q)} = \|u\|_{W_p^{4,2}(\Omega_T)} + \|\theta\|_{W_q^{2,1}(\Omega_T)} \le \Lambda,$$

where Λ depends on $\|(u_0,u_1,\theta_0)\|_{U(p,q)}$ and T.

Proof. We can construct a unique timelocal solution $(u,\theta) \in W_p^{4,2}(\Omega_{\widetilde{T}}) \times W_q^{2,1}(\Omega_{\widetilde{T}})$ of $(\mathrm{TE})_3^L$ for sufficiently small $\widetilde{T} < T$, using the result of Clément and Li [7] (see also [27, Lemma 3.3.7]). Then from the embedding we have $\theta \in \mathrm{C}([0,\widetilde{T}] \times \Omega)$. By combining this regularity result with Lemma 3.5, we obtain $\theta \in \mathrm{C}^{\alpha,\alpha/2}([0,T] \times \Omega)$.

For brevity of notation we denote $c_v - \theta G''(\theta)H(\varepsilon)$ by $c_0(\varepsilon,\theta)$, and $\theta G'(\theta)\partial_t H(\varepsilon) + \nu(A\varepsilon_t) : \varepsilon_t$ by $R(\varepsilon,\theta)$. Then (1.3) can be rewritten as

$$c_0(\varepsilon_0,\theta_0)\theta_t - \Delta\theta = (c_0(\varepsilon_0,\theta_0) - c_0(\varepsilon,\theta))\theta_t + R(\varepsilon,\theta).$$

By the assumptions we have

$$\begin{aligned}\|R(\varepsilon,\theta)\|_{L^q(\Omega_T)} &\le C\|\theta\|_{L^\infty(\Omega_T)}^r \|H_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \|\varepsilon_t\|_{L^q(\Omega_T)} + C\|\varepsilon_t\|_{L^{2q}(\Omega_T)}^2 \\ &\le \Lambda.\end{aligned}$$

From the Hölder continuity it follows that

$$\|c_0(\varepsilon_0,\theta_0) - c_0(\varepsilon,\theta)\|_{L^\infty(\Omega_{T_1})} \le K T_1^{\frac{\alpha}{2}},$$

where K is the Hölder constant independent of T_1. Here $T_1 \ll T$ will be determined later.

Next, we show that $1/c_0(\varepsilon,\theta)(x,T_2)$ is Hölder continuous with respect to the space variable for T_2 fixed in $[0,T]$. We remark that

$$\mathcal{G}(y) := yG''(y) \le M$$

and $\mathcal{G} \in \mathrm{C}^1$ is Lipschitz continuous. Then we have

$$\begin{aligned}
&\left|\frac{1}{c_0}(x,T_2)-\frac{1}{c_0}(x',T_2)\right| \\
&= \left|\frac{\mathcal{G}(\theta(x',T_2))H(\varepsilon(x',T_2))-\mathcal{G}(\theta(x,T_2))H(\varepsilon(x,T_2))}{\{c_v-\mathcal{G}(\theta(x,T_2))H(\varepsilon(x,T_2))\}\{c_v-\mathcal{G}(\theta(x',T_2))H(\varepsilon(x',T_2))\}}\right| \\
&\le \frac{1}{c_v^2}|\{\mathcal{G}(\theta(x',T_2))H(\varepsilon(x',T_2))-\mathcal{G}(\theta(x,T_2))H(\varepsilon(x',T_2))\} \\
&\qquad +\{\mathcal{G}(\theta(x,T_2))H(\varepsilon(x',T_2))-\mathcal{G}(\theta(x,T_2))H(\varepsilon(x,T_2))\}| \\
&\le \frac{1}{c_v^2}|H(\varepsilon(x',T_2))|\,|\mathcal{G}(\theta(x',T_2))-\mathcal{G}(\theta(x,T_2))| \\
&\qquad +\frac{1}{c_v^2}|\mathcal{G}(\theta(x,T_2))|\,|H(\varepsilon(x',T_2))-H(\varepsilon(x,T_2))| \\
&\le \Lambda K|x-x'|^\alpha + CM|x-x'|^\alpha \\
&\le \Lambda|x-x'|^\alpha,
\end{aligned}$$

where Λ is independent of T_2. Therefore $[1/c_0(\varepsilon,\theta)](x,T_2)$ is Hölder continuous for any $T_2 \in [0,T]$. Moreover, we have $\sup_{\Omega_T}[1/c_0(\varepsilon,\theta)] \ge 1/(c_v+M\Lambda)$. These conditions ensure that $\frac{1}{c_0(\varepsilon(T_2),\theta(T_2))}\Delta$ has the maximal regularity property according to (2.3). Hence, taking $T_1 = (\frac{1}{2\Lambda(K,M,T)K})^{\frac{1}{\alpha}}$, we have

$$\begin{aligned}
\|\theta\|_{W_q^{2,1}(\Omega_{T_1})} &\le \Lambda(K,M,T)\Big(\|c_0(\varepsilon_0,\theta_0)-c_0(\varepsilon,\theta)\|_{L^\infty(\Omega_{T_1})}\|\theta_t\|_{L^q(\Omega_{T_1})} \\
&\qquad +\|R(\varepsilon,\theta)\|_{L^q(\Omega_{T_1})}+\|\theta_0\|_{B^{2-2/q}_{q,q}(\Omega)}\Big) \\
&\le \frac{1}{2}\|\theta_t\|_{L^q(\Omega_{T_1})}+\Lambda+\Lambda\|\theta_0\|_{B^{2-2/q}_{q,q}(\Omega)},
\end{aligned}$$

which yields

$$\|\theta\|_{W_q^{2,1}(\Omega_{T_1})} \le \Lambda+\Lambda\|\theta_0\|_{B^{2-2/q}_{q,q}(\Omega)}.$$

Here we remark that

$$\|\theta(T_1)\|_{B^{2-2/q}_{q,q}} \le C(T_1)\|\theta\|_{W_q^{2,1}(\Omega_{T_1})} \le C(T_1)\left(\Lambda+\Lambda\|u_0\|_{B^{2-2/q}_{q,q}}\right)$$

thanks to the embedding $W_q^{2,1}(\Omega_{T_1}) \hookrightarrow BUC([0,T_1],B^{2-\frac{2}{q}}_{q,q})$ (see [4], [18]). Then similarly for the interval $[T_1,2T_1]$ we have

$$\|\theta\|_{W_q^{2,1}(\Omega_{[T_1,2T_1]})} \le \Lambda+\Lambda\|u(T_1)\|_{B^{2-2/q}_{q,q}} \le \Lambda+\Lambda\|u_0\|_{B^{2-2/q}_{q,q}} \le \Lambda.$$

Repeating the same operation yields

$$\|\theta\|_{W_q^{2,1}(\Omega_{[kT_1,(k+1)T_1]})} \le \Lambda.$$

Summing the inequalities from $k=0$ to $k=m$ satisfying $(m+1)T_1 > T$ and $mT_1 \le T$, we conclude that

$$\|\theta\|_{W_q^{2,1}(\Omega_T)} \le \Lambda.$$

Next we estimate the norm $\|u\|_{W_p^{4,2}(\Omega_T)}$. From Lemma 2.2 it follows that

$$\|\nabla\theta\|_{L^\infty(\Omega_T)} + \|\nabla\varepsilon\|_{L^\infty(\Omega_T)} \le \Lambda$$

for $q > 5$. Therefore, by virtue of the maximal regularity (2.1), we have

$$\begin{aligned}
&\|u\|_{W_p^{4,2}(\Omega_T)} \\
&\quad\le C\|(u_0,u_1,0)\|_{U(p,q)} + \|\nabla\cdot(G(\theta)H_{,\varepsilon}(\varepsilon))\|_{L^p(\Omega_T)} + \|\nabla\cdot\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^p(\Omega_T)} \\
&\quad\le C\|(u_0,u_1,0)\|_{U(p,q)} + \Lambda\|\nabla\theta\|_{L^\infty(\Omega_T)}\|G'(\theta)\|_{L^\infty(\Omega_T)}\|H_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \\
&\qquad + \Lambda\|\theta\|^r_{L^\infty(\Omega_T)}\|\nabla\varepsilon\|_{L^\infty(\Omega_T)}\|H_{,\varepsilon\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} + \Lambda\|\nabla\varepsilon\|_{L^\infty(\Omega_T)}\|\overline{H}_{,\varepsilon\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \\
&\quad\le \Lambda(\|(u_0,u_1,0)\|_{U(p,q)}),
\end{aligned}$$

which completes the proof. □

Proof of Theorem 3.1 (*continuation*). The assumption (L4) is satisfied thanks to Lemma 3.6 and estimate (3.3). Then the existence of a solution to problem $(\mathrm{TE})_3^L$ results from Theorem 2.4. Noting that Γ_L is Lipschitz continuous, we can obtain the uniqueness result by repeating the arguments of [22, section 6]. We remark also that the assumption $p \le q$ is required to show (L1); see [21]. Thereby the proof of Theorem 3.1 is completed. □

4. Proof of Theorem 1.1 (existence). The idea of the proof consists of showing that the solution (u_L, θ_L) to $(\mathrm{TE})_3^L$ constructed in section 3 satisfies also the original system (1.2)–(1.5) for sufficiently large truncation size L. To this purpose, assuming that there exists a sufficiently smooth solution of problem (1.2)–(1.5) such that $\theta \ge 0$, we derive for it a sequence of a priori estimates which are independent of L.

LEMMA 4.1 (energy inequality). *Assume that* $\theta \ge 0$ *a.e. in* Ω_T, $K_2 \le 6$, *and* $6r + K_1 \le 6$. *Then for any* $t \in [0,T]$ *a smooth solution of* (1.2)–(1.5) *satisfies*

$$\|\theta(t)\|_{L^1(\Omega)} + \|u_t(t)\|_{L^2(\Omega)} + \|Qu(t)\|_{L^2(\Omega)} \le C(\|(u_0,u_1,\theta_0)\|_{H^2\times L^2\times L^1}). \tag{4.1}$$

Proof. Multiplying (1.2) by u_t and integrating the resulting equation with respect to the space variable, we have

$$\frac{d}{dt}\left(\frac{1}{2}\|u_t\|^2_{L^2} + \frac{\kappa}{2}\|Qu\|^2_{L^2} + \int_\Omega \overline{H}(\varepsilon)dx\right) + \nu\int_\Omega (A\varepsilon_t) : \varepsilon_t dx = -\int_\Omega G(\theta)\frac{\partial}{\partial t}H(\varepsilon)dx.$$

Integrating (1.3) over Ω, we obtain

$$c_v\frac{d}{dt}\int_\Omega \theta dx = \nu\int_\Omega (A\varepsilon_t) : \varepsilon_t dx + \int_\Omega \theta G'(\theta)\frac{\partial}{\partial t}H(\varepsilon)dx + \int_\Omega \theta G''(\theta)\theta_t H(\varepsilon)dx.$$

Combining these equalities, we deduce

$$\begin{aligned}
&\frac{d}{dt}\left(\frac{1}{2}\|u_t\|^2_{L^2} + \frac{\kappa}{2}\|Qu\|^2_{L^2} + c_v\int_\Omega \theta dx + \int_\Omega \overline{H}(\varepsilon)dx\right) \\
&\quad= \int_\Omega \left(\theta G'(\theta)\frac{\partial}{\partial t}H(\varepsilon) + \theta G''(\theta)\theta_t H(\varepsilon) - G(\theta)\frac{\partial}{\partial t}H(\varepsilon)\right)dx \\
&\quad= -\frac{d}{dt}\int_\Omega \overline{G}(\theta)H(\varepsilon)dx,
\end{aligned}$$

where $\overline{G}(\theta) = G(\theta) - \theta G'(\theta)$. Consequently,

$$\frac{d}{dt}\left(\frac{1}{2}\|u_t\|_{L^2}^2 + \frac{\kappa}{2}\|Qu\|_{L^2}^2 + c_v\int_\Omega \theta dx + \int_\Omega \overline{H}(\varepsilon)dx + \int_\Omega \overline{G}(\theta)H(\varepsilon)dx\right) = 0.$$

Here we recall that $\theta \geq 0$ and $H(\varepsilon) \geq 0$. By the structure of $G(\theta)$, the function $\overline{G}(\theta)$ is as follows:

$$\overline{G}(r) = \begin{cases} 0 & \text{if } \theta \in [0, \theta_1], \\ \varphi(\theta) - \theta\varphi'(\theta) & \text{if } \theta \in [\theta_1, \theta_2], \\ C_2(1-r)\theta^r & \text{if } \theta \in [\theta_2, \infty). \end{cases}$$

According to Lemma 2.3 we have $\overline{G}(\theta) \geq 0$. Consequently, it follows from (A)(iii) that

$$\begin{aligned} &\frac{1}{2}\|u_t(t)\|_{L^2}^2 + \frac{\kappa}{2}\|u(t)\|_{H^2}^2 + c_v\|\theta(t)\|_{L^1} \\ &\quad \leq \frac{1}{2}\|u_0\|_{H^2}^2 + \frac{\kappa}{2}\|u_1\|_{L^2}^2 + c_v\|\theta_0\|_{L^1} + \int_\Omega |\overline{H}(\varepsilon_0)|dx + C_3|\Omega| \\ &\qquad + \int_{\{\theta_2 \geq \theta_0 \geq \theta_1\}\cap\Omega} [\varphi(\theta_0) - \theta_0\varphi'(\theta_0)]H(\varepsilon_0)dx + C_2(1-r)\int_{\{\theta_0 > \theta_2\}\cap\Omega} \theta_0^r H(\varepsilon_0)dx, \end{aligned}$$

where $\varepsilon_0 = \varepsilon(u_0)$. Since the smooth function $\varphi(s) - s\varphi'(s)$ is bounded for $s \in [\theta_1, \theta_2]$, it follows that

$$\begin{aligned} \int_{\{\theta_2 \geq \theta_0 \geq \theta_1\}\cap\Omega} [\varphi(\theta_0) - \theta_0\varphi'(\theta_0)]H(\varepsilon_0)dx &\leq C\int_\Omega |\varepsilon_0|^{K_1}dx \\ &\leq C\|u_0\|_{H^2}^{K_1} \end{aligned}$$

for $K_1 \leq 6$,

$$\begin{aligned} \int_{\{\theta_0 > \theta_2\}\cap\Omega} \theta_0^r H(\varepsilon_0)dx &\leq C\|\theta_0\|_{L^1}^r \|\varepsilon_0\|_{L^{\frac{K_1}{1-r}}}^{K_1} \\ &\leq C\|\theta_0\|_{L^1}^r \|u_0\|_{H^2}^{K_1} \end{aligned}$$

for $6r + K_1 \leq 6$, and

$$\int_\Omega |\overline{H}(\varepsilon_0)|dx \leq \|u_0\|_{H^2}^{K_2}$$

for $K_2 \leq 6$. Hence, we conclude the assertion. □

LEMMA 4.2. *Let T be fixed. Assume that $\theta \geq 0$ a.e. in Ω_T and (1.6) holds. Then for any $(u_0, u_1, \theta_0) \in B^{19/8}_{16/5,16/5} \times B^{3/8}_{16/5,16/5} \times L^2 =: U_3$, the solution (u, θ) to (1.2)–(1.5) satisfies*

$$\|\varepsilon\|_{W^{2,1}_{16/5}(\Omega_T)} + \|\nabla\theta\|_{L^2(\Omega_T)} + \|\theta\|_{L^\infty_T L^2} \leq \Lambda, \tag{4.2}$$

where Λ depends on T and $\|(u_0, u_1, \theta_0)\|_{U_3}$. Moreover,

$$\|\varepsilon\|_{L^\infty(\Omega_T)} + \|\theta\|_{L^{10/3}(\Omega_T)} \leq \Lambda. \tag{4.3}$$

Proof. We remark that $\|(u_0,u_1,\theta_0)\|_{H^2\times L^2\times L^1}\le C\|(u_0,u_1,\theta_0)\|_{U_3}$ (see [1]). From the Gagliardo–Nirenberg inequality and Lemma 4.1 it follows that

$$\begin{aligned}\|\varepsilon\|_{L^{5p}(\Omega_T)} &\le C\left\|\|\varepsilon\|^{\frac{4}{5}}_{L^6(\Omega)}\|\varepsilon\|^{\frac{1}{5}}_{W^2_p(\Omega)}\right\|_{L^{5p}_T}\\ &\le C\|\varepsilon\|^{\frac{4}{5}}_{L^\infty_T L^6}\|\varepsilon\|^{\frac{1}{5}}_{W^{2,1}_p(\Omega_T)}\\ &\le C\|u\|^{\frac{4}{5}}_{L^\infty_T H^2}\|\varepsilon\|^{\frac{1}{5}}_{W^{2,1}_p(\Omega_T)}\\ &\le C\|\varepsilon\|^{\frac{1}{5}}_{W^{2,1}_p(\Omega_T)}\end{aligned} \tag{4.4}$$

and

$$\begin{aligned}\|\theta\|_{L^{\frac{8}{3}}(\Omega_T)} &\le C\left\|\|\theta\|^{\frac{1}{4}}_{L^1(\Omega)}\|\theta\|^{\frac{3}{4}}_{H^1(\Omega)}\right\|_{L^\infty_T}\\ &\le C\|\theta\|^{\frac{1}{4}}_{L^\infty_T L^1}\|\theta\|^{\frac{3}{4}}_{L^2_T H^1}\\ &\le \Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L^\infty_T L^2})^{\frac{3}{4}}.\end{aligned} \tag{4.5}$$

It follows from (4.4) that

$$\|\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^{\frac{16}{5}}(\Omega_T)}\le C\|\varepsilon\|^{K_2-1}_{L^{16}(\Omega_T)}\le\Lambda\|\varepsilon\|^{\frac{K_2-1}{5}}_{W^{2,1}_{\frac{16}{5}}(\Omega_T)}\le\frac{1}{4}\|\varepsilon\|_{W^{2,1}_{\frac{16}{5}}(\Omega_T)}+\Lambda$$

for $K_2\in[1,6)$, and

$$\|\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^{\frac{16}{5}}(\Omega_T)}\le M|\Omega_T|^{\frac{5}{16}}\le\Lambda$$

for $K_2\in[0,1)$.

We first consider the case of $K_1\ge 1$. Applying the growth condition and the Young inequality, we have

$$\begin{aligned}&\|G(\theta)H_{,\varepsilon}(\varepsilon)\|_{L^{\frac{16}{5}}(\Omega_T)}\\ &\quad\le\|\theta\|^r_{L^{\frac{8}{3}}(\Omega_T)}\|\varepsilon\|^{K_1-1}_{L^{\frac{16(K_1-1)}{5-6r}}(\Omega_T)}+\sup_{\theta\in[0,\theta_2]}|G(\theta)|\|\varepsilon\|^{K_1-1}_{L^{\frac{16(K_1-1)}{5}}(\Omega_T)}\\ &\quad\le\Lambda\|\theta\|^r_{L^{\frac{8}{3}}(\Omega_T)}\|\varepsilon\|^{K_1-1}_{L^{16}(\Omega_T)}+\Lambda\|\varepsilon\|^{K_1-1}_{L^{16}(\Omega_T)}\end{aligned}$$

for $6r+K_1\le 6$ (and $K_1\le 6$). Then

$$\begin{aligned}&\|\theta\|^r_{L^{8/3}(\Omega_T)}\|\varepsilon\|^{K_1-1}_{L^{16}(\Omega_T)}+\|\varepsilon\|^{K_1-1}_{L^{16}(\Omega_T)}\\ &\quad\le\Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L^\infty_T L_2})^{3r/4}\|\varepsilon\|^{(K_1-1)/5}_{W^{2,1}_{16/5}(\Omega_T)}+\Lambda\|\varepsilon\|^{(K_1-1)/5}_{W^{2,1}_{16/5}(\Omega_T)}\\ &\quad\le\frac{1}{4}\|\varepsilon\|_{W^{2,1}_{16/5}(\Omega_T)}+\Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L^\infty_T L_2})^{15r/4(6-K_1)}+\Lambda\end{aligned}$$

for $6r+K_1<6$ (and $K_1<6$). From the maximal regularity (2.2) it follows that

$$\begin{aligned}&\|\varepsilon\|_{W^{2,1}_{16/5}(\Omega_T)}\\ &\quad\le C\|(u_0,u_1,\theta_0)\|_{U_3}+\|G(\theta)H_{,\varepsilon}(\varepsilon)\|_{L^{16/5}(\Omega_T)}+\|\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^{16/5}(\Omega_T)}\\ &\quad\le C\|(u_0,u_1,\theta_0)\|_{U_3}+\Lambda+\Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L^\infty_T L_2})^{15r/4(6-K_1)}.\end{aligned} \tag{4.6}$$

Next, multiplying (1.3) by θ and integrating over Ω, we have

$$
\begin{aligned}
&\frac{c_v}{2}\frac{d}{dt}\|\theta(t)\|_{L^2}^2 + k\|\nabla\theta\|_{L^2}^2 \\
&\quad = \int_\Omega \theta^2 G''(\theta)\theta_t H(\varepsilon)dx + \int_\Omega \theta^2 G'(\theta)\partial_t H(\varepsilon)dx + \nu\int_\Omega \theta A\varepsilon_t : \varepsilon_t dx \\
&\quad = \int_\Omega G_2'(\theta)\theta_t H(\varepsilon)dx + \int_\Omega G_2(\theta)\partial_t H(\varepsilon)dx + 2\int_\Omega \overline{G}_2(\theta)\partial_t H(\varepsilon)dx \\
&\qquad + \nu\int_\Omega \theta A\varepsilon_t : \varepsilon_t dx \\
&\quad = \frac{d}{dt}\int_\Omega G_2(\theta)H(\varepsilon)dx + 2\int_\Omega \overline{G}_2(\theta)\partial_t H(\varepsilon)dx + \nu\int_\Omega \theta A\varepsilon_t : \varepsilon_t dx,
\end{aligned}
\tag{4.7}
$$

where $G_2(\theta)$ and $\overline{G}_2(\theta)$ are given in the proof of Lemma 3.2. Recall that

$$
G_2(\theta) = \frac{C_2 r(r-1)}{r+1}\theta^{r+1} \le 0 \quad \text{and} \quad \overline{G}_2(\theta) = \frac{2C_2 r}{r+1}\theta^{r+1} \quad \text{for } \theta \ge \theta_2,
$$

and

$$
\sup_{\theta\in[0,\theta_2]} |G_2(\theta)| + \sup_{\theta\in[0,\theta_2]} |\overline{G}_2(\theta)| =: M < \infty.
$$

Then we have

$$
\begin{aligned}
&-\int_\Omega G_2(\theta)H(\varepsilon)dx \\
&\quad = -\int_{\Omega\cap\{\theta\ge\theta_2\}} G_2(\theta)H(\varepsilon)dx - \int_{\Omega\cap\{\theta_1\le\theta\le\theta_2\}} G_2(\theta)H(\varepsilon)dx \\
&\quad \ge -M\int_\Omega |H(\varepsilon)|dx.
\end{aligned}
$$

Hence, integrating (4.7) with respect to the time variable, we obtain

$$
\begin{aligned}
&\frac{c_v}{2}\|\theta\|_{L_T^\infty L^2}^2 + k\|\nabla\theta\|_{L^2(\Omega_T)}^2 \\
&\quad \le \frac{c_v}{2}\|\theta_0\|_{L^2}^2 + \|\overline{G}_2(\theta)\partial_t H(\varepsilon)\|_{L^1(\Omega_T)} + \nu\|\theta A\varepsilon_t : \varepsilon_t\|_{L^1(\Omega_T)} \\
&\qquad + M\sup_{t\in[0,T]}\int_\Omega |H(\varepsilon(t))|dx + \int_\Omega |G_2(\theta_0)H(\varepsilon_0)|dx.
\end{aligned}
$$

By (4.4), (4.5), and the assumptions we infer that

$$
\begin{aligned}
\|\theta^{r+1}\partial_t H(\varepsilon)\|_{L^1(\Omega_T)} &\le \Lambda\|\theta\|_{L^{8/3}(\Omega_T)}^{r+1}\|u\|_{W^{2,1}_{16/5}(\Omega_T)}\|\varepsilon\|_{L^{16}(\Omega_T)}^{K_1-1} \\
&\le \Lambda(\|\nabla\theta\|_{L^2(\Omega_T)} + \|\theta\|_{L_T^\infty L_2})^{\frac{3(r+1)}{4}}\|u\|_{W^{2,1}_{16/5}(\Omega_T)}^{1+(K_1-1)/5},
\end{aligned}
$$

$$
\begin{aligned}
\|\theta A\varepsilon_t : \varepsilon_t\|_{L^1(\Omega_T)} &\le C\|\theta\|_{L^{\frac{8}{3}}(\Omega_T)}\|\varepsilon_t\|_{L^{\frac{16}{5}}(\Omega_T)}^2 \\
&\le \Lambda(\|\nabla\theta\|_{L^2(\Omega_T)} + \|\theta\|_{L_T^\infty L_2})^{\frac{3}{4}}\|\varepsilon_t\|_{\frac{16}{5}(\Omega_T)}^2,
\end{aligned}
$$

$$
\int_\Omega |H(\varepsilon(t))|dx \le C\|u(t)\|_{H^2}^{K_1} \le \Lambda,
$$

and

$$\begin{aligned}\|\theta_0^{r+1}H(\varepsilon_0)\|_{L^1(\Omega)} &\le C\|\theta_0\|_{L^2(\Omega)}^{r+1}\|\varepsilon_0\|_{L^{\frac{2K_1}{1-r}}(\Omega)}^{K_1}\\ &\le C\|\theta_0\|_{L^2(\Omega)}^{r+1}\|u_0\|_{H^2(\Omega)}^{K_1}.\end{aligned}$$

Consequently, we arrive at

$$\begin{aligned}&\|\theta\|_{L_T^\infty L^2}^2+\|\nabla\theta\|_{L^2(\Omega_T)}^2\\ &\quad\le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_3})+\Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L_T^\infty L_2})^{\frac{3(r+1)}{4}}\|\varepsilon\|_{W^{2,1}_{\frac{16}{5}}(\Omega_T)}^{\frac{4}{5}+\frac{K_1}{5}}\\ &\qquad+\Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L_T^\infty L_2})^{\frac{3}{4}}\|\varepsilon_t\|_{L^{\frac{16}{5}}(\Omega_T)}^2.\end{aligned}\tag{4.8}$$

Substituting (4.6) into (4.8) yields

$$\begin{aligned}\|\theta\|_{L_T^\infty L^2}^2+\|\nabla\theta\|_{L^2(\Omega_T)}^2 &\le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_3})\\ &+\Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L_T^\infty L_2})^{\frac{3(r+1)}{4}}\left(\|(u_0,u_1,\theta_0)\|_{U_3}+\|\nabla\theta\|_{L^2(\Omega_T)}^{\frac{15r}{4(6-K_1)}}\right)^{\frac{4}{5}+\frac{K_1}{5}}\\ &+\Lambda(\|\nabla\theta\|_{L^2(\Omega_T)}+\|\theta\|_{L_T^\infty L_2})^{\frac{3}{4}}\left(\|(u_0,u_1,\theta_0)\|_{U_3}+\|\nabla\theta\|_{L^2(\Omega_T)}^{\frac{15r}{4(6-K_1)}}\right)^2.\end{aligned}$$

Here from the assumption $6r+K_1<6$ it follows that

$$\begin{aligned}\frac{3(r+1)}{4}+\frac{15r}{4(6-K_1)}\left(\frac{4}{5}+\frac{K_1}{5}\right)=\frac{30r+3(6-K_1)}{4(6-K_1)}<\frac{5(6-K_1)+3(6-K_1)}{4(6-K_1)}=2,\\ \frac{3}{4}+\frac{30r}{4(6-K_1)}<\frac{3}{4}+\frac{5}{4}=2.\end{aligned}$$

Thus we conclude that

$$\|\theta\|_{L_T^\infty L^2}+\|\nabla\theta\|_{L^2(\Omega_T)}\le\Lambda(\|(u_0,u_1,\theta_0)\|_{U_3})+\Lambda\|\nabla\theta\|_{L^2(\Omega_T)}^{1-}.$$

Here we use $p-$ to denote a number less than p. Hence, by the Young inequality, we have

$$\|\theta\|_{L_T^\infty L^2}+\frac{1}{2}\|\theta\|_{L^2(\Omega_T)}\le\Lambda(\|(u_0,u_1,\theta_0)\|_{U_3}).$$

Substituting the above inequality into (4.6), we deduce also the following:

$$\|\varepsilon\|_{W^{2,1}_{16/5}(\Omega_T)}\le\Lambda(\|(u_0,u_1,\theta_0)\|_{U_3}).$$

Next, we consider the case when $0\le K_1\le 1$ and $0\le r<5/6$. In this case it follows that

$$|H_{,\varepsilon}(\varepsilon)|\le C<\infty.$$

By an argument similar to the one presented above we have

$$\begin{aligned}\|\varepsilon\|_{W^{2,1}_{16/5}(\Omega_T)} &\le \|(u_0,u_1,0)\|_{U_3}+\|G(\theta)H_{,\varepsilon}(\varepsilon)\|_{L^{16/5}(\Omega_T)}\\ &\le \|(u_0,u_1,0)\|_{U_3}+C\|\theta\|_{L^{\frac{16r}{5}}(\Omega_T)}^r+C\sup_{\theta\in[0,\theta_2]}G(\theta)\\ &\le \|(u_0,u_1,0)\|_{U_3}+\Lambda\|\theta\|_{L^{\frac{8}{3}}(\Omega_T)}^r+C.\end{aligned}\tag{4.9}$$

Noting that

$$\|\theta^{r+1}\partial_t H(\varepsilon)\|_{L^1(\Omega_T)} \le \Lambda\|\theta\|^{r+1}_{L^{8/3}(\Omega_T)}\|u\|_{W^{2,1}_{16/5}(\Omega_T)},$$

we obtain

$$\begin{aligned}
&\|\theta\|^2_{L^\infty_T L^2} + \|\nabla\theta\|^2_{L^2(\Omega_T)}\\
&\quad\le \|\theta_0\|^2_{L^2} + \|\theta^{r+1}\partial_t H(\varepsilon)\|_{L^1(\Omega_T)} + \|\theta A\varepsilon_t : \varepsilon_t\|_{L^1(\Omega_T)}\\
&\qquad + M\sup_{t\in[0,T]}\int_\Omega |H(\varepsilon(t))|dx + \int_\Omega |G_2(\theta_0)H(\varepsilon_0)|dx\\
&\quad\le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_3}) + \Lambda\|\theta\|^{r+1}_{L^{8/3}(\Omega_T)}\|u\|_{W^{2,1}_{16/5}(\Omega_T)}\\
&\qquad + C\|\theta\|_{L^{8/3}(\Omega_T)}\|u\|^2_{W^{2,1}_{16/5}(\Omega_T)}\\
&\quad\le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_3}) + \Lambda(\|\nabla\theta\|_{L^2(\Omega_T)} + \|\theta\|_{L^\infty_T L^2})^{3(2r+1)/4}.
\end{aligned}$$

Since $3(2r+1)/4 < 2$, we arrive at the desired estimate (4.2).

The estimate (4.3) follows with the help of the embeddings

$$\|\varepsilon\|_{L^\infty(\Omega_T)} \le \Lambda\|\varepsilon\|_{W^{2,1}_{16/5}(\Omega_T)}$$

and the inequality

$$\|\theta\|_{L^{10/3}(\Omega_T)} \le C\left\|\|\theta\|^{2/5}_{L^2(\Omega)}\|\theta\|^{3/5}_{H^1(\Omega)}\right\|_{L^{10/3}_T} \le C\|\theta\|^{2/5}_{L^\infty_T L^2}\|\theta\|^{3/5}_{L^2H^1}.$$

This completes the proof. □

LEMMA 4.3. *Let T be any fixed. Assume that $\theta \ge 0$ a.e. in Ω_T and (1.6) holds. Then for any $(u_0,u_1,\theta_0) \in B^{5/2}_{4,4}\times B^{1/2}_{4,4}\times H^1 = U_4$ the following estimate holds:*

$$\|\varepsilon\|_{W^{2,1}_4(\Omega_T)} + \|\nabla\theta\|_{L^\infty_T L^2} + \|\theta\|_{W^{2,1}_2(\Omega_T)} \le \Lambda,$$

where the constant Λ depends on T and $\|(u_0,u_1,\theta_0)\|_{U_4}$. Moreover, we have

$$\|\nabla\theta\|_{L^{10/3}(\Omega_T)} + \|\theta\|_{L^{10}(\Omega_T)} + \|\nabla\varepsilon\|_{L^{20}(\Omega_T)} \le \Lambda.$$

Proof. Remark that $U_4 \hookrightarrow U_3$. Using (4.3) we have

$$\|G(\theta)H_{,\varepsilon}(\varepsilon)\|_{L^4(\Omega_T)} \le \begin{cases}\Lambda\|\theta\|^r_{L^{10/3}(\Omega_T)}\|\varepsilon\|^{K_1-1}_{L^\infty(\Omega_T)} \le \Lambda & \text{if } K_1 \ge 1,\\ \Lambda\sup|H_{,\varepsilon}|\|\theta\|^r_{L^{10/3}(\Omega_T)} \le \Lambda & \text{if } K_1 \le 1\end{cases} \tag{4.10}$$

for $r \le 5/6$. Then from the maximal regularity (2.2) it follows that

$$\|\varepsilon\|_{W^{2,1}_4} \le \|(u_0,u_1,\theta_0)\|_{U_4} + \|G(\theta)H_{,\varepsilon}(\varepsilon)\|_{L^4} \le \Lambda. \tag{4.11}$$

Multiplying (1.3) by θ_t and integrating over Ω_T, we get

$$\begin{aligned}
&c_v\|\theta_t\|^2_{L^2(\Omega_T)} + \frac{k}{2}\|\nabla\theta\|^2_{L^\infty_T L^2} \\
&\quad\le \frac{k}{2}\|\theta_0\|^2_{H^1} + \iint_{\Omega_T} \theta_t^2 \theta G''(\theta) H(\varepsilon) dxdt \\
&\qquad + \iint_{\Omega_T} \theta_t \theta G'(\theta)\partial_t H(\varepsilon) dxdt + \iint_{\Omega_T} \theta_t A\varepsilon_t : \varepsilon_t dxdt \\
&\quad\le \frac{k}{2}\|\theta_0\|^2_{H^1} + C\|\theta_t\|_{L^2(\Omega_T)}\|\theta^r H_{,\varepsilon}(\varepsilon)\|_{L^4}\|\varepsilon_t\|_{L^4} + C\|\theta_t\|_{L^2}\|\varepsilon_t\|^2_{L^4} \\
&\quad\le \frac{k}{2}\|\theta_0\|^2_{H^1} + \Lambda(\|(u_0,u_1,\theta_0)\|_{U_4})\|\theta_t\|_{L^2(\Omega_T)} \\
&\quad\le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_4}) + \frac{1}{2}\|\theta_t\|^2_{L^2(\Omega_T)},
\end{aligned}$$

where we applied (3.11), (4.10), and (4.11). Therefore we arrive at

$$\|\varepsilon\|_{W^{2,1}_4(\Omega_T)} + \|\theta_t\|_{L^2(\Omega_T)} + \|\nabla\theta\|_{L^\infty_T L^2} \le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_4}). \tag{4.12}$$

Next, multiplying (1.3) by $\frac{-\Delta\theta}{c_v - \theta G''(\theta)H(\varepsilon)}$ and integrating over Ω, we obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nabla\theta(t)\|^2_{L^2} + \int_\Omega \frac{k|\Delta\theta|^2}{c_v - \theta G''(\theta)H(\varepsilon)} dx \\
&\quad\le \int_\Omega \frac{\Delta\theta}{c_v - \theta G''(\theta)H(\varepsilon)} \left(\theta G'(\theta)\partial_t H(\varepsilon) + \nu A\varepsilon_t : \varepsilon_t\right) dx.
\end{aligned}$$

Here we recall that

$$c_v \le c_v - \theta G''(\theta)H(\varepsilon) \le c_v + M\Lambda,$$

where $0 \le \sup_{\theta\ge 0}(-\theta G''(\theta)) =: M < \infty$. Then integrating the above inequality with respect to the time variable, we conclude that

$$\begin{aligned}
&\|\nabla\theta(t)\|^2_{L^2} + \frac{2k}{c_v + \Lambda M}\|\Delta\theta\|^2_{L^2(\Omega_T)} \\
&\quad\le \|\nabla\theta_0\|^2_{L^2} + \frac{k}{c_v + \Lambda M}\|\Delta\theta\|^2_{L^2(\Omega_T)} \\
&\qquad + \frac{c_v + \Lambda M}{k}\left\|\theta G'(\theta)\partial_t H(\varepsilon) + A\varepsilon_t : \varepsilon_t\right\|^2_{L^2(\Omega_T)} \\
&\quad\le \Lambda + \frac{k}{c_v + \Lambda M}\|\Delta\theta\|^2_{L^2(\Omega_T)} + \Lambda(M)\|\theta^r H_{,\varepsilon}(\varepsilon)\|_{L^4(\Omega_T)}\|\varepsilon_t\|_{L^4(\Omega_T)} \\
&\qquad + \Lambda(M)\|\varepsilon_t\|^2_{L^4(\Omega_T)} \\
&\quad\le \Lambda + \frac{k}{2(1+\Lambda M)}\|\Delta\theta\|^2_{L^2(\Omega_T)}
\end{aligned}$$

due to (4.10) and (4.11). Consequently, we obtain the first assertion.

With the help of Lemma 2.2, we also obtain estimate

$$\|\nabla\theta\|_{L^{10/3}(\Omega_T)} + \|\theta\|_{L^{10}(\Omega_T)} + \|\nabla\varepsilon\|_{L^{20}(\Omega_T)} \le \Lambda\left(\|\theta\|_{W^{2,1}_2(\Omega_T)} + \|\varepsilon\|_{W^{2,1}_4(\Omega_T)}\right) \le \Lambda,$$

which completes the proof. □

LEMMA 4.4. *Let T be arbitrary fixed and $p \in [20/9, 10/3]$ fixed. Assume that $\theta \geq 0$ a.e. in Ω_T and (1.6) holds. Then for any $(u_0, u_1, \theta_0) \in B^{4-2/p}_{p,p} \times B^{2-2/p}_{p,p} \times H^1 =: U_5(p)$, the solution (u, θ) to (1.2)–(1.5) satisfies*

$$\|u\|_{W^{4,2}_p(\Omega_T)} \leq \Lambda,$$

where Λ depends on T and $\|(u_0, u_1, \theta_0)\|_{U_5(p)}$.

Proof. Since the embedding $B^{4-\frac{2}{p}}_{p,p} \hookrightarrow B^{\frac{5}{2}}_{4,4}$ holds for any $\frac{20}{9} \leq p$, by Lemma 4.3 we find that

$$\begin{aligned}\|\varepsilon\|_{W^{2,1}_4(\Omega_T)} + \|\theta\|_{W^{2,1}_2(\Omega_T)} &\leq \Lambda \left(\|(u_0, u_1, \theta_0)\|_{B^{5/2}_{4,4} \times B^{1/2}_{4,4} \times H^1} \right) \\ &\leq \Lambda \left(\|(u_0, u_1, \theta_0)\|_{B^{4-2/p}_{p,p} \times B^{2-2/p}_{p,p} \times H^1} \right).\end{aligned}$$

For any $p \leq \frac{10}{3}$ we have

$$\begin{aligned}\|\nabla \cdot (G(\theta) H_{,\varepsilon}(\varepsilon))\|_{L^p(\Omega_T)} &\leq \Lambda \|\nabla \theta\|_{L^{10/3}(\Omega_T)} \|G'(\theta)\|_{L^\infty(\Omega_T)} \|H_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \\ &\quad + \Lambda \|\theta\|^r_{L^{10}(\Omega_T)} \|\nabla \varepsilon\|_{L^{20}(\Omega_T)} \|H_{,\varepsilon\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \\ &\leq \Lambda\end{aligned}$$

and

$$\|\nabla \cdot \overline{H}_{,\varepsilon}(\varepsilon)\|_{L^p(\Omega_T)} \leq \Lambda \|\nabla \varepsilon\|_{L^{20}(\Omega_T)} \|\overline{H}_{,\varepsilon\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \leq \Lambda,$$

thanks to Lemmas 4.2 and 4.3. Then from the maximal regularity (2.1) it follows that

$$\begin{aligned}\|u\|_{W^{4,2}_p(\Omega_T)} &\leq C \|(u_0, u_1, 0)\|_{U_5(p)} \\ &\quad + C(\|\nabla \cdot (G(\theta) H_{,\varepsilon}(\varepsilon))\|_{L^p(\Omega_T)} + \|\nabla \cdot \overline{H}_{,\varepsilon}(\varepsilon)\|_{L^p(\Omega_T)}) \\ &\leq \Lambda.\end{aligned}$$

This completes the proof. □

LEMMA 4.5. *Let T be arbitrary fixed, $l > 2$ integer, and $p \in (1, \infty)$. Assume that $\theta \geq 0$ a.e. in Ω_T and (1.6) holds. Then for any $(u_0, u_1, \theta_0) \in B^{17/5}_{10/3,10/3} \times B^{7/5}_{10/3,10/3} \times (L^l \cap H^1) =: U_6(l)$, the solution (u, θ) to (1.2)–(1.5) satisfies*

$$\|\theta\|_{L^\infty_T L^l_x} \leq \Lambda,$$

where $\Lambda = \Lambda(T, \|(u_0, u_1, \theta_0)\|_{U_6(l)})$. Moreover, if $(u_0, u_1, \theta_0) \in U_6(\infty)$, then

$$\|\theta\|_{L^\infty(\Omega_T)} \leq \Lambda,$$

where $\Lambda = \Lambda(T, \|(u_0, u_1, \theta_0)\|_{U_6(\infty)})$, and for $(u_0, u_1, \theta_0) \in (B^{3-2/p}_{p,p} \cap B^{17/5}_{10/3,10/3}) \times (B^{1-2/p}_{p,p} \cap B^{7/5}_{10/3,10/3}) \times (L^\infty \cap H^1) =: U_7(p)$, it holds that

$$\|\varepsilon\|_{W^{2,1}_p(\Omega_T)} \leq \Lambda,$$

where $\Lambda = \Lambda(T, \|(u_0, u_1, \theta_0)\|_{U_7(p)})$.

Proof. The same operation as in the proof of Lemma 3.3 yields

$$\begin{aligned}&\frac{c_v}{l} \frac{d}{dt} \|\hat{\theta}\|^l_{L^l} + k(l-1) \int_\Omega \theta^{l-2} |\nabla \theta|^2 dx \\ &\qquad = \int_\Omega \overline{G}_l(\theta) \partial_t H(\varepsilon) dx + \nu \int \theta^{l-1} A \varepsilon_t : \varepsilon_t dx.\end{aligned} \tag{4.13}$$

Here we recall that $G_l(\theta) = \theta^l G'(\theta) - \overline{G}_l(\theta)$, $\overline{G}_l(t) = l\int_0^\theta s^{l-1}G'(s)ds$ and

$$\hat{\theta} = \theta\left(1 - \frac{lG_l(\theta)H(\varepsilon)}{c_v\theta^l}\right)^{1/l} \geq \theta. \tag{4.14}$$

Since $\|H_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} = \Lambda < \infty$ from (4.3), we have

$$\begin{aligned}\left|\int_\Omega \overline{G}_l(\theta)\partial_t H(\varepsilon)dx\right| &\leq C\|\theta^{l-1}\|_{L^1(\Omega)}\|\theta\|_{L^\infty(\Omega)}\|\varepsilon_t\|_{L^\infty(\Omega)}\|H_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega)}\\ &\leq \Lambda\|\theta\|_{L^l(\Omega)}^{l-1}\|\theta\|_{H^2(\Omega)}\|\varepsilon_t\|_{L^\infty(\Omega)}.\end{aligned}$$

Therefore we conclude from (4.13) that

$$\frac{c_v}{l}\frac{d}{dt}\|\hat{\theta}\|_{L^l(\Omega)}^l \leq \Lambda\|\varepsilon_t\|_{L^\infty(\Omega)}\|\theta\|_{H^2(\Omega)}\|\theta\|_{L^l(\Omega)}^{l-1} + C\|\varepsilon_t\|_{L^\infty(\Omega)}^2\|\theta\|_{L^l(\Omega)}^{l-1}. \tag{4.15}$$

Here note that the equality $\partial_t\|\hat{\theta}\|_{L^l(\Omega)}^l = l\|\hat{\theta}\|_{L^l(\Omega)}^{l-1}\partial_t\|\hat{\theta}\|_{L^l(\Omega)}$, the Sobolev embedding, and Lemma 4.4 yield estimates

$$\begin{aligned}\|\varepsilon_t\|_{L_T^2L^\infty} \leq \Lambda\|\varepsilon_t\|_{L_T^2W^1_{10/3}} \leq \Lambda\|u\|_{W^{4,2}_{10/3}(\Omega_T)} \leq \Lambda,\\ \|\theta\|_{L_T^2H^2} \leq \|\theta\|_{W_2^{2,1}(\Omega_T)} \leq \Lambda,\end{aligned}$$

where Λ is independent of l. Thus, integrating (4.15) with respect to the time variable gives

$$\begin{aligned}\|\hat{\theta}\|_{L_T^\infty L^l} &\leq \|\hat{\theta}_0\|_{L^l} + \Lambda\|\varepsilon_t\|_{L_T^2L^\infty}\|\theta\|_{L_T^2H^2} + \Lambda\|\varepsilon_t\|_{L_T^2L^\infty}^2\\ &\leq \Lambda + \|\hat{\theta}_0\|_{L^l}.\end{aligned}$$

In view of the inequality $\hat{\theta}_0 \leq \theta_0\,(1 + lM\Lambda/c_v)^{1/l}$, the desired result can be obtained. For the $W_p^{2,1}$-norm of ε, we find that

$$\begin{aligned}\|\varepsilon\|_{W_p^{2,1}} &\leq C\|(u_0,u_1,0)\|_{U_7(p)} + \Lambda\|\theta\|_{L^\infty(\Omega_T)}^r\|H_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} + \Lambda\|\overline{H}_\varepsilon(\varepsilon)\|_{L^\infty(\Omega_T)}\\ &\leq \Lambda\end{aligned}$$

for $p \in (1,\infty)$, by virtue of the maximal regularity (2.2). This completes the proof. □

Using again Lemma 3.4, we can also prove the Hölder continuity of θ. The Hölder continuity of ε is ensured on account of Lemma 2.2. Hence from Lemma 3.6 we can obtain the bounds in higher Sobolev norms; i.e., for $5 < p, q < \infty$,

$$\|(u,\theta)\|_{V_T(p,q)} = \|u\|_{W_p^{4,2}(\Omega_T)} + \|\theta\|_{W_q^{2,1}(\Omega_T)} \leq \Lambda =: \widehat{\Lambda}, \tag{4.16}$$

where $\widehat{\Lambda}$ is independent of L.

This a priori estimate says that if there exists a solution to problem $(\mathrm{TE})_3$ such that $\theta \geq 0$, then this solution satisfies estimate (4.16). Let us consider now problem $(\mathrm{TE})_3^L$ from section 3, assuming that the truncation size L is sufficiently large such that

$$\left|\nabla\cdot[G(\theta)H_{,\varepsilon}(\varepsilon) + \overline{H}_{,\varepsilon}(\varepsilon)]\right| \leq \widehat{\Lambda}^{K_1+r-1} + \widehat{\Lambda}^{K_2-1} \ll L.$$

In this case we may regard Γ_L as the identity operator because the internal part of Γ_L in (3.1) is smaller than L. Therefore the unique solution (u_L,θ_L) to $(\mathrm{TE})_3^L$ satisfies (4.16) for large L. In other words, the $V_T(p,q)$-norm bound for (u_L,θ_L) does not depend on L. Hence (u_L,θ_L) satisfies also the original system $(\mathrm{TE})_3$.

The positivity of θ follows by the same argument as the proof of Lemma 3.1 in [22]. This completes the proof of Theorem 1.1.

5. 2-D case. In this section we consider the solvability of 2-D system $(\mathrm{TE})_2$. We prove the following theorem.

THEOREM 5.1. *Fix $4 < p \le q < \infty$. Assume that $\min_\Omega \theta_0 \ge 0$, $\nu > 0$, and* (A) *with* (1.7). *Then for any $T > 0$ and $(u_0, u_1, \theta_0) \in U(p, q)$, there exists at least one solution (u, θ) to $(\mathrm{TE})_2$ satisfying $(u, \theta) \in V_T(p, q)$.*

Moreover, if we assume $\min_\Omega \theta_0 = \theta_* > 0$, then there exists a positive constant ω such that

$$\theta \ge \theta_* \exp(-\omega t) \qquad \text{in } \Omega_T.$$

THEOREM 5.2. *In addition to the assumptions of Theorem* 1.1, *suppose that $F(\varepsilon, \theta) \in \mathrm{C}^4(\mathbb{S}^2 \times \mathbb{R}^+, \mathbb{R})$. Then the solution $(u, \theta) \in V_T(p, q)$ to $(\mathrm{TE})_2$ constructed above is unique.*

Proof of Theorem 5.1. With the exception of a priori bounds, the result follows by the same procedure as in the proof of the 3-D case. Thus it remains to check the bounds corresponding to Lemmas 4.1, 4.2, and 4.3 under assumption (A) with (1.7).

LEMMA 5.3 (energy inequality). *Assume that $\theta \ge 0$ a.e. in Ω_T and* (1.7) *holds. Then for any $t \in [0, T]$ the smooth solution of $(\mathrm{TE})_2$ satisfies*

$$\|\theta(t)\|_{L^1(\Omega)} + \|u_t(t)\|_{L^2(\Omega)} + \|Qu(t)\|_{L^2(\Omega)} \le C(\|(u_0, u_1, \theta_0)\|_{H^2 \times L^2 \times L^1}).$$

Proof. The same operation as in the proof of Lemma 4.1 yields

$$\frac{d}{dt}\left(\frac{1}{2}\|u_t\|_{L^2}^2 + \frac{\kappa}{2}\|Qu\|_{L^2}^2 + c_v \int_\Omega \theta dx + \int_\Omega \overline{H}(\varepsilon) dx + \int_\Omega \overline{G}(\theta) H(\varepsilon) dx\right) = 0,$$

where $\overline{G}(\theta) = G(\theta) - \theta G'(\theta)$. Here we recall that $\theta \ge 0$, $H(\varepsilon) \ge 0$, and $\overline{G}(\theta) \ge 0$. Consequently, it follows from (A)(iii) that

$$\begin{aligned}
&\frac{1}{2}\|u_t\|_{L_T^\infty L^2}^2 + \frac{\kappa}{2}\|u\|_{L_T^\infty H^2}^2 + c_v \|\theta\|_{L_T^\infty L^1} \\
&\quad \le \frac{\kappa}{2}\|u_0\|_{H^2}^2 + \frac{1}{2}\|u_1\|_{L^2}^2 + c_v\|\theta_0\|_{L^1} + \int_\Omega \left\{|\overline{H}(\varepsilon_0)| + |\overline{G}(\theta_0) H(\varepsilon_0)|\right\} dx + C_3|\Omega|,
\end{aligned}$$

where $\varepsilon_0 = \varepsilon(u_0)$. From the Sobolev embedding it holds that

$$\|\varepsilon_0\|_{L^s(\Omega)} \le C\|u_0\|_{H^2(\Omega)} \tag{5.1}$$

for any $s \in [1, \infty)$. Then we have

$$\begin{aligned}
\int_\Omega |\overline{G}(\theta_0) H(\varepsilon_0)| dx &\le C\|\theta_0\|_{L^1(\Omega)}^r \|\varepsilon_0\|_{L^{\frac{K_1}{1-r}}(\Omega)}^{K_1} \\
&\le C\|\theta_0\|_{L^1(\Omega)}^r \|u_0\|_{H^2}^{K_1}
\end{aligned}$$

for $r < 1$ and $K_1 < \infty$, and

$$\begin{aligned}
\int_\Omega \overline{H}(\varepsilon_0) dx &\le \|\varepsilon_0\|_{L^{K_2}}^{K_2} \\
&\le C\|u_0\|_{H^2}^{K_2}
\end{aligned}$$

for $K_2 < \infty$. This completes the proof. ∎

LEMMA 5.4. *Let T and $p \in [2,4)$ be fixed. Assume that* (1.7) *holds. Then for any* $(u_0, u_1, \theta_0) \in B^{3-2/p}_{p,p} \times B^{1-2/p}_{p,p} \times L^2 =: U_3'(p)$, *the solution* (u, θ) *to* $(\mathrm{TE})_2$ *satisfies*

$$\|\varepsilon\|_{W^{2,1}_p(\Omega_T)} + \|\nabla\theta\|_{L^2(\Omega_T)} + \|\theta\|_{L^\infty_T L^2} \le \Lambda, \tag{5.2}$$

where Λ *depends on* T *and* $\|(u_0, u_1, \theta_0)\|_{U_3'(p)}$. *Moreover, we have*

$$\|\varepsilon\|_{L^\infty(\Omega_T)} + \|\theta\|_{L^p(\Omega_T)} \le \Lambda. \tag{5.3}$$

Proof. We first show (5.2) for p such that $p < 3$. From the Sobolev inequality (5.1) and Lemma 5.3, it follows that

$$\|\varepsilon\|_{L^s(\Omega_T)} \le \Lambda \|u\|_{L^\infty_T H^2} \le \Lambda$$

for every $s < \infty$, and hence we obtain

$$\|H_{,\varepsilon}\|_{L^s(\Omega_T)} + \|\overline{H}_{,\varepsilon}\|_{L^s(\Omega_T)} \le \Lambda \tag{5.4}$$

for any K_1, $K_2 < \infty$. Moreover, by using the Hölder inequality, we have

$$\|\theta\|_{L^p(\Omega_T)} \le C \left\| \|\theta\|^{1-2/p}_{L^1} \|\theta\|^{2/p}_{L^{2/(3-p)}} \right\|_{L^p_T} \le C \|\theta\|^{1-2/p}_{L^\infty_T L^1} \|\theta\|^{2/p}_{L^2_T H^1} \le \Lambda \|\theta\|^{2/p}_{L^2_T H^1} \tag{5.5}$$

for $p \in [2,3)$.

We fix $\bar{p}$ such that $r + 2 < \bar{p} < 3$. From (5.4), (5.5), and the maximal regularity (2.2) it follows that

$$\begin{aligned}
&\|\varepsilon\|_{W^{2,1}_{\bar{p}}(\Omega_T)} \\
&\quad \le C\|(u_0, u_1, \theta_0)\|_{U_3'(\bar{p})} + C\|G(\theta)H_{,\varepsilon}(\varepsilon)\|_{L^{\bar{p}}(\Omega_T)} + C\|\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^{\bar{p}}(\Omega_T)} \\
&\quad \le \Lambda + C\|\theta\|^r_{L^{\bar{p}}(\Omega_T)} \|H_{,\varepsilon}(\varepsilon)\|_{L^{\frac{\bar{p}}{(1-r)}}(\Omega_T)} + C\|\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^{\bar{p}}(\Omega_T)} \\
&\quad \le \Lambda + \Lambda \|\theta\|^{\frac{2r}{\bar{p}}}_{L^2_T H^1}.
\end{aligned} \tag{5.6}$$

Next, the same operation as in the proof of Lemma 4.2 yields

$$\begin{aligned}
&\frac{c_v}{2}\|\theta\|^2_{L^\infty_T L^2} + k\|\nabla\theta\|^2_{L^2(\Omega_T)} \\
&\quad \le \frac{c_v}{2}\|\theta_0\|^2_{L^2} + \|\overline{G}_2(\theta)\partial_t H(\varepsilon)\|_{L^1(\Omega_T)} + \nu\|\theta A\varepsilon_t : \varepsilon_t\|_{L^1(\Omega_T)} \\
&\qquad + M \sup_{t\in[0,T]} \int_{\Omega_T} |H(\varepsilon(t))|dx + \int_\Omega |G_2(\theta_0)H(\varepsilon_0)|dx.
\end{aligned}$$

By (5.4), (5.5), and (5.6) we have

$$\begin{aligned}
\|\theta^{r+1}\partial_t H(\varepsilon)\|_{L^1(\Omega_T)} &\le \Lambda \|\theta\|^{r+1}_{L^{\bar{p}}(\Omega_T)} \|\varepsilon\|_{W^{2,1}_{\bar{p}}(\Omega_T)} \|H_{,\varepsilon}(\varepsilon)\|_{L^{\frac{\bar{p}}{\bar{p}-(r+2)}}(\Omega_T)} \\
&\le \Lambda \|\theta\|^{\frac{2(r+1)}{\bar{p}}}_{L^2_T H^1} \left(\Lambda + \|\theta\|^{\frac{2r}{\bar{p}}}_{L^2_T H^1} \right)
\end{aligned}$$

for $\bar{p} > r+2$,

$$\begin{aligned}\|\theta A\varepsilon_t : \varepsilon_t\|_{L^1(\Omega_T)} &\le \Lambda\|\theta\|_{L^{\bar{p}}(\Omega_T)}\|\varepsilon_t\|^2_{L^{\frac{2\bar{p}}{\bar{p}-1}}(\Omega_T)} \\ &\le \Lambda\|\theta\|^{\frac{2}{\bar{p}}}_{L^2_T H^1}\left(\Lambda + \|\theta\|^{\frac{4r}{\bar{p}}}_{L^2_T H^1}\right),\end{aligned}$$

$$\int_\Omega |H(\varepsilon(t))|dx \le C\|u(t)\|^{K_1}_{H^2} \le \Lambda,$$

and

$$\begin{aligned}\|\theta_0^{r+1}H(\varepsilon_0)\|_{L^1(\Omega)} &\le C\|\theta_0\|^{r+1}_{L^2(\Omega)}\|\varepsilon_0\|^{K_1}_{L^{\frac{2K_1}{1-r}}(\Omega)} \\ &\le C\|\theta_0\|^{r+1}_{L^2(\Omega)}\|u_0\|^{K_1}_{H^2(\Omega)}.\end{aligned}$$

Consequently, we arrive at

$$\|\theta\|^2_{L^\infty_T L^2} + \|\nabla\theta\|^2_{L^2(\Omega_T)} \le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_3'(p)}) + \Lambda\|\theta\|^{\frac{2(2r+1)}{\bar{p}}}_{L^2_T H^1}.$$

Since $2r+1 < r+2 < \bar{p}$, by using the Young inequality we have

$$\|\theta\|_{L^\infty_T L^2} + \|\nabla\theta\|_{L^2(\Omega_T)} \le \Lambda(\|(u_0,u_1,\theta_0)\|_{U_3'(p)}). \tag{5.7}$$

Substituting (5.7) into (5.6), we obtain (5.2) for $p < 3$.

We shall show the rest of the proof. Taking $p \in [2,4)$, by the same operation as (5.5) we have

$$\|\theta\|_{L^p(\Omega_T)} \le C\|\theta\|^{1-2/p}_{L^\infty_T L^2}\|\theta\|^{2/p}_{L^2_T H^1} \le \Lambda$$

for $p < 4$ thanks to (5.7). Then from (2.2) we conclude that

$$\begin{aligned}\|\varepsilon\|_{W^{2,1}_p} &\le \Lambda + \|\theta\|^r_{L^p}\|H_{,\varepsilon}(\varepsilon)\|_{L^{\frac{p}{(1-r)}}} + \|\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^p} \\ &\le \Lambda.\end{aligned}$$

This completes the proof. □

LEMMA 5.5. *Let T be any fixed. Assume that* (1.7) *holds. Then for any $(u_0,u_1,\theta_0) \in B^{5/2}_{4,4} \times B^{1/2}_{4,4} \times H^1 = U_4'$ the following estimate holds:*

$$\|\varepsilon\|_{W^{2,1}_4(\Omega_T)} + \|\nabla\theta\|_{L^\infty_T L^2} + \|\theta\|_{W^{2,1}_2(\Omega_T)} \le \Lambda,$$

where constant Λ depends on T and $\|(u_0,u_1,\theta_0)\|_{U_4'}$. Moreover, we have

$$\|\nabla\theta\|_{L^4(\Omega_T)} + \|\theta\|_{L^s(\Omega_T)} + \|\nabla\varepsilon\|_{L^s(\Omega_T)} \le \Lambda$$

for any $s < \infty$.

Proof. It follows from Lemma 5.4 and (2.2) that

$$\begin{aligned}\|\varepsilon\|_{W^{2,1}_4(\Omega_T)} &\le C\|(u_0,u_1,\theta_0)\|_{U_4'} + C\|\theta\|^r_{L^{4r}}\|H_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \\ &\quad + C\|\overline{H}_{,\varepsilon}(\varepsilon)\|_{L^\infty(\Omega_T)} \\ &\le \Lambda\end{aligned} \tag{5.8}$$

thanks to $r < 1$. The same operation as in the proof of Lemma 4.3 yields

$$\begin{aligned} c_v\|\theta_t\|^2_{L^2(\Omega_T)} + \frac{k}{2}\|\nabla\theta\|^2_{L^\infty_T L^2} &\le \frac{k}{2}\|\theta_0\|^2_{H^1} + C\|\theta_t\|_{L^2(\Omega_T)}\|\theta^r H_{,\varepsilon}(\varepsilon)\|_{L^4(\Omega_T)}\|\varepsilon_t\|_{L^4(\Omega_T)} \\ &\quad + C\|\theta_t\|_{L^2(\Omega_T)}\|\varepsilon_t\|^2_{L^4(\Omega_T)} \\ &\le \Lambda + \frac{c_v}{2}\|\theta_t\|^2_{L^2(\Omega_T)} \end{aligned}$$

on account of (5.3) and (5.8). Therefore, we arrive at the estimate

$$\|\varepsilon\|_{W^{2,1}_4(\Omega_T)} + \|\theta_t\|_{L^2(\Omega_T)} + \|\nabla\theta\|_{L^\infty_T L^2} \le \Lambda(\|(u_0, u_1, \theta_0)\|_{U'_4}).$$

Moreover, applying the same argument as in the proof of Lemma 4.3, we get

$$\|\Delta\theta\|_{L^2(\Omega_T)} \le \Lambda.$$

This completes the proof of the first assertion. With the help of Lemma 2.2 we obtain the second assertion. This completes the proof of Lemma 5.5. □

From a modification similar to that presented in section 4 we can derive the estimate

$$\|(u,\theta)\|_{V_T(p,q)} = \|u\|_{W^{4,2}_p(\Omega_T)} + \|\theta\|_{W^{2,1}_q(\Omega_T)} \le \Lambda.$$

Hence the proof of Theorem 5.1 are completed. □

Acknowledgment. The authors would like to express deep gratitude to Professor Yoshio Tsutsumi for valuable advice.

REFERENCES

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces* 2nd ed., Academic Press, Amsterdam, 2003.

[2] T. Aiki, *Weak solutions for Falk's model of shape memory alloys*, Math. Methods Appl. Sci., 23 (2000), pp. 299–319.

[3] T. Aiki, A. Kadoya, and S. Yoshikawa, *One-dimensional shape memory alloy problem with small viscosity,* in Mathematical Approach to Nonlinear Phenomena, GAKUTO Internat. Ser. Math. Sci. Appl. 23, Gakkotosho, Tokyo, pp. 1–8.

[4] H. Amann, *Linear and Quasilinear Parabolic Problems. Vol.* I. *Abstract Linear Theory*, Monogr. Math. 89, Birkhäuser Boston, Boston, MA, 1995.

[5] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Appl. Math. Sci. 121, Springer, Berlin, 1996.

[6] N. Bubner and J. Sprekels, *Optimal control of martensitic phase transitions in a deformation-driven experiment on shape memory alloys,* Adv. Math. Sci. Appl., 8 (1998), pp. 299–325.

[7] P. Clément and S. Li, *Abstract parabolic quasilinear equations and application to a groundwater flow problem,* Adv. Math. Sci. Appl., 3 (1993/94), pp. 17–32.

[8] C. M. Dafermos and L. Hsiao, *Global smooth thermomechanical processes in one-dimensional nonlinear thermoviscoelasticity,* Nonlinear Anal., 6 (1982), pp. 435–454.

[9] R. Denk, M. Hieber, and J. Prüss, *$\mathcal{R}$-boundedness, Fourier multipliers, and problems of elliptic and parabolic type,* Mem. Amer. Math. Soc., 166 (2003), no. 788.

[10] F. Falk, *Elastic phase transitions and nonconvex energy functions,* in Free Boundary Problems: Theory and Applications I, K.-H. Hoffmann and J. Sprekels, eds., Longman, London, 1990, pp. 45–59.

[11] F. Falk and P. Konopka, *Three-dimensional Landau theory describing the martensitic phase transformation of shape-memory alloys,* J. Phys. Condens. Matter, 2 (1990), pp. 61–77.

[12] K.-H. Hoffmann and A. Żochowski, *Existence of solutions to some nonlinear thermoelastic systems with viscosity,* Math. Methods Appl. Sci., 15 (1992), pp. 187–204.

[13] M. Hieber and J. Prüss, *Heat kernels and maximal L^p-L^q estimates for parabolic evolution equations,* Comm. Partial Differential Equations, 22 (1997), pp. 1647–1669.

[14] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type,* Trans. Math. Monogr. 23, American Mathmatical Society, Providence, RI, 1967.

[15] G. M. Lieberman, *Second Order Parabolic Differential Equations,* World Scientific, Singapore, 1996.

[16] J. Sprekels and S. Zheng, *Global solutions to the equations of a Ginzburg–Landau theory for structural phase transitions in shape memory alloys,* Phys. D, 39 (1989), pp. 59–76.

[17] J. Sprekels, S. Zheng, and P. Zhu, *Asymptotic behavior of the solutions to a Landau–Ginzburg system with viscosity for martensitic phase transitions in shape memory alloys,* SIAM J. Math. Anal., 29 (1998), pp. 69–84.

[18] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators,* North–Holland, Amsterdam, 1978.

[19] I. Pawłow, *Three-dimensional model of thermomechanical evolution of shape memory materials,* Control and Cybernet., 29 (2000), pp. 341–365.

[20] I. Pawłow and A. Żochowski, *Existence and uniqueness of solutions for a three-dimensional thermoelastic system,* Dissertationes. Math., 406 (2002), pp. 1–46.

[21] I. Pawłow and W. M. Zajączkowski, *Global existence to a three-dimensional nonlinear thermoelasticity system arising in shape memory materials,* Math. Methods Appl. Sci., 28 (2005), pp. 407–442.

[22] I. Pawłow and W. M. Zajączkowski, *Unique global solvability in two-dimensional nonlinear thermoelasticity,* Math. Methods Appl. Sci., 28 (2005), pp. 551–592.

[23] I. Pawłow and W. M. Zajączkowski, *New existence result for a 3-D shape memory model,* in Dissipative Phase Transitions, P. Colli, N. Kenmochi, and J. Sprekels, eds., Ser. Adv. Math. Appl. Sci. 71, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2006, pp. 201–224.

[24] S. Yoshikawa, *Weak solutions for the Falk model system of shape memory alloys in energy class,* Math. Methods Appl. Sci., 28 (2005), pp. 1423–1443.

[25] S. Yoshikawa, *Unique global existence for a three-dimensional thermoelastic system of shape memory alloys,* Adv. Math. Sci. Appl., 15 (2005), pp. 603–627.

[26] S. Yoshikawa, *Small energy global existence for a two-dimensional thermoelastic system of shape memory materials*, in Mathematical Approach to Nonlinear Phenomena, GAKUTO Internat. Ser. Math. Sci. Appl. 23, Gakkotosho, Tokyo, pp. 297–306.

[27] S. Yoshikawa, *Global Solutions for Shape Memory Alloy Systems*, Doctoral Thesis, Tohoku University, Tohoku, Japan, 2006.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1760–1787

NONLINEAR ASYMPTOTIC STABILITY OF THE SEMISTRONG PULSE DYNAMICS IN A REGULARIZED GIERER–MEINHARDT MODEL*

ARJEN DOELMAN†, TASSO J. KAPER‡, AND KEITH PROMISLOW§

Abstract. We use renormalization group (RG) techniques to prove the nonlinear asymptotic stability for the semistrong regime of two-pulse interactions in a regularized Gierer–Meinhardt system. In the semistrong limit the localized activator pulses interact strongly through the slowly varying inhibitor. The interaction is not tail-tail as in the weak interaction limit, and the pulse amplitudes and speeds change as the pulse separation evolves on algebraically slow time scales. In addition the point spectrum of the associated linearized operator evolves with the pulse dynamics. The RG approach employed here validates the interaction laws of quasi-steady two-pulse patterns obtained formally in the literature, and establishes that the pulse dynamics reduce to a closed system of ordinary differential equations for the activator pulse locations. Moreover, we fully justify the reduction to the nonlocal eigenvalue problem (NLEP) showing that the large difference between the quasi-steady NLEP operator and the operator arising from linearization about the pulse is controlled by the resolvent.

Key words. Gierer–Meinhardt, semistrong interaction, renormalization group, geometric singular perturbation, nonlocal eigenvalue problem (NLEP)

AMS subject classifications. 35P15, 35K57, 35K45, 35B25

DOI. 10.1137/050646883

1. Introduction.

Pulse solutions are the building blocks for the analysis of complex patterns in reaction-diffusion equations. Within the proper scaling limit, the dynamics exhibited by many reaction-diffusion systems is governed by the interactions of localized solutions of pulse type. A prototypical example is given by the spatio-temporal chaotic dynamics of the one-dimensional Gray–Scott system for which numerical simulations indicate that the chaotic dynamics originate from the interactions and bifurcations of pulse solutions [13].

In the context of singularly perturbed equations in one spatial dimension, there is a well-developed literature addressing the existence and stability of stationary pulse solutions based on the geometric singular perturbation theory and the Evans function methods (see [15, 5] and the references therein). There is no such general theory for pulse interactions. In fact, strong pulse interactions, and especially the phenomena of pulse-replication and annihilation, have been studied computationally, but are not yet understood mathematically. On the other hand, there are methods to study the behavior of pulses in the weak interaction limit where the pulses are so greatly separated that they can be considered at leading order as copies of a solitary pulse. In this regime the exponentially weak interactions affect only the position of the pulses and have no leading order influence on their shape or stability (see [8, 9, 14, 15] and

*Received by the editors December 6, 2005; accepted for publication (in revised form) September 21, 2006; published electronically February 26, 2007.
http://www.siam.org/journals/sima/38-6/64688.html
†Modeling, Analysis, and Simulation, CWI, 1098 SJ Amsterdam, the Netherlands (A.Doelman@CWI.NL). This author acknowledges support from NWO.
‡Department of Mathematics and Statistics, Boston University, Boston, MA 02215 (tasso@math.bu.edu). This author was supported by NSF-DMS grant 0306523.
§Department of Mathematics, Michigan State University, East Lansing, MI 48824 (kpromisl@math.msu.edu). This author was supported by NSF grants DMS 04055965 and DMS 0510002.

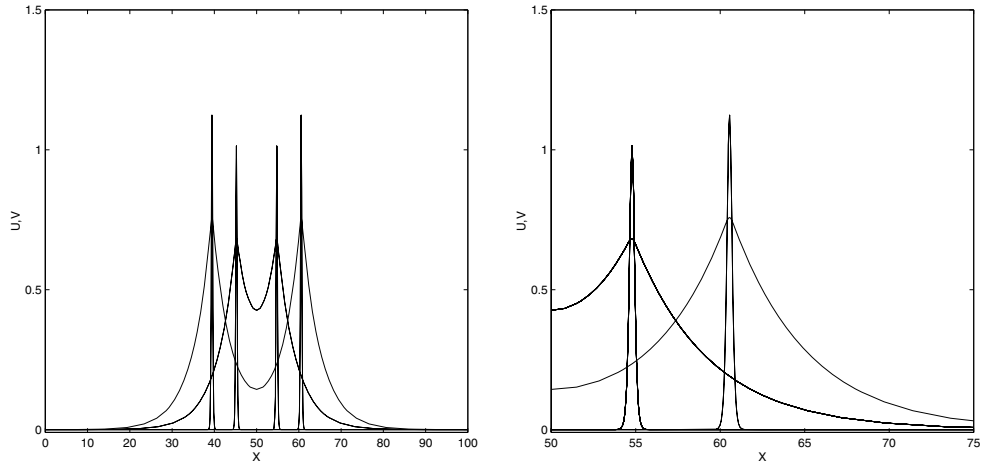


FIG. 1.1. *The two-pulse solution of the Gierer–Meinhardt equation, shown at* $t = 500$ *and* $t = 5000$*, in the slow spatial variable. The figures are obtained from numerical simulation of* (2.1) *with* $\varepsilon^2 = 0.01$ *and* $\mu = 5$.

the references therein).

Recently, an intermediate concept has been introduced in the context of singularly perturbed equations, the semistrong interaction case (see [6, 16] and the references therein). The semistrong regime exists in systems whose components decay at asymptotically distinct rates, so that some of the components of the system approach the trivial background state between pulses, while others do not. Moreover, the pulse positions, amplitudes, and shapes change at rates that are algebraically small in the perturbation parameter (see Figure 1.1) and may bifurcate due to the interactions [6, 16].

Up to now the semistrong pulse interaction has only been studied formally (see Remark 1.3). In this paper we show that the semistrong interaction fits naturally into the framework of the renormalization group (RG) methods developed to study the stability of slowly evolving patterns [14, 12]. For the Gierer–Meinhardt equations, the geometric singular perturbation theory shows that the activator-inhibitor interaction reduces the highly diffusive inhibitor to a local constant within each activator pulse. The value of this constant determines the activator pulse interactions, in particular causing the pulse amplitudes to depend upon the pulse positions, which evolve on an $\mathcal{O}(\frac{1}{\varepsilon^4})$ time scale. The RG analysis makes these statements rigorous, in particular fully justifying the reductions made in the nonlocal eigenvalue problem (NLEP) analysis which arises in the linear stability analysis of the pulses.

The singular perturbation theory typically constructs a family of pulse type patterns which are approximate solutions of a given system of equations [2, 3, 6, 11, 16]. The solutions are characterized by parameters $\vec{p} \in \mathbb{R}^k$, which are often—but not exclusively—pulse locations. The linearizing about the global manifold of slowly evolving pulse patterns for a particular choice of parameters $\vec{p}$ allows one to decompose the phase space into tangential (or active) and normal (or decaying) modes. In the RG approach, rather than using the exact linearization, reduced linearized operators are identified at a discrete family of base points on the manifold. These form a loose covering of approximate tangent planes, much like the scales of a fish form a piecewise linear envelope of the underlying body. In the current setting this

gives two specific advantages: first, in a neighborhood of each base point we identify a temporally constant linearized operator and associated phase space decomposition, and second, we are free to modify the governing linear operator in ways that simplify the analysis. For the Gierer–Meinhardt equations, the singularly perturbed structure of the linearized operators makes them strongly contractive on certain regions of the phase space. This permits a nontrivial replacement of spatially varying potentials with delta functions, affording dramatic simplification to the analysis of the principle linear operator. Indeed we replace the exact linearization with a putatively $\mathcal{O}(\varepsilon^{-2})$ "perturbation." This reduced linear operator gives rise to exactly the NLEP operators introduced previously in the formal linear stability analysis, [2, 6, 16], and justifies the observation that, at the linear level, the inhibitor equation averages perturbations into a mean-field.

The RG method shows that the NLEP operators control the flow in a neighborhood of the pulse configurations, generating a thin absorbing set in the phase space. Moreover, we recover the leading order pulse evolution by projecting the flow onto the tangent plane of the manifold of two-pulse solutions. In this paper we consider only two-pulse solutions. Although there are no new conceptual features, the generalization to N-pulses is technical. In particular, determining the amplitudes of each pulse within an N-pulse configuration requires a nontrivial nonlinear computation, and the stability of the underlying pattern will depend sensitively upon the pulse amplitudes and separations. These issues have been studied numerically in [11] and the construction and interaction of N-pulses on bounded domains has been considered in [16]. However, the nonlinear aspects of the stability approach we developed here generalizes directly from the two-pulse to the N-pulse case. Our methods can also be applied to the weak and semistrong N-pulse interactions in classes of singularly perturbed reaction-diffusion equations as considered in [6] and in [17] on bounded domains. Nevertheless, interesting additional issues may emerge in specific settings, such as the semistrong evolution of pulses in the Gray–Scott equation (see [4]) in which the essential spectrum is asymptotically close to the origin.

The main result of this paper addresses the semistrong evolution of two-pulse solutions Φ, given in (2.5), of the regularized Geirer–Meinhardt equation (2.3). The solutions are parameterized by pulse positions $\Gamma = (\Gamma_1, \Gamma_2)^t$, with pulse separation $\Delta\Gamma = |\Gamma_1 - \Gamma_2|$. In the application at hand the RG approach models the impact of transient initial perturbations on the semistrong pulse evolution, and after the decay of the transients, recovers the formal pattern evolution at leading order. For the Gierer–Meinhardt equation the semistrong regime is comprised of pulses whose separations satisfy $\Delta\Gamma > \Delta\Gamma^*(\mu)$, where $\Delta\Gamma^* = \mathcal{O}(\varepsilon^{-2})$ is defined in (3.37) of Proposition 3.4. For $\Delta\Gamma < \Delta\Gamma^*$ one enters the strong interaction regime, where the two-pulse solution has quasi-stationary eigenvalues which are incompatible with the two-pulse manifold, and pulse splitting bifurcations are observed. Our approach works uniformly for the weak, $\Delta\Gamma \gg \varepsilon^{-2}$, and semistrong interactions, recovering prior results [8, 9, 14, 15], for the weak interaction limit.

We introduce the norm $\|\cdot\|_X$ on $H^1 \times H^1$ defined by

$$\|G\|_X = \varepsilon\|G_1\|_{L^2} + \varepsilon^{-1}\|\partial_\xi G_1\|_{L^2} + \|G_2\|_{H^1}, \tag{1.1}$$

and remark that it controls the L^∞ norm uniformly,

$$\|G_1\|_{L^\infty} \leq (2\|G_1\|_{L^2}\|\partial_\xi G_1\|_{L^2})^{\frac{1}{2}} \leq \varepsilon\|G_1\|_{L^2} + \varepsilon^{-1}\|\partial_\xi G_1\|_{L^2} \leq \|G\|_X. \tag{1.2}$$

We next state our main theorem for the pulses in the semistrong interaction regime.

THEOREM 1.1. *Let ε be sufficiently small, let $\mu > \mu_{\rm Hopf}$, and let the pulse separation satisfy $\Delta\Gamma > \Delta\Gamma^*(\mu)$, where $\mu_{\rm Hopf}$ and $\Delta\Gamma^*(\mu)$ are given in Proposition* 3.4. *The manifold $\mathcal{M}$ of two-pulse solutions* (2.6) *of the regularized Gierer–Meinhardt equation* (2.3) *is asymptotically exponentially stable up to $\mathcal{O}(\varepsilon^3)$. That is, there exists M and $\nu > 0$, independent of ε, such that for all initial data $\vec{U}_0$ sufficiently close to the $\mathcal{M}$, the corresponding solution $\vec{U} = (U, V)^t$ of the regularized Gierer–Meinhardt equations can be decomposed as*

$$\vec{U}(\xi, t) = \Phi_\Gamma + W(\xi, t), \tag{1.3}$$

where the parameters $\Gamma(t)$ of the two-pulse solution Φ evolve at leading order according to (4.17). *Moreover, the remainder W satisfies*

$$\|W\|_X \leq M(e^{-\nu t}\|W_0\|_X + \varepsilon^3). \tag{1.4}$$

In particular, after the perturbation W has decayed to $\mathcal{O}(\varepsilon^3)$, the pulse evolution is given by the ordinary differential equations (4.75) *which are equivalent at leading order to*

$$\frac{d}{dt}\Delta\Gamma = \varepsilon^2\sqrt{\mu}\frac{e^{-\varepsilon^2\Delta\Gamma\sqrt{\mu}}}{1 + e^{-\varepsilon^2\Delta\Gamma\sqrt{\mu}}}. \tag{1.5}$$

Since the pulses are repelling, Theorem 1.1 governs the evolution of all two-pulse solutions in the semistrong or the weak interaction regime. That is, any two-pulse solutions with $\Delta\Gamma(0) > \Delta\Gamma^*(\mu)$ will evolve according to (1.5) for all subsequent time.

Remark 1.2. The pulse dynamics (1.5) were obtained formally in [6].

Remark 1.3. In [2, 3], slowly-modulated two-pulse solutions were constructed for the Gray–Scott model on the infinite line, along with ODEs for the pulse positions, using the method of multiple scales. Various bifurcations, including the bifurcation to self-replicating two-pulse solutions, were identified. Moreover, the NLEP method, which was initially developed for studying the stability of stationary, one-pulse solutions (see [5] and the references therein) was (formally) extended in [2] to the stability analysis of two-pulse solutions. For the generalized Gierer–Meinhardt equations on bounded domains, [11] presents ODEs for N-spike quasi-equilibrium solutions in the semistrong and weak interaction regimes. Also, the NLEP method is employed to formally derive explicitly computable stability criterions. Further formal study of the instabilities (competition and oscillatory) for two-pulse solutions of the Gierer–Meinhardt equations has been reported in [16]. The analysis is primarily on a bounded domain, and results for the infinite line—some of which extend those reported in [2, 3]—are obtained by taking the domain to be large. Semistrong pulse interactions have also been studied in [6] for a large class of reaction-diffusion equations, including for the generalized Gierer–Meinhardt equations the Gray–Scott model, the Thomas equations, the Schnakenberg model, and others. Conditions were derived to determine formally whether adjacent pulses attract or repel, and the interactions between stationary and dynamically-evolving N-pulse solutions were studied.

2. The two-pulse solutions of the Gierer–Meinhardt equations. As proposed the Gierer–Meinhardt model, [10], has an artificial singularity in its nonlinear term, which suggests infinite production of the activator, V, in the absence of the inhibitor, U. While the singular model can be studied by working with exponentially weighted norms which preserve positivity of the inhibitor, the behavior of the

model for small concentrations bears little resemblance to chemical reality. Moreover, the singularity has an exponentially small impact on both the two-pulse construction and their evolution. To avoid clouding the analysis, we truncate the superfluous singularity of the Gierer–Meinhardt reaction term, replacing it with a variation of the classic Rice–Hertzfeld mechanism typical of complex reactions with inhibition steps; see chapter 26 of [1]. In the slow spatial variable x, the regularized Gierer–Meinhardt equation is given by

$$\begin{cases} U_t = \frac{1}{\varepsilon^2} U_{xx} - \mu U + \frac{1}{\varepsilon^2} V^2, \\ V_t = \varepsilon^2 V_{xx} - V + \frac{V^2}{\kappa(U)}, \end{cases} \tag{2.1}$$

where $U(x,t), V(x,t) : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$, $\mu > 0$ is the main (bifurcation) parameter, and $\varepsilon > 0$ is asymptotically small, $0 < \varepsilon \ll 1$. The regularizing function κ takes the form

$$\kappa(s) = \begin{cases} s & \text{if } s > 2\delta, \\ \delta & \text{if } 0 < s < \delta, \end{cases} \tag{2.2}$$

and is smooth for $s \in (\delta, 2\delta)$, with derivative uniformly less than two. In the absence of the inhibitor, U, the production rate of the activator V reduces to V^2/δ, where δ is a small parameter. The regularization introduces an $\mathcal{O}(e^{-\varepsilon^{-2}|\ln \delta|})$ perturbation to the pulse dynamics. The fast spatial scale is defined by $\xi = \frac{x}{\varepsilon}$, so that (2.1) transforms into

$$\begin{cases} U_t = \frac{1}{\varepsilon^4} U_{\xi\xi} - \mu U + \frac{1}{\varepsilon^2} V^2, \\ V_t = V_{\xi\xi} - V + \frac{V^2}{\kappa(U)}. \end{cases} \tag{2.3}$$

We denote the right-hand side of (2.3) by $F(U,V)$. Since the regularizing term has only an exponentially small impact on the pulse construction we carry over the asymptotic results for the singular Gierer–Meinhardt equation without modification.

PROPOSITION 2.1. *The construction and spectral analysis of pulse solutions for the classical GM model given in* [5] *and the construction and formal dynamics of semistrong two-pulses given in* [6] *hold up to exponentially small terms for the regularized models* (2.1)–(2.3).

2.1. Notation. We write $f = g + \mathcal{O}(\varepsilon)$ in norm $\|\cdot\|$ if

$$\|f - g\| \leq c\varepsilon, \tag{2.4}$$

and assume the $\|\cdot\|_X$ norm if no norm is specified. The solution (U,V) of the Gierer–Meinhardt equation is denoted $\vec{U}$. The two-pulse solutions are denoted by $\Phi_\Gamma = (U_0 + \varepsilon^2 U_2 + \cdots, V_0 + \varepsilon^2 V_2 + \cdots)^t$, while the initial data of the Gierer–Meinhardt equation is given by $\vec{U}_0$. We denote by $\|f\|_{\widehat{L}^p}$ and $\|f\|_{\widehat{H}^2}$ the L^p and H^2 norms of the Fourier transform of f. We remark that $\|f\|_{L^\infty} \leq c\|f\|_{\widehat{L}^1}$ and, conversely, $\|f\|_{\widehat{L}^\infty} \leq c\|f\|_{L^1}$, and, in particular, that the delta function resides in $\widehat{L}^\infty$ but is not in L^1. Also the norm $\|\langle x\rangle f\|_{L^1}$ with $\langle x\rangle \equiv 1 + |x|$ controls the L^∞ norm of the derivative of the Fourier transform of f. We denote the mass of a function f by $\overline{f} = \int_{-\infty}^{\infty} f \, d\xi$. The quantity $[\vec{F}]_k$ will denote the kth component of the vector $\vec{F}$ when less cumbersome notation is not available.

2.2. Asymptotic pulse solutions. Within the semistrong pulse regime the two pulses interact strongly through the inhibitor component, U, and weakly through the activator, V. The asymptotic family of semistrong two-pulse solutions is parameterized by the pulse location $\Gamma \in \mathcal{K} = \{(\Gamma_1, \Gamma_2) \Big| \Gamma_1 < \Gamma_2, |\Gamma_1 - \Gamma_2| \geq \Delta\Gamma^*(\mu)\}$, where $\Delta\Gamma^*(\mu)$ is defined in Proposition 3.4. We denote the two-pulse solution by $\Phi_\Gamma(\xi)$ which we expand as

$$\Phi_\Gamma(\xi) = \begin{pmatrix} \Phi_{\Gamma,1} \\ \Phi_{\Gamma,2} \end{pmatrix} = \begin{pmatrix} U_0(\xi;\Gamma) + \varepsilon^2 U_2(\xi;\Gamma) + \varepsilon^4 U_4(\xi;\Gamma) \\ V_0(\xi;\Gamma) + \varepsilon^2 V_2(\xi;\Gamma) \end{pmatrix}, \tag{2.5}$$

and define the manifold $\mathcal{M} \subset H^1 \times H^1$ of two-pulse solutions by

$$\mathcal{M} = \{\Phi_\Gamma \big| \Gamma \in \mathcal{K}\}. \tag{2.6}$$

We first describe the leading order terms $(U_0, V_0)^t$ which were derived in [6]. Full resolution of the pulse dynamics of the renormalization procedure of section 4 requires a more accurate description of the two-pulse solution which requires the construction of the higher order corrections, outlined in Lemma 2.1.

The V-components of the two-pulse solutions are centered around the pulse-positions $\xi = \Gamma_k(t)$, where

$$\Gamma_1(t) = \Gamma_0 - \varepsilon^2 \int_0^t \hat{c}(s)ds, \;\; \Gamma_2(t) = \Gamma_0 + \varepsilon^2 \int_0^t \hat{c}(s)ds. \tag{2.7}$$

In the two-pulse configuration each pulse moves away from their mutual center Γ_0 with equal and opposite speed given by

$$\hat{c} = \frac{1}{2}\sqrt{\mu} \frac{e^{-\varepsilon^2 \Delta\Gamma\sqrt{\mu}}}{1 + e^{-\varepsilon^2 \Delta\Gamma\sqrt{\mu}}}, \tag{2.8}$$

where $\Delta\Gamma = \Delta\Gamma(t) = |\Gamma_1 - \Gamma_2|$; see (1.5).

The leader order term, V_0, of the V component of Φ_Γ is given by the sum of two one-pulses

$$V_0(\xi;\Gamma(t)) = \phi_1 + \phi_2, \tag{2.9}$$

where for $k = 1, 2$ the one-pulse solution is

$$\phi_k(\xi) = \frac{3}{2} A(\Gamma) \operatorname{sech}^2 \frac{1}{2}(\xi - \Gamma_k(t)). \tag{2.10}$$

A key distinction between the semistrong interaction depicted here and the weak pulse interaction is that the pulse amplitude, $A(\Gamma)$, depends nontrivially upon the pulse separation, $\Delta\Gamma = |\Gamma_1 - \Gamma_2|$, via

$$A(\Gamma) = \frac{\sqrt{\mu}}{3} \frac{1}{1 + e^{-2\varepsilon^2 \Delta\Gamma\sqrt{\mu}}}. \tag{2.11}$$

The pulse regions $\mathcal{I}_k = \mathcal{I}_k(t)$, $k = 1, 2$, are defined as regions outside which V_0 is exponentially small, and such that U_0 remains constant at leading order over a pulse region. We set the width of the pulse regions to be $\mathcal{O}(1/\sqrt{\varepsilon})$, i.e., we define

$$\mathcal{I}_k = \left(\Gamma_k(t) - \frac{1}{\sqrt{\varepsilon}}, \Gamma_k(t) + \frac{1}{\sqrt{\varepsilon}}\right), \;\; k = 1, 2. \tag{2.12}$$

The choice of pulse region width is somewhat arbitrary but standard. Another distinguishing feature of the semistrong pulse interaction is that the slowly varying U-component of Φ_Γ is not the sum of two one-pulses. To the left of $\mathcal{I}_1$ and to the right of $\mathcal{I}_2$, $U_0(x,t)$ decays slowly, while in the region between $\mathcal{I}_1$ and $\mathcal{I}_2$ it is cosh-like, but again on the slow, spatial scale,

(2.13)
$$U_0(\xi;\Gamma) = \begin{cases} Ae^{\varepsilon^2\sqrt{\mu}(\xi-\Gamma_1)} & \text{for } \xi < \Gamma_1 - \varepsilon^{-\frac{1}{2}}, \\ A\dfrac{\cosh \varepsilon^2\sqrt{\mu}\,(\xi - (\Gamma_1+\Gamma_2)/2)}{\cosh \varepsilon^2\sqrt{\mu}\Delta\Gamma/2} & \text{for } \Gamma_1 + \varepsilon^{-\frac{1}{2}} < \xi < \Gamma_2 - \varepsilon^{-\frac{1}{2}}, \\ Ae^{-\varepsilon^2\sqrt{\mu}(\xi-\Gamma_2)} & \text{for } \Gamma_2 + \varepsilon^{-\frac{1}{2}} < \xi. \end{cases}$$

As defined above, U_0 would be nonsmooth if extended into the pulse regions $\mathcal{I}_k$. Rather we define the U-component of the two-pulse solution inside $\mathcal{I}_k$ as $U_0+\varepsilon^2 U_2(\xi)$, where $U_0 \equiv A$, and $U_2(\xi)$ is a solution of $U_{\xi\xi} + \phi_k^2 = 0$; see (2.3). Using (2.13) as boundary or matching conditions and the pulse amplitude (2.11), we find

(2.14)
$$U_0 + \varepsilon^2 U_2(\xi;\Gamma) = A + \varepsilon^2 \begin{cases} A[\sqrt{\mu} - 3A](\xi - \Gamma_1) - \int_{\Gamma_1}^{\xi}\int_{\Gamma_1}^{\xi_1} \phi_1^2(\xi_2)\,d\xi_2\,d\xi_1 & \text{for } \xi \in \mathcal{I}_1, \\ A[\sqrt{\mu} - 3A](\xi - \Gamma_2) - \int_{\Gamma_2}^{\xi}\int_{\Gamma_2}^{\xi_1} \phi_2^2(\xi_2)\,d\xi_2\,d\xi_1 & \text{for } \xi \in \mathcal{I}_2, \end{cases}$$

which gives that $U_0 + \varepsilon^2 U_2(\xi) \in C^1 \cap H^2$. The C^1-smoothness of $U_0 + \varepsilon^2 U_2(\xi)$ is equivalent to the amplitude-pulse separation relation (2.11), i.e. $U_0 + \varepsilon^2 U_2(\xi)$ can be smooth only for $A(\Gamma)$ given by (2.11).

Relations (2.7), (2.8), (2.9), (2.10), (2.11), (2.13), and (2.14) give a leading order description of the two-pulse solution $\Phi_\Gamma(\xi)$. The corrections $U_4(\xi)$ and $V_2(\xi)$ can be obtained by a straightforward regular asymptotic expansion and are both defined only in the pulse regions $\mathcal{I}_{1,2}$ (see (2.23) in the proof of Lemma 2.1 below). The residual of Φ_Γ,

$$\mathbf{R} = \mathbf{F}(\Phi_\Gamma) = \begin{pmatrix} F_1(\Phi_\Gamma) \\ F_2(\Phi_\Gamma) \end{pmatrix}, \tag{2.15}$$

is determined by the right-hand side of (2.3), denoted by $(F_1, F_2)^t$, evaluated at Φ_Γ. Obtaining L^1- and L^2-estimates on the residual is a key step to controlling the remainder in the renormalization process.

LEMMA 2.1. *For the residual* $\mathbf{R} = \mathbf{F}(\Phi_\Gamma)$ *defined in* (2.15) *we have*

$$\sup_{\mathbb{R}} |F_2(\Phi_\Gamma)| = \mathcal{O}(\varepsilon^2), \quad \sup_{\mathbb{R}\setminus\mathcal{I}_1\cup\mathcal{I}_2} |F_1(\Phi_\Gamma)| = \mathcal{O}(\varepsilon^4), \quad \sup_{\mathcal{I}_1\cup\mathcal{I}_2} |F_1(\Phi_\Gamma)| = \mathcal{O}(\varepsilon\sqrt{\varepsilon}). \tag{2.16}$$

More specifically

$$R_2(\Gamma) = \varepsilon^2 \hat{c}\,(\phi_1' - \phi_2') + \mathcal{O}(\varepsilon^4) \text{ in } L^2(\mathbb{R}), \tag{2.17}$$

while

$$\|R_1(\Gamma)\|_{L^1} = \mathcal{O}(\varepsilon). \tag{2.18}$$

The $\mathcal{O}(\varepsilon\sqrt{\varepsilon})$ bound on F_1 in (2.16) and the $\mathcal{O}(\varepsilon)$ bound on R_1 in (2.18) deteriorate to $\mathcal{O}(\varepsilon^{-\frac{1}{2}})$ bounds, if we do not introduce the leading order corrections $\varepsilon^4 U_4$ and $\varepsilon^2 V_2$ in (2.5). Moreover, (2.17) no longer holds in that case. On the other hand, the bounds on $R_2(\Gamma)$ given in the lemma are sharp. The bounds on $R_1(\Gamma)$ may be sharpened, but this does not lead to any improvements in the renormalization analysis of section 4.

Proof. In [6], Φ_Γ is constructed as the solution of the classical Gierer–Meinhardt system

$$\begin{cases} -\varepsilon^6 \hat{c}_k U_\xi = U_{\xi\xi} - \varepsilon^4 \mu U + \varepsilon^2 V^2, \\ -\varepsilon^2 \hat{c}_k V_\xi = V_{\xi\xi} - V + \frac{V^2}{U} \end{cases} \tag{2.19}$$

with $\hat{c}_1 = -\hat{c} < 0$ for $\xi < \Gamma_0$ and $\hat{c}_2 = \hat{c} > 0$ for $\xi < \Gamma_0$, and $\hat{c} = \hat{c}(t)$ as in (2.8). Note the factor ε^4 difference between the right-hand sides of the U-equation here and in (2.3). We may employ a regular perturbation expansion, writing

$$\begin{aligned} U(\xi,t) &= U_0(\xi;\Gamma) + \varepsilon^2 U_2(\xi;\Gamma) + \varepsilon^4 U_4(\xi;\Gamma) + \varepsilon^6 U_r(\xi,t;\varepsilon^2), \\ V(\xi,t) &= V_0(\xi;\Gamma) + \varepsilon^2 V_2(\xi;\Gamma) + \varepsilon^4 V_r(\xi,t;\varepsilon^2), \end{aligned} \tag{2.20}$$

where A, U_2, and V_0 are given in (2.9), (2.10), (2.11), (2.13), (2.14), and U_4 and V_2 have already been introduced in (2.5). Likewise, we expand F_1 and F_2 (Proposition 2.1),

$$\begin{aligned} F_1(\Phi_\Gamma) &= \frac{1}{\varepsilon^2}\left[U_{2,\xi\xi} + V_0^2\right] + \left[U_{4,\xi\xi} - \mu U_0 + 2V_0V_2\right] \\ &\quad + \varepsilon^2\left[U_{r,\xi\xi} - \varepsilon^4\mu U_r - F_{1,r}^{\mathrm{inh}}(U_2, V_{0,2,r};\varepsilon^2)\right], \\ F_2(\Phi_\Gamma) &= \left[V_{0,\xi\xi} - V_0 + \frac{V_0^2}{U_0}\right] + \varepsilon^2\left[L_{22}V_2 - \frac{V_0^2U_2}{U_0^2}\right] \\ &\quad + \varepsilon^4\left[L_{22}V_r - F_{2,r}^{\mathrm{inh}}(U_{0,2,4}, \varepsilon^2 U_r, V_{0,2}, \varepsilon^2 V_r;\varepsilon^2)\right], \end{aligned} \tag{2.21}$$

where

$$L_{22}V = V_{\xi\xi} - V + \frac{V_0^2}{U_0}V, \tag{2.22}$$

and the expressions for $F_{1,r}^{\mathrm{inh}}(U_2, V_{0,2,r};\varepsilon^2)$ and $F_{2,r}^{\mathrm{inh}}(U_{0,2,4}, \varepsilon^2 U_r, V_{0,2}, \varepsilon^2 V_r;\varepsilon^2)$ follow directly by substitution of (2.20) in (2.3). We obtain by (2.19) the following equations for U_4 and V_2:

$$U_{4,\xi\xi} = \mu U_0 - 2V_0V_2, \quad L_{22}V_2 = \frac{V_0^2U_2}{U_0^2} - \hat{c}_k V_{0,\xi}, \tag{2.23}$$

for $\xi \in \mathcal{I}_{1,2}$. These equations can be solved uniquely by application of the natural boundary/matching conditions. Note that $U_4(\xi)$ grows as $(\xi - \Gamma_{1,2})^2$ for $|\xi - \Gamma_{1,2}| \gg 1$ and that V_2 decays exponentially to 0 as $|\xi - \Gamma_{1,2}| \gg 1$ (see (2.9)). The equations for the remainders $U_r(\xi,t;\varepsilon^2)$ and $V_r(\xi,t;\varepsilon^2)$ are given by

$$\begin{aligned} U_{r,\xi\xi} - \varepsilon^4\mu U_r &= F_{1,r}^{\mathrm{inh}}(U_2, V_{0,2,r};\varepsilon^2) - \hat{c}_k U_{0,\xi} - \varepsilon^2\hat{c}_k U_{2,\xi} - \varepsilon^4\hat{c}_k U_{4,\xi} - \varepsilon^6\hat{c}_k U_{r,\xi}, \\ L_{22}V_r &= F_{2,r}^{\mathrm{inh}}(U_{0,2,4}, \varepsilon^2 U_r, V_{0,2}, \varepsilon^2 V_r;\varepsilon^2) - \hat{c}_k V_{2,\xi} - \varepsilon^2\hat{c}_k V_{r,\xi}, \end{aligned} \tag{2.24}$$

for $\xi \in \mathbb{R}$. It is a straightforward procedure to check that $|V_r|$ and $|F^{\text{inh}}_{2,r}|$ are uniformly bounded for $\xi \in \mathbb{R}$; in fact, both V_r and $F^{\text{inh}}_{2,r}$ decay exponentially to 0 as $|\xi - \Gamma_{1,2}(t)| \gg 1$. Together with the definitions of V_0 and V_2 ((2.9) and (2.23)), substitution of this result in the second equation of (2.21) yields (2.17). This also implies the results on $F_2(\Phi_\Gamma)$ in (2.16).

Outside the pulse regions $\mathcal{I}_k$, all $V_{0,2,r}$ components are exponentially small, and U_0 is constructed as a solution of the equation $U_{\xi\xi} - \varepsilon^4 \mu U = 0$; see (2.13). Therefore, the correction U_r to U_0 in the U-component of the two-pulse solution also varies like $\varepsilon^2 \xi$, and U_2 and U_4 may be taken to be identically zero outside $\mathcal{I}_k$. This implies by (2.21) and (2.24) that outside $\mathcal{I}_k$

$$F_1(\Phi_\Gamma) = -\varepsilon^2 \hat{c}_k U_{0,\xi} = \mathcal{O}(\varepsilon^4).$$

Since U_r decays for $\xi \to \pm\infty$ with the same slow rate as U_0, we find

$$\int_{\mathbb{R}\setminus\mathcal{I}_1\cup\mathcal{I}_2} |F_1(\Phi_\Gamma)|\, d\xi = \frac{1}{\varepsilon^2} \times \mathcal{O}(\varepsilon^4) = \mathcal{O}(\varepsilon^2).$$

Inside $\mathcal{I}_k$, we conclude from (2.24) and the fact that U_2 grows linearly with $(\xi - \Gamma_i)$ (see (2.14)) that U_r may grow as $(\xi - \Gamma_i)^3$. Nevertheless, both $U_{r,\xi\xi}$ and $F^{\text{inh}}_{1,r}(U_2, V_{0,2,r}; \varepsilon^2)$ only grow linearly in $(\xi - \Gamma_i)$. Since the width of the $\mathcal{I}_k$ interval is $\mathcal{O}(1/\sqrt{\varepsilon})$ (see (2.12)) we deduce from (2.21) and (2.24) that $\sup_{\mathcal{I}_1\cup\mathcal{I}_2} |F_1(\Phi_\Gamma)| = \mathcal{O}(\varepsilon\sqrt{\varepsilon})$ (2.16). Hence, by (2.12),

$$\int_{\mathcal{I}_k} |F_1(\Phi_\Gamma)|\, d\xi = \frac{1}{\sqrt{\varepsilon}} \times \mathcal{O}(\varepsilon\sqrt{\varepsilon}) = \mathcal{O}(\varepsilon),$$

which yields the L^1-bound (2.18). ∎

3. Linearization and the reduced operators. In a neighborhood of the two-pulse manifold $\mathcal{M}$ we decompose the solutions of (2.3) as

$$\begin{pmatrix} U \\ V \end{pmatrix} = \Phi_\Gamma + W(\xi, t), \tag{3.1}$$

where the remainder $W = (W_1, W_2)^t$ and Γ is taken as a function of time. In terms of the remainder introduced in (3.1), the GM equation (2.3) can then be written as

$$W_t + \frac{\partial \Phi}{\partial \Gamma}\dot{\Gamma} = \mathbf{R} + L_\Gamma W + \mathcal{N}(W), \tag{3.2}$$

where R is the residual (2.15) and L_Γ is the linearization of F about Φ_Γ, given by

$$L_\Gamma = \begin{pmatrix} \epsilon^{-4}\partial_\xi^2 - \mu & 2\varepsilon^{-2}\Phi_{\Gamma,2} \\ -\dfrac{\Phi_{\Gamma,2}^2 \kappa'(\Phi_{\Gamma,1})}{\kappa(\Phi_{\Gamma,1})^2} & \partial_\xi^2 - 1 + 2\dfrac{\Phi_{\Gamma,2}}{\kappa(\Phi_{\Gamma,1})} \end{pmatrix}. \tag{3.3}$$

In the linear operator above $\kappa(\Phi_{\Gamma,1}) = \Phi_{\Gamma,1}$ except for those ξ for which $\Phi_{\Gamma,2}(\xi) = \mathcal{O}(e^{-\varepsilon^{-2}|\ln\delta|})$. Thus the perturbation to the linearization introduced by the regularization is compact and exponentially small. The final term, $\mathcal{N}(W)$, representing the nonlinearity is given at leading order by

$$\mathcal{N}(W) = \begin{pmatrix} \epsilon^{-2} W_2^2 \\ \mathcal{O}(W_2^2) + \mathcal{O}(V_0 W_1 W_2) + \mathcal{O}(V_0^2 W_1^2) \end{pmatrix}. \tag{3.4}$$

From the asymptotic form of the pulse solution given in (2.9), (2.11), and (2.13), we calculate that

$$\varepsilon^2 \left\| \frac{\partial U_0}{\partial \Gamma_k} \right\|_{L^1} + \varepsilon \left\| \frac{\partial U_0}{\partial \Gamma_k} \right\|_{L^2} + \left\| \frac{\partial U_0}{\partial \Gamma_k} \right\|_{L^\infty} = \mathcal{O}(\varepsilon^2), \tag{3.5}$$

while

$$\frac{\partial V_0}{\partial \Gamma_k} = -\phi_k' + \mathcal{O}(\varepsilon^2), \tag{3.6}$$

in L^2.

3.1. The reduced linearization. A key step in the RG treatment is the replacement of the exact linear operator with a reduced operator whose spectral and semigroup properties are easier to analyze, yet such that the difference between the exact and the reduced operator, the secularity, does not lead to growth of the remainder W. Due to the contractivity of the L_{11} component of L_Γ, the two-pulse potential which comprises the L_{12} component can be replaced with δ functions located at each pulse position. The mass of the delta function is chosen to equal the mass of the product of the original potential and the function it operates upon. We also replace the exact two-pulse solution Φ_Γ with its leading order approximation $(U_0, V_0)^t$. With these reductions the linearized operator becomes

$$\tilde{L}_\Gamma = \begin{pmatrix} \epsilon^{-4}\partial_\xi^2 - \mu & 2\varepsilon^{-2}\big(\delta_{\Gamma_1} \otimes \phi_1 + \delta_{\Gamma_2} \otimes \phi_2\big) \\ -\dfrac{V_0^2}{A^2} & \partial_\xi^2 - 1 + 2\dfrac{V_0}{A} \end{pmatrix}, \tag{3.7}$$

where the tensor product of f_1 and f_2 is defined by

$$(f_1 \otimes f_2)\, W = (f_2, W)_{L^2} f_1. \tag{3.8}$$

In particular, $\delta_{\Gamma_k} \otimes \phi_k$ represents the tensor product of the δ function centered at $\xi = \Gamma_k$ with ϕ_k. In the analysis below we use the notation

$$\alpha_k(W) = (\phi_k, W)_{L^2}, \tag{3.9}$$

for $k = 1, 2$. The scalar operators that appear in the upper left entry, respectively lower right, of the matrix $\tilde{L}$ (3.7) will be denoted by L_{11}, respectively L_{22}; see (2.22). The reduced operator is ostensibly an $\mathcal{O}(\varepsilon^{-2})$ perturbation of the original operator. However, it is immediately clear that they share the same essential spectrum

$$\sigma_{\text{ess}} = \{\lambda \in \mathbb{R} : \lambda \leq \max(-1, -\mu)\}. \tag{3.10}$$

3.2. The point spectrum. The two-pulse profiles which comprise the manifold $\mathcal{M}$ are not stationary solutions, and as such it is not self-consistent to determine their linear stability in terms of the spectrum of the associated linearized operator. We say that the two-pulse solution Φ_Γ is *spectrally compatible* with the manifold $\mathcal{M}$ if the spectrum of the associated linear operator can be decomposed into a part contained within the left-half complex plane and a finite-dimensional part whose associated eigenspace approximates the tangent plane of $\mathcal{M}$ at Γ.

To determine the point spectrum of $\tilde{L}$ we invert the U component of the eigenvalue equation, and eliminate the inhibitor from the activator equation, reducing the

eigenvalue problem to a scalar equation for the activator component of the eigenfunction. We call this the NLEP equation (see (3.23)) and denote the corresponding linear operator by $\mathcal{L}(\lambda, \Delta\Gamma)$. The NLEP operator controls the point spectrum of $\tilde{L}$ to leading order.

PROPOSITION 3.1. *Up to multiplicity we have* $\sigma_p(\tilde{L}) = \{\lambda | \mathrm{Ker}(\mathcal{L}(\lambda)) \neq 0\}(1 + \mathcal{O}(\varepsilon^2))$. *That is, for each eigenvalue* $\lambda \in \sigma_p(\tilde{L})$ *with corresponding eigenvector* $\Psi = (\Psi_1, \Psi_2)^t$, *there is a* $\lambda_{\mathcal{L}}$ *and corresponding* ψ *such that* $\mathcal{L}(\lambda_{\mathcal{L}})\psi = \lambda_{\mathcal{L}}\psi$, $|\lambda - \lambda_{\mathcal{L}}| = \mathcal{O}(\varepsilon^2)$ *with* $\Psi_2 = \psi(1 + \mathcal{O}(\varepsilon^2))$ *and* Ψ_1 *given by* (3.12) *up to* $\mathcal{O}(\varepsilon^2)$. *Moreover, the small eigenvalues of* $\tilde{L}$ *and* $\mathcal{L}$ *are both exponentially small.*

Proof. The eigenvalue problem for the reduced operator is written as

$$\tilde{L}\Psi = \lambda\Psi, \tag{3.11}$$

where $\Psi = (\Psi_1, \Psi_2)^t$ is a possibly complex two-vector. Since $L_{11} - \lambda$ is invertible for $\lambda \notin (-\infty, -\mu]$, we may solve for Ψ_1 as

$$\Psi_1 = -2\varepsilon^{-2}\big(\alpha_1(L_{11} - \lambda)^{-1}\delta_{\Gamma_1} + \alpha_2(L_{11} - \lambda)^{-1}\delta_{\Gamma_2}\big), \tag{3.12}$$

where the $\alpha_k = (\phi_k, \Psi_2)_{L^2}$ are as yet undetermined. From the Fourier transform we find

$$\widehat{\Psi}_1(k) = \frac{2}{\sqrt{2\pi}}\frac{\varepsilon^2\big(\alpha_1 e^{ik\Gamma_1} + \alpha_2 e^{ik\Gamma_2}\big)}{k^2 + \varepsilon^4(\mu + \lambda)}. \tag{3.13}$$

From the integral relation

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-ik\xi}\frac{\varepsilon^2 e^{ik\Gamma}}{k^2 + \varepsilon^4(\mu + \lambda)} = \sqrt{\frac{\pi}{2(\mu + \lambda)}}e^{-\varepsilon^2\sqrt{\mu+\lambda}|\xi - \Gamma|}, \tag{3.14}$$

we may invert the Fourier transform of Ψ_1 explicitly,

$$\Psi_1(\xi, t) = \alpha_1 H(\lambda, \xi - \Gamma_1) + \alpha_2 H(\lambda, \xi - \Gamma_2), \tag{3.15}$$

where

$$H(\lambda, x) = \frac{1}{\sqrt{\mu + \lambda}}e^{-\varepsilon^2|x|\sqrt{\mu+\lambda}}. \tag{3.16}$$

Eliminating Ψ_1, the equation for Ψ_2 reduces to

$$(L_{22} - \lambda)\Psi_2 = \frac{V_0^2}{A^2}\Psi_1; \tag{3.17}$$

see also (2.22). Since Ψ_1 is a slowly varying function of ξ, while each term in V_0 decays exponentially to zero at an $\mathcal{O}(1)$ rate in ξ, we may reduce the equation for Ψ_2 to

$$(L_{22} - \lambda)\Psi_2 = \frac{V_0^2}{A^2}\left(\alpha_1 H(\lambda, \xi - \Gamma_1) + \alpha_2 H(\lambda, \xi - \Gamma_2)\right) \tag{3.18}$$

$$= \frac{\phi_1^2 + \phi_2^2}{A^2}\left(\alpha_1 H(\lambda, \xi - \Gamma_1) + \alpha_2 H(\lambda, \xi - \Gamma_2)\right) + \mathcal{O}(e^{-\Delta\Gamma}) \tag{3.19}$$

$$= \frac{1}{A^2\sqrt{\mu + \lambda}}\left[\phi_1^2\left(\alpha_1 + \alpha_2 E\right) + \phi_2^2\left(\alpha_1 E + \alpha_2\right)\right] + \mathcal{O}(\varepsilon^2), \tag{3.20}$$

where

$$E = E(\Delta\Gamma;\lambda) = e^{-\varepsilon^2\sqrt{\mu+\lambda}\Delta\Gamma}. \tag{3.21}$$

In the tensor product notation this is written as

$$(L_{22}-\lambda)\Psi_2 = \frac{1}{A^2\sqrt{\mu+\lambda}}\left[\phi_1^2\otimes(\phi_1+E\phi_2)+\phi_2^2\otimes(E\phi_1+\phi_2)\right]\Psi_2. \tag{3.22}$$

We define the NLEP operator as

$$\mathcal{L}(\lambda,\Delta\Gamma) = L_{22} - \frac{1}{A^2\sqrt{\mu+\lambda}}\left[\phi_1^2\otimes(\phi_1+E\phi_2)+\phi_2^2\otimes(E\phi_1+\phi_2)\right]. \tag{3.23}$$

This is a compact perturbation of L_{22} and thus is Fredholm, with the same essential spectrum, but is no longer self-adjoint. Indeed its adjoint exchanges the roles of the potentials in each tensor product. ∎

PROPOSITION 3.2. *Except for the exponentially small eigenvalues, the point spectrum of the NLEP operator $\mathcal{L}$ is given, up to multiplicity, by the zeros of the equation*

$$\mathcal{R}(\lambda) - 3\frac{\sqrt{\mu+\lambda}}{\sqrt{\mu}}\frac{1+e^{-\varepsilon^2\sqrt{\mu}\Delta\Gamma}}{1\pm e^{-\varepsilon^2\sqrt{\mu+\lambda}\Delta\Gamma}} = 0, \tag{3.24}$$

where $\mathcal{R}$ is an explicitly known meromorphic function on $\mathbf{C}\backslash(-\infty,-1]$ given by (3.33).

Proof. The spectrum of the NLEP operator $\mathcal{L}$ can be determined explicitly as the zeros of an analytic equation using the methods developed in [5], which we outline below. We introduce $w_h(\xi)\geq 0$ as the scaled homoclinic solution of

$$w_{\xi\xi} - w + w^2 = 0 \tag{3.25}$$

with its maximum at $\xi = 0$. For $k = 1,2$ we introduce the translates $w_{h,k}(\xi) = w_h(\xi-\Gamma_k)$. Since $\phi_k(\xi) = Aw_{h,k}(\xi)$, (2.10) and (3.22) can be written as

$$\frac{d^2\Psi_2}{d\xi^2} - \left[(1+\lambda) - 2(w_{h,1}+w_{h,2})\right]\Psi_2 = \frac{1}{\sqrt{\mu+\lambda}}\left[w_{h,1}^2(\alpha_1+\alpha_2E)+w_{h,2}^2(\alpha_1E+\alpha_2)\right], \tag{3.26}$$

where $\alpha_k = \alpha_k(\Psi_2)$ (3.9). Since the potential of both the Schrödinger operator on the left-hand side of the equation and the inhomogeneous term on the right-hand side consists of disjoint parts localized about Γ_1 and Γ_2, it is natural to decompose Ψ_2 into

$$\Psi_2 = \psi_1(\xi) + \psi_2(\xi), \tag{3.27}$$

where ψ_k is localized about Γ_k and decays exponentially as ξ moves away from Γ_k. Equation (3.26) is equivalent, up to exponentially small terms, to the coupled system

$$\begin{cases} \dfrac{d^2\psi_1}{d\xi^2} - \left[(1+\lambda) - 2w_{h,1}\right]\psi_1 = \dfrac{w_{h,1}^2}{\sqrt{\mu+\lambda}}(\alpha_1+\alpha_2E), \\ \dfrac{d^2\psi_2}{d\xi^2} - \left[(1+\lambda) - 2w_{h,2}\right]\psi_2 = \dfrac{w_{h,2}^2}{\sqrt{\mu+\lambda}}(\alpha_1E+\alpha_2). \end{cases} \tag{3.28}$$

We define $\bar{\psi} = \bar{\psi}(\xi;\lambda)$ as the uniquely determined bounded solution of

$$\frac{d^2\psi}{d\xi^2} - [(1+\lambda) - 2w_h]\,\psi = w_h^2, \tag{3.29}$$

and its translates $\bar{\psi}_k(\xi)$ are defined by $\bar{\psi}_k(\xi) = \bar{\psi}(\xi - \Gamma_k)$. The functions $\bar{\psi}$ can be determined explicitly; see [5].

We first consider the solution of (3.29) for $\lambda \notin \sigma_{\rm red} = \{\frac{5}{4}, 0, -\frac{3}{4}\} \cup (-\infty, -1]$, the spectrum of the operator

$$L_{\rm red} = \frac{d^2}{d\xi^2} - (1 - 2w_h(\xi)). \tag{3.30}$$

Clearly,

$$\psi_1(\xi) = C_1\bar{\psi}_1(\xi), \;\; \psi_2(\xi) = C_2\bar{\psi}_2(\xi) \tag{3.31}$$

for some constants C_k that depend on λ and $\Delta\Gamma$. Recalling that here $\alpha_k = (\phi_k, \Psi_2)_{L^2}$ and using (3.27), we find

$$\begin{aligned} \alpha_k &= \int_{-\infty}^{\infty} \phi_k(\psi_1 + \psi_2)d\xi = \int_{-\infty}^{\infty} Aw_{h,k}(C_1\bar{\psi}_1 + C_2\bar{\psi}_2)d\xi \\ &= AC_k \int_{-\infty}^{\infty} w_{h,k}\bar{\psi}_k d\xi = AC_k \int_{-\infty}^{\infty} w_h\bar{\psi}d\xi \end{aligned} \tag{3.32}$$

up to asymptotically small corrections. The quantity

$$\mathcal{R}(\lambda) \equiv \int_{-\infty}^{\infty} w_h\bar{\psi}d\xi \tag{3.33}$$

is meromorphic for $\lambda \in \mathbb{C}\backslash(-\infty, -1]$, with poles at $\lambda = \frac{5}{4}$ and $\lambda = -\frac{3}{4}$; see [5]. Note that in [5] a more general function, $\mathcal{R}(\lambda;\beta_1,\beta_2)$, has been defined and studied; (3.33) is related to [5] by $\mathcal{R}(\lambda) = 216\mathcal{R}(\lambda;2,2)$. The system (3.28) can be written as

$$\begin{cases} \dfrac{d^2\psi_1}{d\xi^2} - [(1+\lambda) - 2w_{h,1}]\,\psi_1 = \dfrac{Aw_{h,1}^2\mathcal{R}}{\sqrt{\mu+\lambda}}\,(C_1 + C_2E)\,, \\ \dfrac{d^2\psi_2}{d\xi^2} - [(1+\lambda) - 2w_{h,2}]\,\psi_2 = \dfrac{Aw_{h,2}^2\mathcal{R}}{\sqrt{\mu+\lambda}}\,(C_1E + C_2)\,. \end{cases} \tag{3.34}$$

Comparing the equations for $\psi_{1,2}(\xi)$ to (3.29), we obtain the following relations for C_1 and C_2:

$$C_1 = \frac{A\mathcal{R}}{\sqrt{\mu+\lambda}}\,(C_1 + C_2E)\,, \;\; C_2 = \frac{A\mathcal{R}}{\sqrt{\mu+\lambda}}\,(C_1E + C_2)\,, \tag{3.35}$$

or, equivalently,

$$\begin{pmatrix} \dfrac{A\mathcal{R}}{\sqrt{\mu+\lambda}} - 1 & \dfrac{AE\mathcal{R}}{\sqrt{\mu+\lambda}} \\ \dfrac{AE\mathcal{R}}{\sqrt{\mu+\lambda}} & \dfrac{A\mathcal{R}}{\sqrt{\mu+\lambda}} - 1 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3.36}$$

For (3.26) to have nontrivial solutions the determinant of the matrix on the left-hand side of (3.36) must be zero. Isolating $\mathcal{R}(\lambda)$ from the resulting expression and using

(2.11) and (3.21), we obtain (3.24) whose zeros are the eigenvalues of the NLEP equation (3.22), up to multiplicity, outside of $\sigma_{\rm red}$. These eigenvalues lie on curves $\lambda(\Delta\Gamma)$, parametrized by the pulse separation.

For $\lambda \in \sigma_{\rm red}$ we analyze the eigenvalue equation case by case. For $\lambda = -\frac{3}{4}$ or $\frac{5}{4}$, (3.29) does not have a bounded solution since the right-hand side is not orthogonal to the kernel of the $L_{\rm red}$, and so these values cannot be eigenvalues. However, the eigenfunction $\frac{d}{d\xi}w_h$ of $L_{\rm red}$ at $\lambda = 0$ is L^2 orthogonal to w_h. Equation (3.32) implies that $\alpha_1 = \alpha_2 = 0$ and the system (3.28) has a double eigenvalue at $\lambda = 0$ with a two-dimensional eigenspace spanned by $\{\frac{d}{d\xi}w_h(\xi - \Gamma_1), \frac{d}{d\xi}w_h(\xi - \Gamma_2)\}$. These eigenvalues do not occur as solutions of (3.24), rather they correspond to exponentially small eigenvalues of the original NLEP system (3.22), whose corresponding eigenfunctions, derived in Lemma 3.7, form the key spectral projection onto the active tangent plane. ∎

Remark 3.3. Proposition 3.2 is equivalent to Principle Result 5.3 of [16].

We identify conditions on μ and Γ such that the reduced linearized operator, $\tilde{L}_\Gamma$, is spectrally compatible with the manifold $\mathcal{M}$ of two-pulse solutions.

PROPOSITION 3.4. *For each* $\Delta\Gamma \in (0, \infty)$, *the* NLEP *eigenvalue problem* (3.22) *has two exponentially small eigenvalues, denoted* $\lambda_\pm$, *and* 4 *or* 6 *eigenvalues* $\lambda_+^{j_+}(\Delta\Gamma)$ *and* $\lambda_-^{j_-}(\Delta\Gamma)$, $j_\pm = 1, \ldots, J_\pm, J_\pm = J_\pm(\mu) = 2$ or 3. *There exists a unique* $\mu_{\rm Hopf} > 0$ *such that for all* $\mu > \mu_{\rm Hopf}$, *there is a* $\Delta\Gamma^*(\mu)$ *and a* $\nu > 0$ *such that*

$$\mathrm{Re}[\lambda_\pm^{j_\pm}(\Delta\Gamma)] < -\nu < 0 \ \text{ for all } \ \Delta\Gamma \geq \Delta\Gamma^*(\mu),\ j_\pm = 1, \ldots, J_\pm.$$

For all $\mu > \mu_{\rm TP} \approx 0.62$ (*the tangent point*), $\Delta\Gamma^*$ *takes the exact form*

$$\Delta\Gamma^*(\mu) = \frac{1}{\varepsilon^2\sqrt{\mu}}\log 3, \tag{3.37}$$

while for $\mu \in (\mu_{\rm Hopf}, \mu_{\rm TP})$, $\Delta\Gamma^*(\mu)$ *increases with decreasing* μ, *with* $\Delta\Gamma^*(\mu) \to \infty$ *as* $\mu \downarrow \mu_{\rm Hopf}$.

Since the two pulses of $\Phi_\Gamma(\xi)$ move away from each other (see (2.7), (2.8)), this result implies that the spectrum of the NLEP operator $\mathcal{L}$ remains in the stable half-plane for all $t \geq 0$ if $\Delta\Gamma(0) > \Delta\Gamma^*$.

Proof. We can distinguish two limits, $\Delta\Gamma \to \infty$ and $\Delta\Gamma \downarrow 0$. The first case represents the situation in which the two pulses of $\Phi_\Gamma(\xi)$ are so far apart that the two-pulse solution can be considered as two one-pulse solutions, i.e., the two-pulse solution is in the weak interaction limit. In this limit, (3.24) reduces to

$$\mathcal{R}(\lambda) = 3\frac{\sqrt{\mu + \lambda}}{\sqrt{\mu}}, \tag{3.38}$$

for both $\lambda_\pm(\Delta\Gamma)$. This is the relation that determines the point spectrum of the solitary one-pulse solution of (2.1), independent of the regularization. It was shown in Theorem 5.11 of [5] that there exists a unique $\mu_{\rm Hopf} > 0$ such that all solutions of (3.38) have $\mathrm{Re}(\lambda) < 0$ for $\mu > \mu_{\rm Hopf}$ and that (3.22) always has incompatible eigenvalues if $\mu < \mu_{\rm Hopf}$. Numerical evaluation shows that $\mu_{\rm Hopf} \approx 0.36$. Moreover, (3.38) has 2 or 3 nontrivial eigenvalues, i.e., $\lambda \neq 0$, depending on μ; the third (compatible) eigenvalue is created in an edge bifurcation as μ increases through $\mu_{\rm edge} \approx 0.77$ [5]. There also are 2 or 3 curves $\lambda_+^{j_+}(\Delta\Gamma)$ and $\lambda_-^{j_-}(\Delta\Gamma)$, i.e., $j_\pm = 1, \ldots, J_\pm$, $J_\pm(\mu) = 2$, respectively 3, for $\mu < \mu_{\rm edge}$, respectively $> \mu_{\rm edge}$. The eigenvalues $\lambda_\pm^3(\Delta\Gamma)$ are real and $\lambda_\pm^3(\Delta\Gamma) < -\frac{3}{4}$.

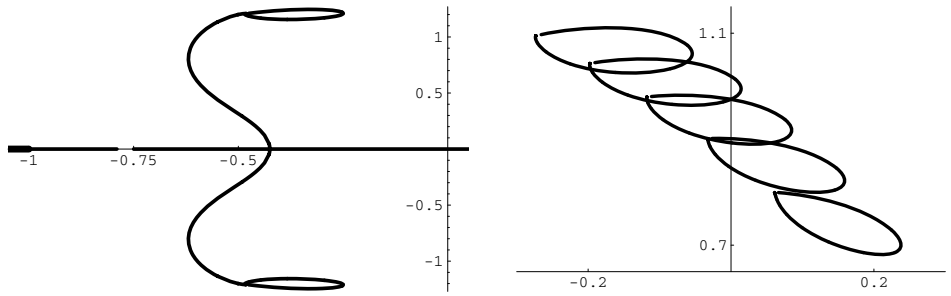


FIG. 3.1. (a) *The orbits of the zeroes* $\lambda(\Delta\Gamma)$ *of* (3.24) *plotted parametrically in the complex plane as a function of* $\Delta\Gamma$ *for* $\mu = 1$. *The eigenvalues* $\lambda_+^{1,2}(\Delta\Gamma)$ *are closed loops attached to the homoclinic limit* $\lambda \approx -0.48 + 1.20i$; *the curves* $\lambda_-^{1,2}(\Delta\Gamma)$ *approach the homoclinic point in the limit* $\Delta\Gamma \to \infty$, *but* $\lambda_-^{1,2}(\Delta\Gamma)$ *collide on the real axis, becoming real as* $\Delta\Gamma$ *decreases approaching the limits* $\lambda_-^{1,2}(\Delta\Gamma) \to -\frac{3}{4}, \frac{5}{4}$ *as* $\Delta\Gamma \to 0$. *The third pair of eigenvalues satisfies* $\lambda_\pm^3(\Delta\Gamma) < -\frac{3}{4}$ *with* $\lambda_-^3(\Delta\Gamma)$ *disappearing into the essential spectrum of* (3.22) *as* $\Delta\Gamma$ *decreases through a critical value. All eigenvalues* $\lambda_\pm^{1,2,j}(\Delta\Gamma)$ *have negative real part for* $\Delta\Gamma > \Delta\Gamma^*(1)$ *given by* (3.37). (b) *The closed* $\lambda_+^1(\Delta\Gamma)$*-loops for five values of* μ: $\mu = 0.7 > \mu_{\rm TP}$, $\mu = 0.6, 0.5, 0.4 \in (\mu_{\rm Hopf}, \mu_{\rm TP})$, *and* $\mu = 0.3 < \mu_{\rm Hopf}$. *The* $\Delta\Gamma$*-region* $(\Delta\Gamma_+^{1,*}(\mu), \Delta\Gamma_+^{2,*}(\mu))$ *in which* $Re[\lambda_+^1(\Delta\Gamma)] > 0$ *grows as* $\mu \in (\mu_{\rm Hopf}, \mu_{\rm TP})$ *decreases, so that* $\Delta\Gamma^*(\mu) = \Delta\Gamma_+^{2,*}(\mu)$ *for* $\mu < \mu^* \in (\mu_{\rm Hopf}, \mu_{\rm TP})$.

For small values of $\Delta\Gamma$ there are two mechanisms to generate incompatible point spectrum, one which occurs for $\mu > \mu_{\rm TP}$ and the other for $\mu \in (\mu_{\rm Hopf}, \mu_{\rm TP})$. The first occurs when the eigenvalues λ_-^1 and λ_-^2 collide and become real. Indeed, in the limit $\Delta\Gamma \downarrow 0$, it follows from Proposition 3.2 that $\lambda_+^{1,2}(\Delta\Gamma)$ again approaches a solution of (3.38), i.e., the $\lambda_+^{1,2}(\Delta\Gamma)$-branches are closed curves. On the other hand, $|\mathcal{R}(\lambda_-^{1,2}(\Delta\Gamma))|$ becomes unbounded in this limit. By evaluation of (3.24), we see that $\mathcal{R}(\lambda)$ becomes unbounded as $\lambda_-^1(\Delta\Gamma) \to -\frac{3}{4}$, the stable pole of $\mathcal{R}(\lambda)$, and as $\lambda_-^2(\Delta\Gamma) \to +\frac{5}{4}$, the other, unstable pole of $\mathcal{R}(\lambda)$. The passage of the real eigenvalue $\lambda_-^2(\Delta\Gamma)$ through zero corresponds to $\Delta\Gamma$ given by (3.37) since $\mathcal{R}(0) = 6$ [5]. In particular, the eigenvalue problem (3.22) has incompatible eigenvalues for all $\Delta\Gamma < \log 3/(\varepsilon^2\sqrt{\mu})$, for $\mu > \mu_{\rm TP}$.

In the second case, the $\lambda_+^{1,2}(\Delta\Gamma)$-branches may cross through the imaginary axis. For the tangent point value, $\mu = \mu_{\rm TP}$, the $\lambda_+^{1,2}$ curves are tangent to the imaginary axis. For $\mu \in (\mu_{\rm Hopf}, \mu_{\rm TP})$, a part of the closed, complex conjugate $\lambda_+^{1,2}(\Delta\Gamma)$-curves lies in the unstable half-plane, while the endpoints of the curve, i.e., the eigenvalues associated to the stationary homoclinic one-pulse limit, lie in the stable half-plane; see Figure 3.1(b). More specifically, $\mathrm{Re}[\lambda_+^{1,2}(\Delta\Gamma)] > 0$ for $\Delta\Gamma \in (\Delta\Gamma_+^{1,*}(\mu), \Delta\Gamma_+^{2,*}(\mu))$, where

$$\lim_{\mu\to\mu_{\rm Hopf}} \Delta\Gamma_+^{1,*}(\mu) = 0, \quad \lim_{\mu\to\mu_{\rm Hopf}} \Delta\Gamma_+^{2,*}(\mu) = \infty,$$
$$\lim_{\mu\to\mu_{\rm TP}} \Delta\Gamma_+^{1,*}(\mu) = \lim_{\mu\to\mu_{\rm TP}} \Delta\Gamma_+^{2,*}(\mu) \approx \frac{1.32}{\varepsilon^2},$$

so that $\Delta\Gamma^*(\mu) = \Delta\Gamma_+^{2,*}(\mu) > \log 3/(\varepsilon^2\sqrt{\mu})$ for $\mu \in (\mu_{\rm Hopf}, \mu_+^*)$ for a certain $\mu_+^* \in (\mu_{\rm Hopf}, \mu_{\rm TP})$. □

The orbits of the eigenvalues λ of (3.22) as function of $\Delta\Gamma$ can be determined by a direct evaluation of $\mathcal{R}(\lambda)$ [5]; see Figure 3.1.

Remark 3.5. Competition instabilities and synchronous oscillatory instabilities were identified for the Gierer–Meinhardt equations in [17, 16]; see, especially, section

5.2 of [16]. The presence of these two instabilities is related to the two multipliers in the NLEP, as also found here.

Remark 3.6. Proposition 3.4 implies that $\Phi_\Gamma(\xi)$ is not spectrally compatible with the manifold $\mathcal{M}$ if $\Delta\Gamma(0) < \Delta\Gamma^*(\mu)$. However, this lower bound on the admissible pulse separation distance does not limit the semistrong character of the pulse interaction in $\Phi_\Gamma(\xi)$, since the U-component of $\Phi_\Gamma(\xi)$ evolves on the slow $\varepsilon^2\xi$ space scale. To quantify the lower bound on pulse separation we determine the corresponding maximum value of the minimum $U_{\min}(t)$ of the inhibitor U between the two pulses $\Gamma_{1,2}$; see also Figure 1.1. Since $U_{\min}(t;\mu)$ decreases monotonically in time (by (2.11) and (2.13)), we find that a spectrally compatible two-pulse solution must satisfy

$$U_{\min}(0) < U^*_{\min}(\mu) = \frac{A(\Delta\Gamma^*)}{\cosh \varepsilon^2\sqrt{\mu}\Delta\Gamma^*/2} = \frac{3}{16}\sqrt{3\mu},$$

if $\mu > \mu_{\mathrm{TP}}$ (3.37). In the context of Figure 1.1, in which $\mu = 5$, it follows that $U_{\min}(0)$ must be less than $0.72\ldots$. The evolution shown there is thus governed by Theorem 1.1.

Remark 3.7. The lower bound (3.37) on the pulse separation distance does not contradict the pulse-splitting behavior observed in the Gierer–Meinhardt equation [7], in which a stable two-pulse solution is observed with an $\mathcal{O}(1)$ pulse separation distance at the onset of splitting. It is shown in [7] that pulse splitting only occurs for $\mu = \mathcal{O}(1/\varepsilon^4)$. For these values of μ, $\Delta\Gamma^*(\mu) = \mathcal{O}(1)$ (3.37), which implies that the two V-pulses of $\Phi_\Gamma(\xi)$ are no longer well separated. Thus, the lower bound (3.37) agrees with the analysis of [7], since it implies that μ must be $\mathcal{O}(1/\varepsilon^4)$ in order to have two-pulse solutions that are not well separated.

3.3. The resolvent estimates and the semigroup. To establish estimates on the semigroup generated by the reduced linearization $\tilde{L}$ we begin with preliminary bounds on the resolvents of L_{11} and $\mathcal{L}$ in the norms defined in section 2.1. A key point is that the resolvent of L_{11} is strongly contractive on zero-mass functions.

LEMMA 3.1. *Let $\lambda \in \mathbf{C}$ be an $\mathcal{O}(1)$ distance from $\sigma(L_{11})$ and set $g = (L_{11}-\lambda)^{-1}f$. Then the following estimates hold uniformly in λ:*

$$\varepsilon\|g\|_{L^2} + \varepsilon^{-1}\|\partial_\xi g\|_{L^2} \le c\varepsilon^2\,\|f\|_{\widehat{L}^\infty}. \tag{3.39}$$

Moreover, for small total mass, $\overline{f}$ we have the improved estimate,

$$\varepsilon\|g\|_{L^2} + \varepsilon^{-1}\|\partial_\xi g\|_{L^2} \le c\left(\varepsilon^2|\overline{f}| + \varepsilon^4\left\|<x>f\right\|_{L^1}\right). \tag{3.40}$$

Proof. We take the Fourier transform of the equation $(L_{11}-\lambda)g = f$, obtaining

$$\widehat{g}(k) = \frac{1}{\sqrt{2\pi}}\frac{\varepsilon^4\widehat{f}(k)}{k^2+\varepsilon^4(\lambda+\mu)}. \tag{3.41}$$

Assuming that $f \in \widehat{L}^\infty$, the bound

$$\left(\int_{-\infty}^{\infty}\left|\frac{\varepsilon^4}{k^2+\varepsilon^4(\lambda+\mu)}\right|^2 dk\right)^{\frac{1}{2}} \le c\varepsilon, \tag{3.42}$$

for some $c > 0$, shows that $\|g\|_{L^2} \le c\varepsilon\|f\|_{\widehat{L}^\infty}$. Replacing $\widehat{f}$ with $ik\widehat{f}$ in (3.41) and calculating an integral similar to (3.42) gives $\|\partial_\xi g\|_{L^2} \le c\varepsilon^3\|f\|_{\widehat{L}^\infty}$. Together these

results yield (3.39). In the case that f has small mass, the identify $\widehat{f}(0) = \overline{f}$ and the fact that the norm $\|\langle x\rangle f\|_{L^1}$ controls the L^∞ norm of the k-derivative of the Fourier transform of f imply that $\widehat{f}$ is uniformly Lipschitz and small at zero, and so we have the estimate

$$|\widehat{f}(k)| \leq c\frac{|\overline{f}| + |k|}{1+|k|}\|\langle x\rangle f\|_{L^1}. \tag{3.43}$$

This inequality, used in (3.41), leads to the bound (3.40). □

We define $\mathcal{V}$ to be the eigenspace associated to the two exponentially small eigenvalues $\lambda_\pm^*$ of $\mathcal{L}^\dagger$, the adjoint of $\mathcal{L}$.

LEMMA 3.2. *Assume that $\lambda \in \mathbf{C}$ is an $\mathcal{O}(1)$ distance from $\sigma(\mathcal{L})\backslash\{\lambda_+, \lambda_-\}$. Then we have the following estimate, uniformly in λ, and for $\Gamma \in \mathcal{K}$:*

$$\|(\mathcal{L}-\lambda)^{-1} f\|_{H^1} \leq c\|f\|_{L^2} \tag{3.44}$$

for all $f \perp \mathcal{V}$.

Proof. The NLEP operator $\mathcal{L}$ is a finite rank perturbation of L_{22}, a self-adjoint Schrödinger operator, and hence is Fredholm. Moreover, away from its point spectrum, $\mathcal{L}-\lambda$ is boundedly invertible with $\mathcal{O}(1)$ inverse, uniformly in $\Delta\Gamma$ for $\Gamma \in \mathcal{K}$. If $f \perp \mathcal{V}$, then $\mathcal{L}-\lambda$ is uniformly invertible for λ in a neighborhood of $\lambda_\pm$. To obtain uniformity in λ for large $|\lambda|$ we observe that the resolvent of $\mathcal{L}$ can be explicitly constructed in terms of the resolvent of the selfadjoint operator L_{22} and that this later quantity decays like $(\mathrm{dist}(\lambda, \sigma(L_{22})))^{-2}$. That the resolvent of $\mathcal{L}$ maps into H^1 follows from a classic argument by contradiction. □

To study the resolvent of $\tilde{L}$ we project off the eigenspace $\{\Psi_+, \Psi_-\}$ associated to its small eigenvalues, $\lambda_\pm$. We introduce the space $X_\Gamma = \{\vec{U} \,|\, \|\vec{U}\|_X < \infty$ and $\pi_\Gamma \vec{U} = 0\}$, where the spectral projection is given in terms of the adjoint eigenfunctions $\Psi_\pm^\dagger$ by

$$\pi_\Gamma \vec{U} = \frac{(\vec{U}, \Psi_-^\dagger)}{(\Psi_-, \Psi_-^\dagger)}\Psi_- + \frac{(\vec{U}, \Psi_+^\dagger)}{(\Psi_+, \Psi_+^\dagger)}\Psi_+. \tag{3.45}$$

The complimentary projection is $\tilde{\pi}_\Gamma = I - \pi_\Gamma$. Assuming the spectral compatibility of Φ_Γ, the space X_Γ is associated to temporally decaying solutions of the semigroup generated by $\tilde{L}_\Gamma$, while $\tilde{X}_\Gamma = \mathcal{R}\pi_\Gamma$ is the eigenspace associated to the two exponentially small eigenvalues $\lambda_\pm$. To characterize the projections we need asymptotics for these eigenfunctions.

LEMMA 3.3. *The small eigenvalue eigenfunctions have the following asymptotic form:*

$$\Psi_\pm = \begin{pmatrix} 0 \\ \phi_1' \pm \phi_2' \end{pmatrix} + \text{ exponentially small}, \tag{3.46}$$

$$\Psi_\pm^\dagger = \begin{pmatrix} 0 \\ \phi_1' \pm \phi_2' \end{pmatrix} + \mathcal{O}\left(\varepsilon^4\right), \tag{3.47}$$

in the X-norm.

Proof. The expansion for the eigenfunctions follows from classical results. For the adjoint eigenfunctions we present the case for a single pulse; the generalization to two-pulses is straightforward. The adjoint operator is given by

$$\tilde{L}_\Gamma^\dagger = \begin{pmatrix} L_{11} & -\dfrac{\phi_1^2}{A^2} \\ 2\varepsilon^{-2}\phi_1 \otimes \delta_{\Gamma_1} & L_{22} \end{pmatrix}, \tag{3.48}$$

where L_{11} and L_{22} are given in (3.7). Writing $\Psi^\dagger = (\Psi_1^\dagger, \Psi_2^\dagger)^t$ and taking λ_+^* exponentially small, we solve for the second component of $\Psi^\dagger$, noting that ϕ_1 is in the range of L_{22} since it is orthogonal to its kernel, ϕ_1',

$$\Psi_2^\dagger = \beta\phi_1' - 2\varepsilon^{-2}\Psi_1^\dagger(\Gamma_1)L_{22}^{-1}\phi_1, \tag{3.49}$$

where β is a free parameter. Solving for $\Psi_1^\dagger$ we have

$$\Psi_1^\dagger = \frac{\beta}{A^2}L_{11}^{-1}\phi_1^2\phi_1' - \frac{2\Psi_1^\dagger(\Gamma_1)}{\varepsilon^2 A^2}L_{11}^{-1}\left(\phi_1^2 L_{22}^{-1}\phi_1\right). \tag{3.50}$$

The function $\phi_1^2\phi_1'$ has zero mass, so from (3.40) we have

$$\|L_{11}^{-1}\phi_1^2\phi_1'\|_{L^\infty} \le c\varepsilon^4. \tag{3.51}$$

It can be verified that $\phi_1^2 L_{22}^{-1}\phi_1$ is a positive $\mathcal{O}(1)$ function, and thus we know $L_{11}^{-1}\left(\phi_1^2 L_{22}^{-1}\phi_1\right)\big|_{\xi=\Gamma_1}$ is nonzero and $\mathcal{O}(1)$. Evaluating (3.50) at $\xi = \Gamma_1$ and solving for $\Psi_1^\dagger(\Gamma_1)$ shows that $\Psi_1^\dagger(\Gamma_1) = \mathcal{O}(\varepsilon^6)$. Substituting this back into (3.50) and choosing $\beta = 1$ yields the equivalent of (3.47) in the one-pulse case. ∎

With these results we may estimate the resolvent of $\tilde{L}_\Gamma$ restricted to X_Γ.

PROPOSITION 3.8. *Let λ be an $\mathcal{O}(1)$ distance from $\sigma(\tilde{L})\backslash\{\lambda_+, \lambda_-\}$ and denote $G = (\tilde{L} - \lambda)^{-1}F$. For $F \in X_\Gamma$, we have the following estimates on the resolvent of $\tilde{L}$, holding uniformly in λ, and in $\Gamma \in \mathcal{K}$:*

$$\|G\|_X \le c\left(\varepsilon^2\|F_1\|_{L^1} + \|F_2\|_{L^2}\right). \tag{3.52}$$

If, in addition, the mass of F_1 is small, then we have the improved estimate

$$\|G\|_X \le c\left(\varepsilon^2|\overline{F}_1| + \varepsilon^4\| <x> F_1\|_{L^1} + \|F_2\|_{L^2}\right). \tag{3.53}$$

Proof. By analogy with the eigenvalue problem we solve for G_1:

$$G_1 = (L_{11} - \lambda)^{-1}F_1 + \alpha_1 H(\lambda, \xi - \Gamma_1) + \alpha_2 H(\lambda, \xi - \Gamma_2), \tag{3.54}$$

where H is given by (3.16), and $\alpha_k = (G_2, \phi_k)_{L^2}$, for $k = 1, 2$. The second component of G satisfies

$$(L_{22} - \lambda)G_2 = F_2 + \frac{V_0^2}{A^2}\left((L_{11} - \lambda)^{-1}F_1 + \alpha_1 H(\xi - \Gamma_1) + \alpha_2 H(\xi - \Gamma_2)\right). \tag{3.55}$$

Approximating the product $V_0^2 H$ as in the eigenvalue problem, we find the equation

$$(\mathcal{L} - \lambda)\, G_2 = F_2 + \frac{V_0^2}{A^2}(L_{11} - \lambda)^{-1}F_1, \tag{3.56}$$

where the NLEP operator $\mathcal{L}$ is defined in (3.23).

From the asymptotics on $\Psi_\pm^\dagger$ the condition $F \in X_\Gamma$ is equivalent to the right-hand side of (3.56) being orthogonal to $\mathcal{V}$. From Proposition 3.1 the point spectrum of $\tilde{L}$, less its exponentially small eigenvalues, agrees with the point spectrum of $\mathcal{L}$, less its exponentially small eigenvalues, up to $\mathcal{O}(\varepsilon^2)$. So λ is an $\mathcal{O}(1)$ distance from $\sigma(\mathcal{L})\backslash\{\lambda_\pm\}$ and the estimate (3.44) applied to (3.56) yields

$$\begin{aligned}\|G_2\|_{H^1} &\le c\left(\|F_2\|_{L^2} + \|V_0^2(L_{11} - \lambda)^{-1}F_1\|_{L^2}\right)\\ &\le c\left(\|F_2\|_{L^2} + \|(L_{11} - \lambda)^{-1}F_1\|_{L^\infty}\right).\end{aligned} \tag{3.57}$$

From (3.39) and (1.2) we find that

$$\|G_2\|_{H^1} \leq c\left(\|F_2\|_{L^2} + \varepsilon^2\|F_1\|_{L^1}\right). \tag{3.58}$$

If F_1 has small mass, then by applying (3.40) we have the improved estimate

$$\|G_2\|_{H^1} \leq c\left(\|F_2\|_{L^2} + \varepsilon^2|\overline{F}_1| + \varepsilon^4\left\|<x>F_1\right\|_{L^1}\right). \tag{3.59}$$

From (3.54) and (3.39) we find that

$$\|G_1\|_{L^2} \leq c\left(\varepsilon\|F_1\|_{L^1} + \|G_2\|_{L^2}\|H(\lambda)\|_{L^2}\right), \tag{3.60}$$

but since

$$|\widehat{H}(k,\lambda)| \leq c\left|\frac{\varepsilon^2}{k^2 + \varepsilon^4(\lambda+\mu)}\right|, \tag{3.61}$$

we have $\|H(\lambda)\|_{L^2} \leq c\varepsilon^{-1}$ and we obtain

$$\|G_1\|_{L^2} \leq c\left(\varepsilon\|F_1\|_{L^1} + \varepsilon^{-1}\|F_2\|_{L^2}\right). \tag{3.62}$$

A similar argument yields

$$\|\partial_\xi G_1\|_{L^2} \leq c\left(\varepsilon^3\|F_1\|_{L^1} + \varepsilon\|F_2\|_{L^2}\right), \tag{3.63}$$

which verifies (3.52).

If F_1 has small mass, then applying (3.40) to (3.57) yields the improved estimate

$$\|G_2\|_{H^1} \leq c\left(\|F_2\|_{L^2} + \varepsilon^2|\overline{F}_1| + \varepsilon^4\left\|<x>F_1\right\|_{L^1}\right). \tag{3.64}$$

Following the arguments laid out in (3.60)–(3.63) yields (3.53). □

Since $\tilde{L}$ is an analytic operator we can generate its semigroup from the Laplace transform of the resolvent. We fix the contour C in the complex plane as depicted in Figure 3.3 and generate the semigroup S associated to $\tilde{L}|_{X_\Gamma}$ via the contour integral

$$S(t)F = \frac{1}{2\pi i}\int_C e^{\lambda t}(\lambda - \tilde{L})^{-1}F\, d\lambda, \tag{3.65}$$

where we assume that $F \in X_\Gamma$. The semigroup inherits the following properties from the resolvent.

PROPOSITION 3.9. *Let $\mu > \mu_{\text{Hopf}}$ and $\Delta\Gamma > \Delta\Gamma^*(\mu)$ be given and let $\nu > 0$ be as given by Proposition* 3.4. *The solution $\vec{U}$ of $\vec{U} = S(t)F$, where $F \in X_\Gamma$, satisfies*

$$\|\vec{U}\|_X \leq Me^{-\nu t}\left(\varepsilon^2\|F_1\|_{L^1} + \|F_2\|_{L^2}\right), \tag{3.66}$$

for some $M > 0$ independent of $\Delta\Gamma > \Delta\Gamma^*(\mu)$. If, in addition, F_1 has small mass, then we have the improved estimate

$$\|\vec{U}\|_X \leq Me^{-\nu t}\left(\varepsilon^2|\overline{F}_1| + \varepsilon^4\| <x>F_1\|_{L^1} + \|F_2\|_{L^2}\right). \tag{3.67}$$

Proof. By Proposition 3.4, the conditions on μ and Γ imply that $\sigma(\tilde{L})\backslash\{\lambda_+, \lambda_-\}$ is contained within the interior of the contour C, and $\text{dist}(\sigma(\tilde{L}), C) = \mathcal{O}(1)$. The estimates on the semigroup follow directly from the contour integral representation (3.65) of $S(t)$, the resolvent estimates (3.52)–(3.53), and the uniformity of these estimates over the contour C. □

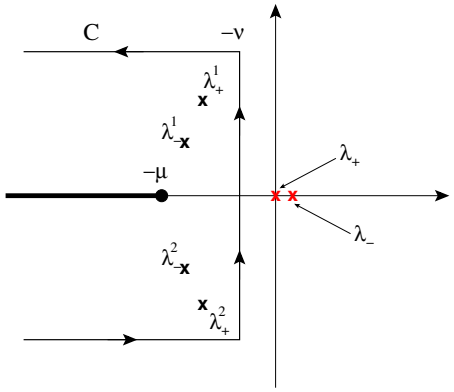


FIG. 3.2. *The spectrum $\sigma(\tilde{L})$ of the reduced operator as determined by Proposition* 3.3*, and the contour C used to generate the semigroup S. Depiction is for the case $\mu > \mu_{\mathrm{Hopf}}$ and $\Delta\Gamma > \Delta\Gamma^*(\mu)$ for which $\lambda_{\pm}^{1,2}$ are within the left-half plane and $J_{\pm} = 2$; see Proposition* 3.4*. The eigenspace corresponding to the small point spectrum $\{\lambda_{\pm}\}$ is projected away and is not contained within the contour.*

4. Nonlinear stability via the RG method. We adapt the RG method developed in [14] to the singular perturbation setting of the Gierer–Meinhardt equations. We assume at time t_0 that our initial data $\vec{U}_0$ satisfies

$$\|\Phi_{\Gamma_*} - \vec{U}_0\| \leq \delta, \tag{4.1}$$

for some $\Gamma_* \in \mathcal{K}$. The following proposition, adapted from Proposition 2.2 of [14], permits us to choose our base point Γ_0 about which we develop our local coordinate system.

PROPOSITION 4.1. *Fix $\delta \ll 1$. Given $\vec{U}_0$ and $\Gamma_* \in \mathcal{K}$ satisfying $\|W_*\|_X \leq \delta$, for $W_* \equiv \Phi_{\Gamma_*} - \vec{U}_0$, there exists $M > 0$, independent of $\vec{U}_0$ and Γ_*, and a smooth function $\mathcal{H} : X \mapsto \mathcal{K}$ such that $\Gamma = \Gamma_* + \mathcal{H}(W_*)$ satisfies*

$$W_0 \equiv \vec{U}_0 - \Phi_\Gamma \in X_\Gamma. \tag{4.2}$$

Moreover, if $W_* \in X_{\tilde{\Gamma}}$ for some $\tilde{\Gamma} \in \mathcal{K}$, then

$$|\Gamma - \Gamma_*| \leq M_0 \|W_*\|_X |\Gamma_* - \tilde{\Gamma}|. \tag{4.3}$$

Proof. Since

$$W_0 = W_* + \Phi_\Gamma - \Phi_{\Gamma_*}, \tag{4.4}$$

the condition (4.2) is equivalent to

$$0 = \pi_\Gamma W_0 = \pi_\Gamma \left(W_* + \Phi_\Gamma - \Phi_{\Gamma_*} \right). \tag{4.5}$$

Since $\Psi_\pm^\dagger$, are approximately spanned by $(0, \phi_1')^t$ and $(0, \phi_2')^t$, and $\Phi_{\Gamma,2} = V_0 + \mathcal{O}(\varepsilon^2)$, our equations $\Lambda = (\Lambda_1, \Lambda_2)^t$ are equivalent, up to $\mathcal{O}(\varepsilon^2)$, to

$$\Lambda_1(\Gamma, W_*) \equiv \Big(W_{2,*} + V_0(\Gamma) - V_0(\Gamma_*), \phi_1'(\cdot, \Gamma_1) \Big)_{L^2} = 0, \tag{4.6}$$

$$\Lambda_2(\Gamma, W_*) \equiv \Big(W_{2,*} + V_0(\Gamma) - V_0(\Gamma_*), \phi_2'(\cdot, \Gamma_2) \Big)_{L^2} = 0. \tag{4.7}$$

Since $\Lambda(\Gamma_*, 0) = 0$ and the Γ gradient of Λ given by

$$\nabla_\Gamma \Lambda\Big|_{(\Gamma=\Gamma_*, W_*=0)} = \begin{pmatrix} -\|\phi_1'\|_{L^2} & 0 \\ 0 & -\|\phi_1'\|_{L^2} \end{pmatrix} + \mathcal{O}(\varepsilon^2) \tag{4.8}$$

is uniformly invertible, the implicit function theorem guarantees the existence of a smooth function $\mathcal{H}$ which provides the solution of (4.2) and in a neighborhood about the manifold $\mathcal{M}$ defined in (2.6). The interval of existence of $\mathcal{H}$ may be chosen uniformly in Γ since the solution of (4.2) behaves smoothly as $\Delta\Gamma \to \infty$.

If, in addition, we have $W_* \in X_{\tilde{\Gamma}}$, then $(W_{2,*}, \phi_k'(\tilde{\Gamma}_k))_{L^2} = \mathcal{O}(\varepsilon^4)$ for $k = 1, 2$. We see that

$$\left|\Big(W_{2,*}, \phi_k'(\Gamma_k)\Big)_{L^2}\right| \le \left|\Big(W_{2,*}, \phi_k'(\tilde{\Gamma}_k) - \phi_k'(\Gamma_k)\Big)_{L^2}\right| \le M_0 \|W_*\|_{L^2} |\tilde{\Gamma} - \Gamma|. \qquad \square \tag{4.9}$$

4.1. The projected equations. To begin the RG procedure we freeze $\Gamma = \Gamma_0$ in X_{Γ_0}, where Γ_0 is the base point provided by Proposition 4.1, and change variables as

$$\vec{U}(t) = \Phi_\Gamma + W, \tag{4.10}$$

where $W \in X_{\Gamma_0}$ and $\Gamma = \Gamma(t)$. Comparing to (3.2), we write the evolution for the remainder W as

$$W_t + \frac{\partial \Phi}{\partial \Gamma}\dot{\Gamma} = \mathbf{R} + \tilde{L}_{\Gamma_0} W + \left(L_\Gamma - \tilde{L}_{\Gamma_0}\right) W + \mathcal{N}(W), \tag{4.11}$$

$$W(\xi, 0) = W_0, \tag{4.12}$$

where $W_0 = W_* + \Phi_{\Gamma_0} - \Phi_{\Gamma_*}$. The terms $\Delta L \equiv L_\Gamma - \tilde{L}_{\Gamma_0}$ include both the approximations made to the linear operator and the secular growth implicit in the sliding of Γ away from Γ_0.

To enforce $W \in X_{\Gamma_0}$ we impose the nondegeneracy condition $\frac{\partial}{\partial t}\pi_0 W = 0$, where $\pi_0 = \pi_{\Gamma_0}$ is given by (3.45). Since π_0 is independent of time, the nondegeneracy condition is equivalent to $\pi_0 W_t = 0$, and, moreover, as π_0 commutes with $\tilde{L}_{\Gamma_0}$ it follows that $\pi_0 \tilde{L}_{\Gamma_0} W = \tilde{L}_{\Gamma_0} \pi_0 W = 0$. The nondegeneracy condition is thus equivalent to the pair of equations obtained by projecting onto $\Psi_+^\dagger$ and $\Psi_-^\dagger$,

$$\left(\frac{\partial \Phi}{\partial \Gamma}\dot{\Gamma}, \Psi_\pm^\dagger\right)_{L^2} = \left(\mathbf{R} + \Delta L\, W + \mathcal{N}(W), \Psi_\pm^\dagger\right)_{L^2}. \tag{4.13}$$

From the form of the semistrong pulse solutions, and assuming momentarily that $\dot{\Gamma} = \mathcal{O}(\varepsilon^2)$, we calculate

$$\frac{\partial \Phi_\Gamma}{\partial \Gamma}\dot{\Gamma} = \begin{pmatrix} 0 \\ \phi_1'\dot{\Gamma}_1 + \phi_2'\dot{\Gamma}_2 \end{pmatrix} + \begin{pmatrix} \mathcal{O}(\varepsilon^3) \\ \mathcal{O}(\varepsilon^4) \end{pmatrix} \tag{4.14}$$

componentwise in the L^2 norm. Using the form of the adjoint eigenvector (3.47) and (3.6), (4.13) may be written as

$$\begin{pmatrix} \|\phi_1'\|_{L^2}^2 + \mathcal{O}(\varepsilon^4) & \|\phi_2'\|_{L^2}^2 + \mathcal{O}(\varepsilon^4) \\ \|\phi_1'\|_{L^2}^2 + \mathcal{O}(\varepsilon^4) & -\|\phi_2'\|_{L^2}^2 + \mathcal{O}(\varepsilon^4) \end{pmatrix} \dot{\Gamma} = \begin{pmatrix} \left(\mathbf{R} + \Delta L\, W + \mathcal{N}(W), \Psi_+^\dagger\right)_{L^2} \\ \left(\mathbf{R} + \Delta L\, W + \mathcal{N}(W), \Psi_-^\dagger\right)_{L^2} \end{pmatrix}. \tag{4.15}$$

Again using the asymptotic form of the adjoint eigenfunctions $\Psi^\dagger_\pm$ we may neglect the contribution from $\Psi^\dagger_{\pm,1}$ in the inner products on the right-hand side of (4.15). In particular, from the L^1 bounds on R_1 from (2.18), we have

$$(R_1, \Psi^\dagger_{\pm,1})_{L^2} \leq \|R_1\|_{L^1} \|\Psi^\dagger_{\pm,1}\|_{L^\infty} = \mathcal{O}(\varepsilon^5). \tag{4.16}$$

Inverting the matrix on the left-hand side and using the expansions for $\Psi^\dagger_{\pm,2}$, we arrive at the equations of motion for Γ which show explicitly the coupling between the remainder W and the pulse evolution,

$$\dot\Gamma_k = -\frac{(R_2 + [\Delta L\ W]_2 + \mathcal{N}_2, \phi_k'(\cdot;\Gamma_0))_{L^2}}{(\phi_k', \phi_k')_{L^2}} + \mathcal{O}(\varepsilon^5, \varepsilon^4 \|W\|_X). \tag{4.17}$$

To simplify the equation for the evolution of the remainder W, we introduce the reduced residual

$$\tilde{\mathbf{R}} = \tilde\pi_\Gamma \left(\mathbf{R} - \frac{\partial \Phi_\Gamma}{\partial \Gamma} \dot\Gamma \right), \tag{4.18}$$

and observe from the asymptotic description (2.17) of R_2 that the projection removes the leading order term from the second component of the residual. By Lemma 2.1, the reduced residual enjoys the estimates

$$\|\tilde R_1\|_{L^1} \leq \mathcal{O}(\varepsilon), \tag{4.19}$$

$$\|\tilde R_2\|_{L^2} \leq \mathcal{O}(\varepsilon^4). \tag{4.20}$$

The evolution for the remainder W is now given by

$$W_t = \tilde{\mathbf{R}} + \tilde L_0 W + \tilde\pi_0 \left(\Delta L\ W + \mathcal{N} \right), \tag{4.21}$$

$$W(\xi, t_0) = W_0, \tag{4.22}$$

where $\tilde L_0 = \tilde L_{\Gamma_0}$ and $\tilde\pi_0 = I - \pi_{\Gamma_0}$. The point of the reduction of the Gierer–Meinhardt equation (2.3) to the projected residual equation (4.21), in the case of two-pulse dynamics, is that the asymptotically relevant and the asymptotically negligible terms are now evident. The evolution for W is controlled by the first two terms on the right-hand side of (4.21); we will show that the last two terms are asymptotically irrelevant, until $\Gamma - \Gamma_0$ is so large that the secularity implicit in ΔL forces an update of Γ_0.

4.2. Decay of the remainder. We identify the duration of each renormalization interval and quantify the decay of the remainder W over this interval. To control the dynamics we introduce the quantities

$$T_0(t) = \sup_{t_0 < s < t} e^{\nu(s-t_0)} \|W(s)\|_X, \tag{4.23}$$

$$T_1(t) = \sup_{t_0 < s < t} |\Gamma(s) - \Gamma_0|. \tag{4.24}$$

The first enforces the decay of the remainder W, and the second measures the distance the pulse positions have moved from their frozen base point. The variation of constants formula applied to (4.21) yields the solution

$$W(\xi, t) = S(\Delta t) W_0 + \int_{t_0}^t S(t-s)(\tilde R + \tilde\pi_0 \left(\Delta L\ W + \mathcal{N} \right)) \, ds, \tag{4.25}$$

where we have introduced $\Delta t = t - t_0$.

To estimate the distance that the pulse locations Γ have moved from the base point Γ_0 we examine the equations (4.17). We break ΔL into secular and reductive parts $\Delta L = \Delta L_s + \Delta L_r$, where $\Delta L_s = L_\Gamma - L_{\Gamma_0}$ and $\Delta L_r = L_{\Gamma_0} - \tilde{L}_{\Gamma_0}$, and remark that

$$\|[\Delta L\ W]_2\|_{L^2} \le \|\,[\Delta L_s\, W]_2\,\|_{L^2} + \|\,[\Delta L_r\, W]_2\,\|_{L^2} \tag{4.26}$$

$$\le c\left(|\Gamma - \Gamma_0| + \varepsilon^2\right)\|W\|_{L^2} \tag{4.27}$$

$$\le c(T_1(t) + \varepsilon^2)e^{-\nu(t-t_0)}T_0(t), \tag{4.28}$$

where the estimates on ΔL_s and ΔL_r are described in more detail below. From the form (3.4) of the regularized nonlinearity it is straightforward to obtain the estimate

$$|(\mathcal{N}_2, \phi_k')_{L^2}| \le c\|W\|_X^2. \tag{4.29}$$

With these bounds in hand, the drift of the pulses is controlled by their speed,

$$T_1(t) \le \int_{t_0}^{t_0+\Delta t} |\dot{\Gamma}(s)|ds$$

$$\le \int_{t_0}^{t_0+\Delta t} c\left(\|R_2\|_{L^2} + (\varepsilon^2 + T_1(t))e^{-\nu(s-t_0)}T_0(t) + e^{-2\nu(s-t_0)}T_0^2(t)\right)ds \tag{4.30}$$

$$\le c(\varepsilon^2\Delta t + (\varepsilon^2 + T_1)T_0 + T_0^2). \tag{4.31}$$

For T_0 small enough we can eliminate T_1 from the right-hand side, and neglecting T_0 in the sum $T_0 + \Delta t$, we obtain

$$T_1 \le c(\varepsilon^2\Delta t + T_0^2). \tag{4.32}$$

Turning to bounds on the remainder, we estimate the irrelevant terms first. The secular term takes the form

$$\Delta L_s = \begin{pmatrix} 0 & 2\varepsilon^{-2}\Big(V_0(\cdot;\Gamma) - V_0(\cdot;\Gamma_0)\Big) \\ V_{12}(\cdot;\Gamma) - V_{12}(\cdot,\Gamma_0) & V_{22}(\cdot;\Gamma) - V_{22}(\cdot;\Gamma_0) \end{pmatrix}, \tag{4.33}$$

where V_{12} and V_{22} denote the potentials in the $\tilde{L}_{12}$ and $\tilde{L}_{22}$ components of $\tilde{L}$. Since each potential, V_0, V_{12}, and V_{22}, decays rapidly away from the pulse locations, the difference between the potential centered at pulse locations Γ_0 and at Γ scales like $\Gamma - \Gamma_0$ in any reasonable norm. In particular,

$$\|V_0(\cdot,\Gamma) - V_0(\cdot,\Gamma_0)\|_{H^1} + \|V_0(\cdot,\Gamma) - V_0(\cdot,\Gamma_0)\|_{L^1} \le c|\Gamma - \Gamma_0|, \tag{4.34}$$

and similarly for V_{12} and V_{22}. Using these estimates it follows directly that

$$\|[\Delta L_s\, W]_1\|_{L^1} \le c\varepsilon^{-2}\|V_0(\cdot,\Gamma) - V_0(\cdot,\Gamma_0)\|_{L^1}\|W_2\|_{L^\infty} \tag{4.35}$$

$$\le c\varepsilon^{-2}T_1\|W\|_X. \tag{4.36}$$

Similarly

$$\|[\Delta L_s\, W]_2\|_{L^2} \le c\|V_{12}(\cdot,\Gamma) - V_{12}(\cdot,\Gamma_0)\|_{H^1}\|W_1\|_X$$

$$+ \|V_{22}(\cdot,\Gamma) - V_{22}(\cdot,\Gamma_0)\|_{H^1}\|W_2\|_X \tag{4.37}$$

$$\le cT_1\|W\|_X. \tag{4.38}$$

Combining these estimates with the unweighted semigroup estimate (3.66) we find that

$$\|S(t-s)\tilde{\pi}_0(\Delta L_s\ W(s))\|_X \leq Me^{-\nu(t-s)}T_1(s)\|W\|_X. \tag{4.39}$$

The small mass version of the semigroup estimate plays a key role in controlling the reductive term $\Delta L_r\ W$ given by

$$\Delta L_r = \begin{pmatrix} 0 & 2\varepsilon^{-2}\Big(V_0 - \delta_{\Gamma_1}\otimes\phi_1(\Gamma_0) - \delta_{\Gamma_2}\otimes\phi_2(\Gamma_0)\Big) + \mathcal{O}(1) \\ \mathcal{O}(\varepsilon^2 V_0^2) & \mathcal{O}(\varepsilon^2 V_0) \end{pmatrix}, \tag{4.40}$$

where here the $\mathcal{O}$ indicates pointwise estimates. The first component of the reductive term can be decomposed as

$$[\Delta L_r\, W]_1 = \varepsilon^{-2}\left(\Sigma_1 + \Sigma_2\right), \tag{4.41}$$

where

$$\Sigma_k = \phi_k - \delta_{\Gamma_1}\otimes\phi_k + \mathcal{O}(\varepsilon^2\phi_k). \tag{4.42}$$

That is, each Σ_k decays at an $\mathcal{O}(1)$ exponential rate away from $x = \Gamma_k$, and, moreover, at leading order each component $\Sigma_k W_2$ is mass-free for $k = 1, 2$. In particular, the total mass of the 1 component of $\Delta L_r\, W$ arises only from the higher order corrections,

$$|\overline{[\Delta L_r\, W]_1}| \leq c\|W_2\|_X. \tag{4.43}$$

In the weighted norms we estimate

$$\begin{aligned} \|<x-\Gamma_k>\Sigma_k W_2\|_{L^2} &\leq \varepsilon^{-2}(\|\langle x-\Gamma_k\rangle\delta_{\Gamma_k}\|_{L^2} + \|\langle x-\Gamma_k\rangle\phi_k\|_{L^2})\|W_2\|_{L^\infty} \\ &\leq c\varepsilon^{-2}\|W\|_X, \end{aligned} \tag{4.44}$$

for $k = 1, 2$. For the second component we observe that

$$\|[\Delta L_r W]_2\|_{H^1} \leq c\varepsilon^2\|W\|_{H^1}, \tag{4.45}$$

so from the weighted semigroup estimate (3.67) we find

$$\|S(t-s)\tilde{\pi}_0(\Delta L_r\ W(s))\|_X \leq Me^{-\nu(t-s)}\varepsilon^2\|W\|_X \tag{4.46}$$

independent of the pulse spacing $\Delta\Gamma$.

Finally, for the nonlinear term given by (3.4), it is easy to verify

$$\|S(t-s)\tilde{\pi}_0\mathcal{N}\|_X \leq Me^{-\nu(t-s)}(\|W_2^2\|_{L^1} + \|W_2^2\|_{L^2} + \|W_1^2\|_{L^2}) \tag{4.47}$$

$$\leq Me^{-\nu(t-s)}\|W\|_X^2. \tag{4.48}$$

From the bounds on the reduced residual (4.19)–(4.20) and the semigroup estimate we obtain

$$\|S(t-s)\tilde{R}\|_X \leq M\varepsilon^3 e^{-\nu(t-s)}. \tag{4.49}$$

Taking the X-norm of variation of constants solution for W, (4.25), and using the estimates outlined above we obtain

(4.50)
$$\|W(t)\|_X \leq M\left(e^{-\nu\Delta t}\|W(t_0)\|_X + \int_{t_0}^{t} e^{-\nu(t-s)}\left[\varepsilon^3 + \left(\varepsilon^2 + T_1(s)\right)\|W(s)\|_X + \|W(s)\|_X^2\right] ds\right).$$

To estimate the decay of $\|W(t')\|_X$ for $t' \in (t_0, t)$ we evaluate (4.50) at $t = t'$, multiply by $e^{\nu(t'-t_0)}$, and take the sup over $t' \in (t_0, t)$ obtaining

$$T_0(t) \leq M\left(T(t_0) + \int_{t_0}^{t}\left[\varepsilon^3 e^{\nu(s-t_0)} + \left(\varepsilon^2 + T_1(t)\right)T_0(t) + e^{-\nu s}T_0(t)^2\right] ds\right) \tag{4.51}$$

$$\leq M(T_0(t_0) + \varepsilon^3 e^{\nu\Delta t} + \left(\varepsilon^2 + T_1(t)\right)\Delta t\, T_0(t) + T_0(t)^2). \tag{4.52}$$

From (4.32) we may eliminate T_1 from the T_0 estimate,

$$T_0(t) \leq M(T_0(t_0) + \varepsilon^3 e^{\nu\Delta t} + \varepsilon^2(\Delta t + (\Delta t)^2)T_0(t) + T_0(t)^2 + \Delta t\, T_0^3). \tag{4.53}$$

For $\Delta t \ll \min\left\{\varepsilon^{-1}, T_0^{-1}\right\}$ the term $M\varepsilon^2((\Delta t)^2 + \Delta t) < \frac{1}{2}$, and we may eliminate the linear term in T_0 from the right-hand side. In addition we may absorb the cubic term in T_0 into the quadratic one. With these reductions (4.53) becomes

$$T_0 \leq 2M(T_0(t_0) + \varepsilon^3 e^{\nu\Delta t} + T_0^2). \tag{4.54}$$

The quadratic equation in T_0,

$$0 = T_0(t_0) + \varepsilon^3 e^{\nu\Delta t} - \frac{1}{2M}T_0 + T_0^2, \tag{4.55}$$

has two positive real roots so long as $T_0(t_0) + \varepsilon^3 e^{\nu\Delta t} \ll 1$. The smaller of these roots, r_0, takes the form

$$r_0 = 2M(T_0(t_0) + \varepsilon^3 e^{\nu\Delta t}) + \mathcal{O}(T_0(t_0) + \varepsilon^3 e^{\nu\Delta t})^2, \tag{4.56}$$

while the larger is

$$r_1 = \frac{1}{2M} + \mathcal{O}(T_0(t_0) + \varepsilon^3 e^{\nu\Delta t}). \tag{4.57}$$

Thus if $T_0(t_0) \ll 1$ and $\varepsilon^3 e^{\nu\Delta t} \ll 1$, then there is an *excluded region*, either $0 < T_0 < r_0$ or $r_1 < T_0 < \infty$. Since $T_0(t_0) < r_0$ and T_0 is continuous in t, we see that

$$T_0(t) \leq r_0 \leq M(T_0(t_0) + \varepsilon^3 e^{\nu\Delta t}) \tag{4.58}$$

so long as

$$\Delta t \leq \frac{3\beta|\log\varepsilon|}{\nu} \tag{4.59}$$

for any fixed $\beta < 1$. This condition on Δt prevents the secularity from dominating the linear operator; in particular, it is a stronger condition on Δt than that imposed after (4.53). This implies that

$$\|W(t)\|_X \leq M(e^{-\nu(t-t_0)}\|W(t_0)\|_X + \varepsilon^3), \quad \text{for } t \in \left(t_0, t_0 + \frac{3\beta|\log\varepsilon|}{\nu}\right) \tag{4.60}$$

and, in particular, for $t_1 = t_0 + \Delta t$ we have

$$\|W(t_1)\|_X \leq M(\varepsilon^{3\beta}\|W(t_0)\|_X + \varepsilon^3). \tag{4.61}$$

4.3. The RG iteration. We break the time evolution into a series of initial value problems, tracking the decay of the remainder over the long-time scale of many RG iterations. We fix $\beta < 1$ and $\Delta t = \frac{3\beta|\log\varepsilon|}{\nu}$. The renormalization times are defined sequentially,

$$t_n = t_{n-1} + \Delta t. \tag{4.62}$$

We break the evolution of W into disjoint intervals $I_n = [t_n, t_{n+1})$. On each interval I_n we solve the initial value problem (4.21) with initial data $W(t_n) \in X_{\Gamma_n}$, with the quantities $T_{0,n}$ and $T_{1,n}$ corresponding to (4.23)–(4.24) over I_n. The renormalization map, $\mathcal{G}$, takes the initial data $W_{n-1} = W(t_{n-1})$ for the initial value problem on interval I_{n-1} and returns the initial data $W_n = W(t_n)$ for the initial value problem on the interval I_n,

$$\mathcal{G}W_{n-1} = W_n. \tag{4.63}$$

Arguing inductively, the initial data and the new base point Γ_n are obtained from $W(t_n^-)$, the end-value of the evolution of W over I_{n-1}, by applying Proposition 4.1. Indeed, we know that $W(t_n^-) \in X_{\Gamma_{n-1}}$ and so from (4.3) we have

$$|\Gamma_n - \Gamma(t_n^-)| \le M_0\|W(t_n^-)\|_X|\Gamma(t_n^-) - \Gamma(t_{n-1})| \le M_0\|W(t_n^-)\|_X T_{1,n-1}(t). \tag{4.64}$$

From the estimates on Δt and $T_{1,n-1}$, we bound the jump in Γ at renormalization by

$$|\Gamma_n - \Gamma(t_n^-)| \le M_0\left(|\log\varepsilon|\varepsilon^2 + T_{0,n-1}^2\right)\|W(t_n^-)\|_X. \tag{4.65}$$

The solution at time $t = t_n$ is independent of the decomposition,

$$\vec{U}(t_n) = \Phi_{\Gamma(t_n^-)} + W(t_n^-) = \Phi_{\Gamma_n} + W_n, \tag{4.66}$$

and we may bound the jump in W at each renormalization

$$\begin{aligned}\|W(t_n^-) - W(t_n)\|_X &= \|\Phi_{\Gamma(t_n^-)} - \Phi_{\Gamma_n}\|_X \le c|\Gamma_n - \Gamma(t_n^-)| \\ &\le M_0\left(|\log\varepsilon|\varepsilon^2 + T_{0,n-1}^2\right)\|W(t_n^-)\|_X,\end{aligned} \tag{4.67}$$

where we used the fact that U_0 is $\mathcal{O}(1)$ X-Lipschitz in Γ, as follows from (3.5) and (1.2). From (4.58), using the equality $T_{0,n-1}(t_{n-1}) = \|W_{n-1}\|_X$, we have the estimate

$$T_{0,n-1} \le M_1(\|W_{n-1}\|_X + \varepsilon^{3(1-\beta)}). \tag{4.68}$$

Combining the estimates (4.67) and (4.68) with (4.61), we obtain a bound on $\mathcal{G}W_{n-1} = W_n$,

$$\begin{aligned}\|\mathcal{G}W_{n-1}\|_X \le (1 + M_0[|\log\varepsilon|\varepsilon^2 + M_1^2(\|W(t_{n-1})\|_X + \varepsilon^{3(1-\beta)})^2]) \\ \times M(\varepsilon^{3\beta}\|W(t_{n-1})\|_X + \varepsilon^3).\end{aligned} \tag{4.69}$$

Neglecting the terms involving positive powers of ε within the first parenthesis on the left-hand side, we may bound $\|W(t_n)\|_X$ by η_n, the solution of the map

$$\eta_n = M(1 + M_2\eta^2)(\varepsilon^{3\beta}\eta_{n-1} + \varepsilon^3), \tag{4.70}$$

with $\eta_0 = \|W(\cdot, t_0)\|_X$, and $M_2 = M_0M_1^2$. It is easy to see for $\eta_0 = \mathcal{O}(1)$ and ε sufficiently small that

$$\eta_n \to \frac{M}{1 - \varepsilon^{3\beta}M}\varepsilon^3, \tag{4.71}$$

as $n \to \infty$. Since $\|W(\cdot, t_n)\|_X \le \eta_n$, the estimate (4.61) yields the result (1.4) in Theorem 1.1.

4.4. Long-time asymptotics. To recover the asymptotic pulse motion, we consider the situation where t is sufficiently large so that $\|W\|_X \leq M\varepsilon^3$. In this regime we see from (4.32) that $T_1 \leq c\varepsilon^2|\log\varepsilon|$, and hence from (4.27) that

$$\|\Delta L\ W\|_{L^2} \leq c\varepsilon^2|\log\varepsilon|\|W\|_{L^2} \leq c\varepsilon^5|\log\varepsilon|. \tag{4.72}$$

Moreover, from the form (3.4) of the nonlinearity we readily verify that

$$\|\mathcal{N}_2(W)\| \leq c\|W\|_X^2 = \mathcal{O}(\varepsilon^6). \tag{4.73}$$

In this regime the estimates (4.72) and (4.73) on the secularity and the nonlinearity show that the remainder W has an asymptotically small influence on the pulse evolution equations (4.17), which reduce to

$$\dot{\Gamma}_k(t) = -\frac{(R_2, \phi_k'(\cdot;\Gamma(t)))_{L^2}}{\|\phi_k'\|_{L^2}^2} + \mathcal{O}(|\log\varepsilon|\varepsilon^5). \tag{4.74}$$

Furthermore, the asymptotic form (2.17) for the second component of the remainder shows that

$$\dot{\Gamma}_k(t) = \varepsilon^2\hat{c}(\Gamma)\frac{(\phi_1' - \phi_2', \phi_k')_{L^2}}{\|\phi_k'\|_{L^2}^2} + \mathcal{O}(\varepsilon^4) = (-1)^{k+1}\varepsilon^2\hat{c}(\Gamma) + \mathcal{O}(\varepsilon^4), \tag{4.75}$$

where $\hat{c}(\Gamma)$ is, by construction, the position-dependent formal pulse speed given by (2.8). In particular, the pulse separation $\Delta\Gamma = \Gamma_1 - \Gamma_2$ grows as given by (1.5) while the amplitudes increase according to (2.11).

REFERENCES

[1] P. W. Atkins, *Physical Chemistry*, 5th edition, W. H. Freeman and Company, New York, 1994.

[2] A. Doelman, W. Eckhaus, and T. J. Kaper, *Slowly modulated two-pulse solutions in the Gray–Scott model,* I: *Asymptotic construction and stability*, SIAM J. Appl. Math., 61 (2000), pp. 1080–1102.

[3] A. Doelman, W. Eckhaus, and T. J. Kaper, *Slowly-modulated two-pulse solutions in the Gray–Scott model,* II: *Geometric theory, bifurcations and splitting dynamics*, SIAM J. Appl. Math., 61 (2001), pp. 2036–2062.

[4] A. Doelman, S. Ei, T. Kaper, and K. Promislow, *Algebraic Stability of the Semi-Strong Interaction of Two-Pulse Solutions in the Weakly Damped Gray–Scott Equations*, in preparation.

[5] A. Doelman, R. A. Gardner, and T. J. Kaper, *Large stable pulse solutions in reaction-diffusion equations*, Indiana Univ. Math. J., 50 (2001), pp. 443–507.

[6] A. Doelman and T. J. Kaper, *Semistrong pulse interactions in a class of reaction-diffusion equations*, SIAM J. Appl. Dyn. Sys., 2, (2003), pp. 53–96.

[7] A. Doelman and H. van der Ploeg, *Homoclinic stripe patterns*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 65–104.

[8] S. Ei, *The motion of weakly interacting pulses in reaction-diffusion systems*, J. Dynam. Differential Equations, 14, (2002), pp. 85–137.

[9] S. Ei, M. Mimura, and M. Nagayama, *Pulse-pulse interactions in reaction-diffusion systems*, Phys. D, 165 (2002), pp. 176–198.

[10] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.

[11] D. Iron and M. J. Ward, *The dynamics of multispike solutions to the one-dimensional Gierer–Meinhardt model*, SIAM J. Appl. Math, 62 (2002), pp. 1924–1951.

[12] R. Moore and K. Promislow, *Renormalization group reduction of pulse dynamics in thermally loaded optical parametric oscillators*, Phys. D, 206 (2005), pp. 62–81.

[13] Y. Nishiura and D. Ueyama, *Spatio-temporal chaos for the Gray–Scott model*, Phys. D, 150 (2001), pp. 137–162.

[14] K. Promislow, *A renormalization method for modulational stability of quasi-steady patterns in dispersive systems*, SIAM J. Math. Anal., 33 (2002), pp. 1455–1482.
[15] B. Sandstede, *Stability of travelling waves*, in *Handbook of Dynamical Systems*, II, B. Fiedler, ed., North-Holland, Amsterdam, 2002, pp. 983–1055.
[16] W. Sun, M. J. Ward, and R. Russell, *The slow dynamics of two-spike solutions for the Gray–Scott and Gierer–Meinhardt systems: Competition and oscillatory instabilities*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 904–953.
[17] M. J. Ward and J. Wei, *Hopf bifurcation of spike solutions for the shadow Gierer–Meinhardt model*, European J. Appl. Math, 14 (2003), pp. 677–711.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1788–1807

DIFFUSION LIMIT OF A SEMICONDUCTOR BOLTZMANN–POISSON SYSTEM*

NADER MASMOUDI† AND MOHAMED LAZHAR TAYEB‡

Abstract. The paper deals with the diffusion limit of the initial-boundary value problem for the multidimensional semiconductor Boltzmann–Poisson system. Here, we generalize the one-dimensional results obtained in [5] to the case of several dimensions using global renormalized solutions. The method of moments and a velocity averaging lemma are used to prove the convergence of the renormalized solutions to the semiconductor Boltzmann–Poisson system towards a global weak solution of the drift-diffusion-Poisson model.

Key words. kinetic transport equations, semiconductor Boltzmann–Poisson system, drift-diffusion model, entropy dissipation, moment method, velocity averaging lemma, renormalized solution

AMS subject classifications. 82D37, 35Q40

DOI. 10.1137/050630763

1. Introduction and main results. In this paper, we study the diffusion limit of the initial-boundary value problem for the semiconductor Boltzmann–Poisson system (see [24, 27]). The model we consider here is associated with a linear low density approximation of the electron-phonon collisions. In other words it is a low density approximation of the physically correct Fermi–Dirac system. When the potential is given and is smooth enough, Poupaud [27] has proved the convergence of the rescaled Boltzmann equation towards a linear drift-diffusion model. Let us recall that the drift-diffusion equation is a standard model for semiconductors physics and suited for numerical computations since it does not involve the kinetic variable v. We refer to [9, 14, 15, 24] for a discussion about drift-diffusion models.

In the one-dimensional case, the convergence results of [27] are extended in [5] to the semiconductor Boltzmann system with a Poisson coupling. In [5] the solutions considered are defined in a weak sense [1, 4, 5]. The entropy inequality and a hybrid-Hilbert expansion are used to approximate the entropy production term due to the boundary and allow the proof of the convergence of the rescaled Boltzmann equation towards the drift-diffusion for self-consistent potential. The method is based essentially on the fact that solutions to the limit system are smooth, which gives useful uniform bounds on all terms of the Hilbert expansion and then allows one to obtain a strong convergence and also to exhibit a convergence rate. The multidimensional case is different. Indeed, if we want to work with global solutions, we can only deal with solutions to the semiconductor Boltzmann–Poisson which are defined in the renormalized sense (see [11, 25]). Indeed, due to the presence of the Poisson term and the Boltzmann collision term in the equation for the density, we cannot prove global uniform bounds in any L^p space for $p > 1$. On one hand, we can see that if we remove the collision term, then we can easily get a priori estimates for f in any

*Received by the editors May 5, 2005; accepted for publication (in revised form) August 15, 2006; published electronically March 2, 2007.
http://www.siam.org/journals/sima/38-6/63076.html
†Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012 (masmoudi@cims.nyu.edu). The author was partially supported by an NSF grant DMS 0403983.
‡Département de Mathématiques, Faculté des Sciences de Tunis, Le Campus Universitaire El Manar, 1060, Tunisia (lazhar.tayeb@fst.rnu.tn).

$L^\infty((0,T);L^p(dxdv))$, $1 \le p \le \infty$. On the other hand, if we remove the Poisson term, then we can get a priori estimates for f in any $L^\infty((0,T);L^p(dxM^{1-p}dv))$, $1 \le p < \infty$. Hence, we can see that, mathematically, these two terms cannot be treated in the same spaces. This is one of the major mathematical difficulties of this model.

Before recalling the Boltzmann–Poisson system, let us mention that drift-diffusion models can also be derived from other singular limits. We refer, for instance, to [28] where the drift-diffusion model is derived from a Vlasov–Fokker–Planck system.

1.1. Formulation of the problem. In this paper we study the parabolic limit of the rescaled Boltzmann–Poisson system. Hence, the rescaled system, defined on the phase space $\Omega = \omega \times \mathbb{R}^d$ where $d \ge 1$, reads as follows:

$$\partial_t f^\varepsilon + \frac{1}{\varepsilon}\left(v.\nabla_x f^\varepsilon - \nabla_x(\phi^\varepsilon + \tilde{\phi}_b).\nabla_v f^\varepsilon\right) - \frac{Q(f^\varepsilon)}{\varepsilon^2} = 0, \quad (x,v) \in \Omega, \tag{1}$$

where ε is a small parameter related to the mean free path and $f^\varepsilon(t,x,v)$ denotes the electron distribution function. The time variable t is nonnegative. The position x belongs to an open set ω of $\mathbb{R}^d$, assumed to be smooth and bounded and the velocity v belongs to $\mathbb{R}^d$. This equation has to be complemented with initial and boundary conditions which take into account how particles are injected in the semiconductor device. We assume that the boundary $\partial\omega$ is sufficiently smooth. We denote by $n(x)$ the outward unit normal vector at the position $x \in \partial\omega$ and $d\sigma_x$ the Lebesgue measure on $\partial\omega$. The outgoing and incoming parts are defined as

$$\Gamma^\pm = \{(x,v) \in \partial\Omega;\ \pm v.n(x) > 0\}.$$

The initial data is assumed to be known and depend on the mean free path ε:

$$f^\varepsilon(0,x,v) = f_0^\varepsilon(x,v), \qquad (x,v) \in \Omega. \tag{2}$$

The incoming boundary data is assumed to be a well-prepared function [4, 5, 27], in the sense that

$$f^\varepsilon(t,x,v) = f_b(t,x,v) := \rho_b(t,x)M(v), \qquad (x,v) \in \Gamma^-, \tag{3}$$

where M is the normalized Maxwellian

$$M(v) = \frac{e^{-|v|^2/2}}{(2\pi)^{d/2}}$$

and $\rho_b(t,x)$ is a boundary data. The precise assumptions we choose on the initial and boundary conditions will be detailed later on. The linear operator Q describes physical conservation properties during collisions. Here, we only assume that the charge is conserved during the collision [2, 24]. A typical model for such a situation is the linear approximation of the electron-phonon interaction, given by

$$Q(f)(v) = \int_{\mathbb{R}^d} \sigma(v,v')(M(v)f(v') - M(v')f(v))dv'. \tag{4}$$

The cross section σ is assumed to be symmetric (micro-reversibility principle) and bounded from above and below:

$$\begin{cases} \sigma(v,v') = \sigma(v',v), \quad (v,v') \in \mathbb{R}^{2d}, \\ \exists\ \sigma_1,\ \sigma_2 > 0 \quad / \quad 0 < \sigma_1 \le \sigma \le \sigma_2. \end{cases} \tag{5}$$

Here, we are making another approximation by assuming that σ is bounded from below and above instead of taking delta measures concentrated on balls of constant kinetic energies (see, [3, 6]).

The mean free path is defined to be an average of the collision frequency $\lambda(v)$ given by

$$\lambda(v) = \int_{\mathbb{R}^d} \sigma(v, v')M(v')dv'.$$

Here, for all v, we have $\sigma_1 \leq \lambda(v) \leq \sigma_2$. Hence, the mean free path in (1) is of order $1/\varepsilon$. We refer to [4, 5, 27] for the detailed properties of these kinds of collision kernels. We assume that the potential ϕ^ε is self consistent:

$$\left\{ \begin{array}{l} -\Delta_x \phi^\varepsilon = \displaystyle\int_{\mathbb{R}^d} f^\varepsilon dv, \\ \phi^\varepsilon{}_{|\partial\omega} = 0. \end{array} \right. \tag{6}$$

The potential $\tilde{\phi_b}$ is given in $\overline{\omega}$. It takes into account the distribution of positive background charges.

We define the charge density ρ^ε and the current density j^ε associated to the distribution f^ε by

$$\rho^\varepsilon(t,x) = \int_{\mathbb{R}^d} f^\varepsilon(t,x,v)dv, \qquad j^\varepsilon(t,x) = \frac{1}{\varepsilon}\int_{\mathbb{R}^d} v f^\varepsilon(t,x,v)dv.$$

1.2. Assumptions and preliminaries. Throughout the paper we shall make the following assumptions and notations:

A1: $f_0^\varepsilon \geq 0$, $\displaystyle\int_\Omega f_0^\varepsilon(1+|v|^2+|\log f_0^\varepsilon|) \leq C$ and $\phi^\varepsilon(t=0)$ is bounded in $H^1(\omega)$.

A2: $(\sqrt{\rho_b}, \partial_t \rho_b) \in L^\infty_{loc}(\mathbb{R}^+;\, H^{1/2}(\partial\omega) \times L^\infty(\partial\omega))$ and the density is bounded from above and below; there exist $\underline{c}$ and $\overline{c}$ such that $0 < \underline{c} \leq \rho_b(.,x) \leq \overline{c}$, for $x \in \partial\omega$.

A3: $\tilde{\phi_b} \geq 0$ and $(\tilde{\phi_b}, \partial_t \tilde{\phi_b}) \in L^\infty_{loc}(\mathbb{R}^+;\, W^{1,\infty}(\overline{\omega}) \times L^\infty(\overline{\omega}))$.

We define the total charge (or mass), the kinetic energy, and two distances to the local equilibrium by

$$\mathcal{M}^\varepsilon(t) = \int_\Omega f^\varepsilon(t,x,v)dxdv, \qquad \mathcal{K}^\varepsilon(t) = \int_\Omega \frac{|v|^2}{2} f^\varepsilon(t,x,v)dxdv,$$

$$\mathcal{R}^\varepsilon(t) = \frac{1}{2}\int_0^t \int_\Omega \left(\sqrt{f^\varepsilon} - \sqrt{\rho^\varepsilon M}\,\right)^2 dxdvds \tag{7}$$

$$\mathcal{R}_1{}^\varepsilon(t) = \frac{1}{2}\int_0^t \int_\Omega \left(f^\varepsilon - \rho^\varepsilon M\,\right)\left(\log f^\varepsilon - \log(\rho^\varepsilon M)\,\right) dxdvds.$$

The entropy and entropy fluxes through the inflow and outflow boundaries are defined by

$$\begin{aligned} \mathcal{E}^\varepsilon(t) &= \frac{1}{2}\left\|\nabla_x \phi^\varepsilon(t)\right\|^2_{L^2(\omega)} + \int_\Omega f^\varepsilon \left(\log f^\varepsilon + \frac{|v|^2}{2} + \tilde{\phi_b}\right)(t), \\ I^\pm_\varepsilon(t) &= \int_0^t \int_{\Gamma^\pm} f^\varepsilon \left(\log f^\varepsilon + \frac{|v|^2}{2} + \phi_b\right) |v.n(x)|\, d\sigma_x dvds. \end{aligned} \tag{8}$$

We also define the quasi-Fermi level (defined on the boundary $\partial\omega$)

$$\mathcal{E}_F(t,x) = \log\left(\frac{\rho_b(t,x)}{(2\pi)^{d/2}}\right) + \tilde{\phi}_b(t,x). \tag{9}$$

Let us recall two lemmas about the collision kernel (see [27]).

LEMMA 1.1 (H-theorem). *Assume that* (5) *holds, then the operator Q is bounded in $L^1(dv)$ and satisfies for all $f \in L^1(dv)$, $f \geq 0$, and $f(|\log f| + |v|^2) \in L^1(dv)$*

$$\int_{\mathbb{R}^d} Q(f) = 0 \quad \textit{and} \quad \mathcal{H}(f) = \int_{\mathbb{R}^d} Q(f)\log\left(\frac{f}{M}\right) \leq -\frac{\sigma_1}{2}\int_{\mathbb{R}^d}\left(\sqrt{f} - \sqrt{\rho\, M}\right)^2,$$

where $\rho = \int f(v)dv$. Moreover,

$$\mathcal{H}(f) = 0 \Leftrightarrow Q(f) = 0 \Leftrightarrow f(v) = \rho\, M(v).$$

LEMMA 1.2. *Assume that* (5) *holds, then*

1. *$-Q$ is a bounded, symmetric, nonnegative operator on $L^2(\mathbb{R}^d;\ M^{-1}dv)$,*
2. *$\mathcal{K}erQ = \mathbb{R}\, M$,*
3. *$-Q$ is coercive on $\mathcal{R}(Q) = \mathcal{K}erQ^{\perp}$.*

1.3. Statement of the result. Our motivation in this work is to prove the convergence of renormalized solutions $(f^\varepsilon, \phi^\varepsilon)$ to (1–6) towards $(\rho M, \phi)$, where (ρ, ϕ) satisfies the following drift-diffusion-Poisson (DD-P) system [14, 18, 27]:

$$\text{(DD-P)} \quad \begin{cases} \partial_t\rho + \nabla_x.J(\rho,\phi) = 0, \\ J(\rho,\phi) = -\mathbf{D}[\nabla_x\rho + \rho\nabla_x(\phi + \tilde{\phi}_b)], \\ \mathbf{D} = -\displaystyle\int_{\mathbb{R}^d} v \otimes\ Q^{-1}(vM)dv > 0, \\ -\Delta_x\phi = \rho, \\ \rho(t=0) = \rho_0, \quad (\rho,\phi)\,_{|\partial\omega} = (\rho_b, 0). \end{cases}$$

DEFINITION 1.3. *We say that (ρ,ϕ) is a weak solution to the drift-diffusion-Poisson system (DD-P) if*

$$\begin{aligned} &\rho \in L^\infty(0,T;\ \mathrm{LlogL}(\omega)) \cap L^2(0,T;\ L^2(\omega)), \\ &\sqrt{\rho} \in L^2(0,T;\ H^1(\omega)), \\ &\partial_t\rho \in L^1(0,T;\ W^{-1,1}(\omega)), \\ &\phi \in L^2(0,T;\ H^1_0(\omega)). \end{aligned}$$

and (ρ,ϕ) satisfies (DD-P) in the weak sense.

We recall here the definition of the space $\mathrm{LlogL}(\omega)$,

$$\mathrm{LlogL}(\omega) = \{f | f \geq 0 \,\text{and}\, \int_\omega [\, f\,(1 + |\log f|)] \text{ is finite}\}, \tag{10}$$

and that $\rho \in L^\infty(0,T;\ \mathrm{LlogL}(\omega))$ if and only if $\int_\omega \rho(t)\,(1 + |\log\rho(t)|)\,dx \leq C$, where C is independent of $t \in (0,T)$.

We also point out that due to the fact that $\partial_t\rho \in L^1(0,T;\ W^{-1,1}(\omega))$, we deduce that ρ is continuous in time with values in $W^{-1,1}(\omega)$ and hence the initial data for ρ makes sense. The main result of this paper is the following theorem.

THEOREM 1.4. *Assume that assumptions* A1, A2, *and* A3 *hold. Let* $(f^\varepsilon, \phi^\varepsilon)$ *be a renormalized solution of* (1–6) *(in the sense of Theorem* 2.2*). Then,*

$$\begin{array}{ll} f^\varepsilon \to \rho M & \text{in } L^1((0,T)\times\Omega), \\ \phi^\varepsilon \to \phi & \text{in } L^2((0,T); W^{1,p}(\omega)) \quad \forall\, p<2, \end{array} \tag{11}$$

where (ρ, ϕ) *is a weak solution of the (DD-P). Moreover,*

$$\phi \in L^\infty(0,T;\ H_0^1(\omega)) \cap L^2(0,T;\ H^2(\omega)).$$

The proof of this theorem is as follows. In section 2, we prove the existence of renormalized solutions to the semiconductor Boltzmann–Poisson system. In section 3, we establish some a priori uniform estimates. These estimates generalize the estimates obtained in [5]. To get the convergence we argue in a different manner as in [5]. Indeed, in the one-dimensional case the energy estimate of section 3 and the convergence are deduced from a hybrid-Hilbert expansion which is based on the regularity of the limiting system. In the present case, the solution to the (DD-P) is not regular enough and the solutions of the initial system are only renormalized. Instead, the method of moment and velocity averaging are used to pass to the limit ($\varepsilon \to 0$). In section 4, we use a velocity averaging lemma to prove the compactness of the charge density ρ. In section 5, we pass to the limit weakly in the equation. In section 6, we recover the boundary condition for ρ. Finally, section 7 is devoted to the proof of the regularity estimates on (ρ, ϕ) and that the limit solution (ρ, ϕ) is a weak solution of (DD-P).

2. Existence of renormalized solutions. For the existence of renormalized solutions to the full Boltzmann–Poisson system we refer to [11, 25]. It is noteworthy that even though the Boltzmann kernel we are considering here is linear, the combination of the Boltzmann term and the Poisson term makes the existence of weak solutions to (1–6) with uniform bounds a difficult problem and we were not able to construct such kinds of solutions. This is coming from the fact that the Poisson term can be well treated in $L^p(dv)$ type of spaces whereas the linear Boltzmann term can be well treated in $L^p(M^{1-p}dv)$. This incompatibility is responsible for the lack of estimates. We notice then that the entropy bound given in (15) is not enough to give a sense to $\nabla_x \phi^\varepsilon f_\varepsilon$.

Before stating an existence theorem for (1–6) let us give a definition for renormalized solution or more precisely the definition we are going to use.

DEFINITION 2.1. *We say that* $(f^\varepsilon, \phi^\varepsilon)$ *is a renormalized solution to the semiconductor Boltzmann–Poisson system if it satisfies the following:*

1. $\forall\, \beta \in \mathcal{C}^1(\mathbb{R}^+)$, $|\beta(t)| \le C(\sqrt{t}+1)$ *and* $|\beta'(t)| \le C$, $\beta(f^\varepsilon)$ *is a weak solution of*

$$\begin{cases} \varepsilon\partial_t\beta(f^\varepsilon) + v.\nabla_x\beta(f^\varepsilon) - \nabla_v.(\nabla_x(\phi^\varepsilon + \tilde{\phi}_b)\beta(f^\varepsilon)) = \beta'(f^\varepsilon)\frac{Q(f^\varepsilon)}{\varepsilon} \\ \beta(f^\varepsilon)_{|\Gamma^-} = \beta(f_b^\varepsilon), \\ \beta(f^\varepsilon)(t=0) = \beta(f_0^\varepsilon). \end{cases} \tag{12}$$

2. $\forall\, \lambda > 0$, $\theta_{\varepsilon,\lambda} = \sqrt{f^\varepsilon + \lambda M}$ *satisfies*

$$\varepsilon\partial_t\theta_{\varepsilon,\lambda} + v.\nabla_x\theta_{\varepsilon,\lambda} - \nabla_v.[\nabla_x(\phi^\varepsilon + \tilde{\phi}_b)\theta_{\varepsilon,\lambda}] = \frac{Q(f^\varepsilon)}{2\varepsilon\theta_{\varepsilon,\lambda}} + \frac{\lambda M}{2\theta_{\varepsilon,\lambda}} v.\nabla_x(\phi^\varepsilon + \tilde{\phi}_b). \tag{13}$$

THEOREM 2.2. *The semiconductor Boltzmann–Poisson system* (1–6) *has a renormalized solution in the sense of Definition* 2.1 *which satisfies, in addition,*

1. *The continuity equation*

$$\partial_t \rho^\varepsilon + \nabla_x . j^\varepsilon = 0 \tag{14}$$

2. *The entropy inequality*

$$\left[\int_\Omega f^\varepsilon \left(\tilde{\phi}_b + \frac{|v|^2}{2} + \log f^\varepsilon\right) + \frac{1}{2}\|\nabla\phi^\varepsilon\|_{L^2}^2\right]_0^t - \frac{1}{\varepsilon^2}\int_0^t\int_\Omega Q(f^\varepsilon)\log\left(\frac{f^\varepsilon}{M}\right)$$
$$\leq \int_0^t\int_\Omega \partial_t\tilde{\phi}_b f^\varepsilon - \frac{1}{\varepsilon}\int_0^t\int_{\Gamma^+\cup\Gamma^-} f^\varepsilon\left(\phi_b + \frac{|v|^2}{2} + \log f^\varepsilon\right)(v.n(x)). \tag{15}$$

Proof. For the convenience of the reader, we give an idea of the proof in Appendix A. We also refer to [25] for more details. □

3. Uniform energy estimates.

LEMMA 3.1. *Assume that assumptions* A1, A2, *and* A3 *are satisfied. Then, any renormalized solution* $(f^\varepsilon, \phi^\varepsilon)$ *of the semiconductor Boltzmann–Poisson system* (1–6) *satisfies*

$$\mathcal{M}^\varepsilon(t) + \mathcal{K}^\varepsilon(t) + \|\nabla\phi^\varepsilon(t)\|_{L^2}^2 + \frac{\mathcal{R}_1^\varepsilon(t)}{\varepsilon^2} + \int_0^t \|j^\varepsilon(s)\|_{L^1} ds \leq C_T \tag{16}$$

uniformly in ε, *where* $\mathcal{M}^\varepsilon$, $\mathcal{K}^\varepsilon$, *and* $\mathcal{R}_1^\varepsilon$ *are defined in* (7).

Proof. Starting from the entropy inequality (15), one can write the entropy dissipation in the following form:

$$\int_{\mathbb{R}^d} Q(f^\varepsilon)\log\left(\frac{f^\varepsilon}{M}\right)dv$$
$$= -\frac{1}{2}\int_{\mathbb{R}^{2d}} \sigma M M' \left(\log\frac{f^\varepsilon(v')}{M(v')} - \log\frac{f^\varepsilon(v)}{M(v)}\right)\left(\frac{f^\varepsilon(v')}{M(v')} - \frac{f^\varepsilon(v)}{M(v)}\right)dvdv'.$$

Using the Jensen inequality,

$$\int_{\mathbb{R}^d} Q(f^\varepsilon)\log\left(\frac{f^\varepsilon}{M}\right) \leq -\frac{\sigma_1}{2}\int_{\mathbb{R}^d} M\left(\log\rho^\varepsilon - \log\frac{f^\varepsilon(v)}{M(v)}\right)\left(\rho^\varepsilon - \frac{f^\varepsilon(v)}{M(v)}\right).$$

Applying the relation

$$(a-b)\log(a/b) \geq \left(\sqrt{a} - \sqrt{b}\right)^2 \tag{17}$$

we obtain

$$-\int_0^t\int_\Omega Q(f^\varepsilon)\left(\log f^\varepsilon + \frac{|v|^2}{2}\right) \geq \sigma_1\mathcal{R}_1^\varepsilon(t) \geq \sigma_1\mathcal{R}^\varepsilon(t). \tag{18}$$

Moreover, we also have to approximate the entropy production term by the boundary $(I_\varepsilon^+ - I_\varepsilon^-)(t)$, defined in (7). We write this quantity as follows:

$$(I_\varepsilon^+ - I_\varepsilon^-) = \int_0^t\int_{\Gamma^+} |v.n|\left[f^\varepsilon(v)\log\left(\frac{f^\varepsilon(v)}{f^\varepsilon(-v)}\right) + (f^\varepsilon(v) - f^\varepsilon(-v))\,\mathcal{E}_F(s,x)\right].$$

Using the inequality $a\log(a/b) \geq a - b$ for $a, b > 0$, we obtain

$$(I_\varepsilon^+ - I_\varepsilon^-)(t) \geq \int_0^t\int_{\Gamma^+} |v.n(x)|[f^\varepsilon(v) - f^\varepsilon(-v)]\,(1 + \mathcal{E}_F(t,x))d\sigma_x\,dv\,ds. \tag{19}$$

Then, we can replace (15) according to (18) and (19) and obtain

(20)
$$[\mathcal{E}^\varepsilon(t)]_0^t + \frac{\sigma_1}{\varepsilon^2}\mathcal{R}_1^\varepsilon(t) \le \int_0^t \int_\Omega \partial_t \tilde{\phi_b} f^\varepsilon - \frac{1}{\varepsilon}\int_0^t \int_{\Gamma^+} |v.n(x)|[f^\varepsilon(v) - f^\varepsilon(-v)](1+\mathcal{E}_F),$$

where

$$\mathcal{E}^\varepsilon(t) = \frac{1}{2}\left\|\nabla_x \phi^\varepsilon(t)\right\|^2_{L^2(\omega)} + \int_\Omega f^\varepsilon \left(\log f^\varepsilon + \frac{|v|^2}{2} + \tilde{\phi_b}\right)(t)$$

and

$$\mathcal{E}_F(t,x) = \log\left(\frac{\rho_b(t,x)}{(2\pi)^{d/2}}\right) + \tilde{\phi}_b(t,x).$$

We extend the quasi-Fermi level on $\bar{\omega}$ (denoted by $\tilde{\mathcal{E}_F}$) and replace ρ_b by its harmonic extension (in $\bar{\omega}$) $\tilde{\rho}_b$. According to assumptions A2 and A3, $\nabla_x\tilde{\mathcal{E}_F}$ and $\partial_t\tilde{\mathcal{E}_F}$ defined on ω by

$$\begin{aligned}
\partial_t\tilde{\mathcal{E}_F} &= (\partial_t\tilde{\rho}_b + \tilde{\rho}_b\,\partial_t\tilde{\phi_b})/\tilde{\rho}_b,\\
\nabla_x\tilde{\mathcal{E}_F} &= (\nabla_x\tilde{\rho}_b + \tilde{\rho}_b\,\nabla_x\tilde{\phi_b})/\tilde{\rho}_b
\end{aligned}$$

are bounded. By multiplying (14) by $(1+\tilde{\mathcal{E}_F}(t,x))$ and integrating by parts, we obtain

(21)
$$\begin{aligned}
\frac{1}{\varepsilon}\int_0^t \int_{\Gamma^+} (1+\mathcal{E}_F)(f^\varepsilon(v) - f^\varepsilon(-v))|v.n(x)| &= \int_0^t\int_\omega \partial_t\tilde{\mathcal{E}_F}\ \rho^\varepsilon + \int_0^t\int_\omega \nabla_x\tilde{\mathcal{E}_F}.j^\varepsilon\\
&- \left[\int_\omega (1+\tilde{\mathcal{E}_F})\rho^\varepsilon\right]_0^t
\end{aligned}$$

and then (20) is equivalent to

$$[\mathcal{E}^\varepsilon(t)]_0^t + \frac{\sigma_1}{\varepsilon^2}\mathcal{R}_1^\varepsilon(t) \le \left[\int_\omega (1+\tilde{\mathcal{E}_F})\rho^\varepsilon\right]_0^t - \int_0^t\int_\omega \nabla_x\tilde{\mathcal{E}_F}.j^\varepsilon - \int_0^t\int_\omega \frac{\partial_t\tilde{\rho}_b}{\tilde{\rho}_b}\ \rho^\varepsilon$$

which implies, according to A1, A2, and A3, that

$$\mathcal{E}^\varepsilon(t) + \frac{\mathcal{R}_1^\varepsilon(t)}{\varepsilon^2} \le C_T\left(1 + \mathcal{M}^\varepsilon(t) + \int_0^t \mathcal{M}^\varepsilon(s)ds + \int_0^t \|j^\varepsilon(s)\|_{L^1}ds\right),$$

where C_T depends only on T. Let us estimate the current density in the following way:

$$\begin{aligned}
\int_0^t \|j^\varepsilon(s)\|_{L^1}ds &= \frac{1}{\varepsilon}\int_0^t\int_\omega \left|\int_{\mathbb{R}^d} v\left(\sqrt{f^\varepsilon} - \sqrt{\rho^\varepsilon M}\right)\left(\sqrt{f^\varepsilon} + \sqrt{\rho^\varepsilon M}\right)\right|\\
&\le \frac{1}{\varepsilon}\sqrt{\mathcal{R}^\varepsilon(t)}\left(\int_0^t\int_\Omega |v|^2\left(\sqrt{f^\varepsilon} + \sqrt{\rho^\varepsilon M}\right)^2\right)^{1/2}.
\end{aligned}$$

The Young's inequality $(\alpha\, a^2 + \frac{1}{4\alpha}b^2 \ge ab \quad \forall\,\alpha > 0)$, gives

(22)
$$\int_0^t \|j^\varepsilon(s)\|_{L^1}\,ds \le \frac{\alpha}{\varepsilon^2}\mathcal{R}^\varepsilon(t) + \frac{C_T}{4\alpha}\int_0^t (\mathcal{M}^\varepsilon(s) + \mathcal{K}^\varepsilon(s))ds,$$

where α does not depend on ε (for example, $\alpha = 1/2$). Then, one can deduce that

$$\mathcal{E}^\varepsilon(t) + \frac{\mathcal{R}_1^\varepsilon(t)}{2\varepsilon^2} \le C_T \left(1 + \mathcal{M}^\varepsilon(t) + \int_0^t \mathcal{M}^\varepsilon(s)ds + \int_0^t \mathcal{K}^\varepsilon(s)ds\right) \tag{23}$$

and bound $\mathcal{M}^\varepsilon$ and $\mathcal{K}^\varepsilon$ in terms of $\mathcal{E}^\varepsilon$ using

$$\int_\Omega f^\varepsilon \log\left(\frac{f^\varepsilon}{e^{2C_T} e^{-|v|^2/4}}\right) \ge \mathcal{M}^\varepsilon - |\omega| e \int_{\mathbb{R}^d}^{e^{2C_T}} e^{-|v|^2/4}. \tag{24}$$

Hence, we deduce that

$$\mathcal{M}^\varepsilon(t) + \mathcal{K}^\varepsilon(t) + \|\nabla_x \phi^\varepsilon(t)\|_{L^2}^2 + \frac{\mathcal{R}_1^\varepsilon(t)}{\varepsilon^2} \le C_T \left(1 + \int_0^t \mathcal{M}^\varepsilon(s)ds + \int_0^t \mathcal{K}^\varepsilon(s)ds\right).$$

The Gronwall inequality leads to a uniform bound of $\mathcal{M}^\varepsilon$, $\mathcal{K}^\varepsilon$, $\varepsilon^{-2}\mathcal{R}_1^\varepsilon$, and $\|\nabla_x \phi^\varepsilon\|_{L^2}$. Then we get the L^1-bound on j^ε using (22). □

COROLLARY 3.2. *The renormalized solution satisfies*

$$\int_\Omega f^\varepsilon(1 + |v|^2 + |log f^\varepsilon|) + \int_0^t \int_{\Gamma^+} f^\varepsilon(1 + |v|^2 + |log f^\varepsilon|)|v.n(x)| \le C_T.$$

Moreover, f^ε and its trace $f^\varepsilon_{|_{\Gamma^+}}$ are weakly, relatively compact in $L^1((0,T)\times\Omega)$ and $L^1((0,T)\times\Gamma^+, |v.n(x)|dtd\sigma_x dv)$, respectively.

Proof. Let us remark that

$$\int f^\varepsilon |\log f^\varepsilon| = \int_{f^\varepsilon \ge 1} f^\varepsilon \log f^\varepsilon - \int_{f^\varepsilon \le 1} f^\varepsilon \log f^\varepsilon. \tag{25}$$

Estimates (16), (23), and (24) imply that

$$|\mathcal{E}^\varepsilon(t)| \le C_T,$$

$$\begin{aligned} \int_{f^\varepsilon \le 1} f^\varepsilon |\log f^\varepsilon| &= -\int_{f^\varepsilon \le 1} f^\varepsilon \log(f^\varepsilon / e^{-|v|^2}) + \int |v|^2 f^\varepsilon \\ &\le \int |v|^2 f^\varepsilon dxdv + |\omega| \int e^{-|v|^2} dv \le C_T, \end{aligned}$$

and

$$\int_{f^\varepsilon \ge 1} f^\varepsilon \log f^\varepsilon \le |\mathcal{E}^\varepsilon(t)| + \int_{f^\varepsilon \le 1} f^\varepsilon |\log f^\varepsilon| \le C_T.$$

The Dunford–Pettis theorem [13] implies the weak compactness of f^ε in the mentioned space. We obtain the bound and the weak compactness of $f^\varepsilon_{|_{\Gamma^+}}$ by a similar argument. Indeed, from the entropy bound we can deduce that

$$\int_0^t \int_{\Gamma^+} |v.n(x)| \left[\frac{f^\varepsilon(v)}{\rho_b M} \log\left(\frac{f^\varepsilon(v)}{\rho_b M}\right)\right] \rho_b M(v) dv d\sigma dt \le C\varepsilon \tag{26}$$

and then we can argue as previously. □

Corollary 3.2 will be used to approximate uniformly f^ε by bounded function. Indeed, let β_δ be an approximation of the identity, namely $\beta_\delta(s) = \frac{1}{\delta}\beta(\delta s)$, where β

is a $\mathcal{C}_0^\infty$ function satisfying $\beta(s) = s$ for $s \le 1$, $0 \le \beta'(s) \le 1$ for all s and $\beta(s) = 2$ for $s \ge 3$. As a consequence of the equi-integrability of f^ε, we deduce that $(\beta_\delta(f^\varepsilon))_{\delta,\varepsilon}$ is weakly relatively compact in $L^1((0,T)\times\Omega)$. Indeed, we have

$$\int_0^T \int_\Omega |\beta_\delta(f^\varepsilon) - f^\varepsilon| \le C \int_{f^\varepsilon \ge 1/\delta} f^\varepsilon \to 0 \quad \text{as } \delta \to 0$$

and

$$\int_0^T \int_{\Gamma^+} |\beta_\delta(f^\varepsilon) - f^\varepsilon|(v.n(x)) \le \int_0^T \int_{\Gamma^+ \cap \{f^\varepsilon \ge 1/\delta\}} |f^\varepsilon|(v.n(x)) \to 0 \quad \text{as } \delta \to 0$$

uniformly in ε. Up to extraction of a subsequence, we also have

$$\begin{array}{rll} \beta_\delta(f^\varepsilon) - f^\varepsilon & \to 0 & \text{a. e. as } \delta \to 0, \\ \beta'_\delta(f^\varepsilon) & \to 1 & \text{a. e. as } \delta \to 0. \end{array}$$

More precisely,

$$\sup_{\varepsilon<1} \|\beta_\delta(f^\varepsilon) - f^\varepsilon\|_{L^1_{t,x,v}} \to 0 \qquad \text{as } \delta \to 0. \tag{27}$$

PROPOSITION 3.3. *The renormalized solution $(f^\varepsilon, \phi^\varepsilon)$ satisfies the following:*

1. *ρ^ε is weakly, relatively compact in $L^1((0,T)\times\omega)$.*
2. *$\frac{Q(f^\varepsilon)}{\varepsilon}$ is weakly, relatively compact in $L^1((0,T)\times\Omega)$.*
3. *$\nabla\phi^\varepsilon$ is relatively compact in $L^2(0,T;\ L^p(\omega))$ for all $1 \le p < 2$.*

Proof. Let $\log^+ s = \max(0, \log s)$. Applying the Jensen inequality we get

$$\rho^\varepsilon \log^+ \rho^\varepsilon = \int \left(\frac{f^\varepsilon}{M} M dv\right) \left(\log^+ \int \frac{f^\varepsilon}{M} M dv\right) \le \int f^\varepsilon \log^+ \frac{f^\varepsilon}{M} dv.$$

The uniform energy bound (16) and Corollary 3.2 lead to

$$\int_0^t \int_\omega \rho^\varepsilon (1 + \log^+ \rho^\varepsilon) \le C_T$$

which implies the $L^1((0,T)\times\omega)$ weak compactness of the sequence ρ^ε. Let us define

$$r_\varepsilon := \frac{\sqrt{f^\varepsilon} - \sqrt{\rho^\varepsilon M}}{\varepsilon \sqrt{M}}.$$

Hence, from the energy bound (16), we deduce that r_ε is bounded in $L^2((0,T)\times\Omega,\ M dt dx dv)$. Extracting a subsequence if necessary, we denote by r its weak limit. Using r_ε, we can rewrite

$$f^\varepsilon = \rho^\varepsilon M + 2\varepsilon M \sqrt{\rho^\varepsilon}\, r_\varepsilon + \varepsilon^2 r_\varepsilon^2 M$$

and

$$\frac{Q(f^\varepsilon)}{\varepsilon} = 2\sqrt{\rho^\varepsilon}\, Q(r_\varepsilon M) + \varepsilon\, Q(r_\varepsilon^2 M),$$

where $r_\varepsilon^2 M$ and $r_\varepsilon M$ are, respectively, bounded in L^1 and $L^2(M^{-1}dv)$. The operator Q is bounded in L^1 and $L^2(M^{-1}dv)$. This implies that $\frac{Q(f^\varepsilon)}{\varepsilon}$ is bounded in $L^1_{t,x,v}$ and

$$\frac{Q(f^\varepsilon)}{\varepsilon} = 2\sqrt{\rho^\varepsilon}\, Q(r_\varepsilon M) + \mathbf{O}(\varepsilon)_{L^1(0,T)\times\Omega}.$$

Moreover, let $\alpha > 0$, then

$$\int_A |\sqrt{\rho^\varepsilon}\, Q(r_\varepsilon M)| \le C \int_A \rho^\varepsilon M + \frac{1}{4C} \int_A \frac{Q^2(r_\varepsilon M)}{M}$$

and choose C such that $\frac{1}{4C} \| \frac{Q^2(r_\varepsilon M)}{M} \|_{L^1_{t,x,v}} \le \alpha/2$. For such fixed α and C the equi-integrability of $\rho^\varepsilon M$ implies

$$\exists\, \delta > 0, \forall\, A \subset (0,T) \times \Omega, \quad |A| < \delta \Rightarrow C \int_A \rho^\varepsilon M \le \alpha/2.$$

So, $\sqrt{\rho^\varepsilon}\, Q(r_\varepsilon M)$ is equi-integrable. Besides, if $\Omega_R := \omega \times B(0,R)$, then

$$\int_0^T \int_{\Omega_R^c} |\sqrt{\rho^\varepsilon}\, Q(r_\varepsilon M)| \le \|\rho^\varepsilon\|^{1/2}_{L^1_{t,x}} \left(\int_{B(O,R)^c} M dv \right)^{1/2} \left(\int_0^T \int_{\Omega_R^c} Q(r_\varepsilon M)^2 M^{-1} \right)^{1/2} \to 0$$

as R goes to infinity uniformly in ε. This proves the weak compactness of $\frac{Q(f^\varepsilon)}{\varepsilon}$ in $L^1((0,T) \times \Omega)$.

The third assertion of the proposition is a consequence of the Lions–Aubin theorem (see [22, Lemma 5.1]). Indeed, using that $\nabla_x \phi^\varepsilon$ is bounded in $L^2(0,T;\, L^2(\omega))$ and $\Delta_x \phi^\varepsilon$ is bounded in $L^2(0,T,\, L^1(\omega))$ we deduce that $\nabla_x \phi^\varepsilon$ is bounded in $L^2(0,T;\, L^2(\omega) \cap W^{r,q}(\omega))$ for some $q > 1$ and $0 < r < 1$ such that $\frac{1-r}{d} > 1 - \frac{1}{q}$. And, using that $\partial_t \nabla_x \phi^\varepsilon = \nabla_x (\Delta_x)^{-1} \nabla_x . (j^\varepsilon)$ we see that $\partial_t \nabla_x \phi^\varepsilon$ is bounded in $L^1(W^{-s,p}_{loc}(\omega))$ for some $p > 1$ and $s > d - \frac{d}{p}$. Hence, $\nabla_x \phi^\varepsilon$ is compact in $L^2(0,T;\ L^q_{loc}(\omega))$ for some $1 < q < 2$. And using that $\nabla_x \phi^\varepsilon$ is bounded in $L^2(0,T;\ L^2(\omega))$ we deduce the compactness in $L^2(0,T;\ L^p(\omega))$ for all $p < 2$. □

PROPOSITION 3.4. *r_ε is such that $\varepsilon |r_\varepsilon|^2 |v|^2 M$ is bounded in $L^1((0,T) \times \Omega)$ and $\sqrt{\varepsilon} |r_\varepsilon|^2 |v| M$ is bounded in $L^1((0,T) \times \Omega)$.*

Proof. The proof uses Young's inequality (see [7] and [23] where a similar argument is used to control the distance to the Maxwellian). Let us denote $r(z) = z \log(1+z)$ and

$$r^*(p) = \sup_{z > -1} (pz - r(z))$$

its Legendre transform. Hence $r^*(p)$ behaves like e^p when p goes to $+\infty$. Moreover, $r^*(p)$ has a superquadratic homogeneity, namely for $0 < \alpha < 1$ and $p > 0$, we have $r^*(\alpha p) \le \alpha^2 r^*(p)$. We also denote $z_\varepsilon = \frac{f^\varepsilon}{\rho_\varepsilon M} - 1$ and $z_\varepsilon = 0$ if $\rho_\varepsilon = 0$. Hence

$$\varepsilon |r_\varepsilon|^2 |v|^2 \le \frac{1}{\varepsilon} \rho_\varepsilon |z_\varepsilon| |v|^2. \tag{28}$$

By the Young inequality, we have

$$\begin{aligned} \frac{1}{\varepsilon} \rho_\varepsilon |v|^2 |z_\varepsilon| &\le \frac{4\rho_\varepsilon}{\varepsilon^2} \left[r^* \left(\frac{\varepsilon}{4} |v|^2 \right) + r(|z_\varepsilon|) \right] \\ &\le \frac{4\rho_\varepsilon}{\varepsilon^2} \left[\varepsilon^2 r^* \left(\frac{|v|^2}{4} \right) + r(z_\varepsilon) \right] \end{aligned}$$

which is clearly bounded in $L^1((0,T) \times \Omega, dt dx M dv)$ by using the growth of r^* and the entropy dissipation bound $\mathcal{R}_1^\varepsilon(t) \le C\varepsilon^2$. This proves the first assertion. Interpolating with the fact that r_ε is bounded in $L^2((0,T) \times \Omega, dt dx M dv)$, we deduce the second bound. This ends the proof of the proposition. □

4. Compactness by velocity averaging.

PROPOSITION 4.1. *The density ρ^ε is relatively compact in $L^1((0,T)\times\omega)$: there exists $\rho\in L^1((0,T)\times\omega)$ such that, up to extraction of a subsequence if necessary,*

$$\rho^\varepsilon\to\rho\quad \text{in } L^1 \text{and a.e.}$$

Using (27), it suffices to show the compactness of $(\beta_\delta(f^\varepsilon))_\varepsilon$ for all (fixed) $\delta>0$. This is a consequence of the averaging lemma (see [17, 12]) and the continuity equation.

Let us recall the following averaging lemma.

LEMMA 4.2. *Assume that h^ε is bounded in $L^2((0,T)\times\Omega)$, that h_0^ε and h_1^ε are bounded in $L^1((0,T)\times\Omega)$, and that*

$$\varepsilon\,\partial_t h^\varepsilon + v.\nabla_x h^\varepsilon = h_0^\varepsilon + \nabla_v . h_1^\varepsilon. \tag{29}$$

Then, for all $\psi\in\mathcal{C}_0^\infty(\mathbb{R}^d)$,

$$\left\|\int_{\mathbb{R}^d}(h^\varepsilon(t,x+y,v)-h^\varepsilon(t,x,v))\,\psi(v)dv\right\|_{L^1_{t,x}}\to 0 \text{ when } y\to 0 \text{ uniformly in } \varepsilon \tag{30}$$

where $h^\varepsilon(t,x,v)$ has been prolonged by 0 for $x\notin\omega$.

Remark 4.1. The above lemma only gives the compactness in the x variable of the averages in v of $h^\varepsilon(t,x,v)$. This is due to the presence of an ε in front of the time derivative in (29). We also refer to [18] and [7] for similar averaging lemmas where there is only gain of regularity in the x variable.

This lemma can be deduced from the proof of Theorem 1.8 of [8] or from the proof of Theorem 6 of [12] (see also the proof of Theorem 3 of [12]). The only difference here is the presence of the time derivative which comes with the factor ε in front and hence does not imply regularity in time as in Theorem 6 of [12]. Actually, following the proof of Theorem 3 of [12] with $q=m=1$, $p=2$, $\tau=0$, and writing the problem in the whole space $\mathbb{R}_t\times\mathbb{R}^d_x$, we can prove that $\int_{\mathbb{R}^d}\psi(v)h^\varepsilon(t,x,v)dv$ is in the Besov space $L^{r,\infty}((0,T);B^{s,r}_{\infty,\infty})$, where $r=\frac{5}{3}$ and $s=\frac{1}{5}$. For the definition of the Besov space $B^{s,r}_{\infty,\infty}$, we refer to [12]. A sketch of the proof will be given in Appendix B. This, of course, yields the compactness stated in (30).

Proof of Proposition 4.1. Let δ be a (fixed) nonnegative parameter. Let us verify that the rescaled Boltzmann equation (in the renormalized sense) satisfies the assumptions of Lemma 4.2. Indeed, $\beta_\delta(f^\varepsilon)$ is a weak solution of

$$\varepsilon\,\partial_t\beta_\delta(f^\varepsilon) + v.\nabla_x\beta_\delta(f^\varepsilon) = h_0^\varepsilon + \nabla_v.h_1^\varepsilon,$$

where

$$h_0^\varepsilon = \frac{Q(f^\varepsilon)}{\varepsilon}\beta_\delta'(f^\varepsilon)\qquad\text{and}\qquad h_1^\varepsilon = \nabla(\phi^\varepsilon+\tilde{\phi_b})\,\beta_\delta(f^\varepsilon).$$

The sequences $(\beta_\delta(f^\varepsilon))_\varepsilon$ is bounded in $L^\infty\cap L^1((0,T)\times\Omega)$ and hence in $L^2((0,T)\times\Omega)$. Moreover, h_0^ε is weakly, relatively compact in $L^1_{t,x,v}$ and by applying Holder's inequality and using the uniform bound of $\beta_\delta(f^\varepsilon)$ in L^2 (for fixed δ), we obtain

$$\|h_1^\varepsilon\|_{L^1((0,T)\times\omega;L^2(\mathbb{R}^d))}\le\frac{C}{\sqrt{\delta}}\sup_{t\le T}\|\nabla(\phi^\varepsilon+\tilde{\phi_b})(t)\|_{L^2_x}.$$

Since we are using a compactly supported function to localize in space, the L^2 bound in v also yields an L^1 bound.

Assumptions of Lemma 4.2 are satisfied and hence we get the L^1-compactness in x of $\int_{\mathbb{R}^d} \psi(v)\beta_\delta(f^\varepsilon)dv$, namely (30) holds with h^ε replaced by $\beta_\delta(f^\varepsilon)$.

Next, using that $(\beta_\delta(f^\varepsilon))_\varepsilon$ is bounded in $L^\infty(0,T;\, L^1((1+|v|^2)dxdv))$, we see that we can take $\psi(v)$ to be constant equal to 1 in (30) and hence we deduce that, after also sending δ to 0 and using the equi-integrability of f^ε, that

$$\|\rho^\varepsilon(t,x+y)-\rho^\varepsilon(t,x)\|_{L^1_{t,x}} \to 0 \text{ when } y \to 0 \text{ uniformly in } \varepsilon.$$

Finally, using that $\partial_t\rho^\varepsilon = -\nabla_x .j^\varepsilon$ is bounded in $L^1((0,T);W^{-1,1}(w))$, we deduce that ρ^ε is relatively compact in $L^1((0,T)\times\omega)$ and Proposition 4.1 is proved. □

5. Passing to the limit. Using the previous section, there exists $\rho \in L^1((0,T)\times\omega)$ such that

$$\rho^\varepsilon \to \rho \quad \text{in } L^1_{t,x} \quad \text{and a. e.}$$

The inequality $\left(\sqrt{a}-\sqrt{b}\right)^2 \le |a-b|$ leads to

$$\sqrt{\rho^\varepsilon} \to \sqrt{\rho} \quad \text{in } L^2_{t,x} \quad \text{and a. e.}$$

The entropy dissipation given by (15) leads to

$$f^\varepsilon \to \rho\, M \quad \text{in } L^1_{t,x,v} \quad \text{and a. e.} \tag{31}$$

Moreover, we have

$$\frac{Q(f^\varepsilon)}{\varepsilon} = \left(2\sqrt{\rho^\varepsilon}\, Q(r_\varepsilon\, M) + \varepsilon\, Q(r_\varepsilon^2 M)\right) \quad \rightharpoonup \quad 2\sqrt{\rho}\, Q(r\, M) \quad \text{in } L^1,$$

where r is the weak limit of r^ε in $L^2((0,T)\times\Omega,\, M(v)dtdxdv)$. So, one can pass to the limit in (13) for $\lambda > 0$, up to extraction of a subsequence, and get

$$\begin{aligned} v.\nabla_x\sqrt{(\rho+\lambda)M} - \nabla_v.(\nabla_x(\phi+\tilde{\phi_b})\sqrt{(\rho+\lambda)M}) \;&=\; \frac{\sqrt{\rho}\, Q(r\, M)}{\sqrt{(\rho+\lambda)M}} \\ &\quad + \frac{\lambda M\, v.\nabla_x(\phi+\tilde{\phi_b})}{2\sqrt{(\rho+\lambda)M}}, \end{aligned} \tag{32}$$

where $\nabla_x\phi$ is the $\mathrm{L}^2_{t,x}$-weak limit of $\nabla_x\phi^\varepsilon$. Sending λ to 0, we infer that

$$\left(\nabla_x\sqrt{\rho} + \frac{1}{2}\sqrt{\rho}\,\nabla_x(\phi+\tilde{\phi_b})\right).vM = Q(r\, M).$$

Using that $Q(r\, M)$ is bounded in $L^2((0,T)\times\Omega,\, M^{-1}(v)dtdxdv)$, we deduce that $\nabla_x\sqrt{\rho} + \frac{1}{2}\sqrt{\rho}\,\nabla_x(\phi+\tilde{\phi_b})$ is bounded in $L^2((0,T)\times\omega)$. Besides, the current density is given by

$$j^\varepsilon = 2\sqrt{\rho^\varepsilon}\int_{\mathbb{R}^d} r^\varepsilon\, vMdv + \varepsilon\int r_\varepsilon^2 vMdv.$$

Using Proposition 3.4, we deduce that

$$j^\varepsilon = 2\sqrt{\rho^\varepsilon}\int_{\mathbb{R}^d} r^\varepsilon\, vMdv + \mathbf{O}(\sqrt{\varepsilon})_{L^1((0,T)\times\Omega)} \rightharpoonup 2\sqrt{\rho}\int_{\mathbb{R}^d} r\, vMdv \quad \text{in } L^1_{t,x}.$$

The function $vM \in \mathcal{R}(Q) := \mathcal{K}erQ^{\perp}$, therefore

$$\int r\, vM\, dv = \int Q(r\, M)Q^{-1}(vM)\frac{dv}{M}$$

and

$$j^{\varepsilon} \rightharpoonup J(\rho,\phi) := 2\sqrt{\rho}\left[\int_{\mathbb{R}^d}(v\otimes Q^{-1}(vM))dv\right]\left(\nabla_x\sqrt{\rho}+\frac{1}{2}\sqrt{\rho}\,\nabla_x(\phi+\tilde{\phi}_b)\right).$$

Passing to the limit ($\varepsilon \to 0$) in (14) and the Poisson equation $-\Delta\phi^{\varepsilon} = \rho^{\varepsilon}$, we obtain

$$\begin{cases} \partial_t\rho + \nabla_x J(\rho,\phi) = 0, \\ J(\rho,\phi) = 2\sqrt{\rho}\left[-\mathbf{D}\left(\nabla_x\sqrt{\rho}+\frac{1}{2}\sqrt{\rho}\,\nabla_x(\phi+\tilde{\phi}_b)\right)\right], \\ \mathbf{D} = -\displaystyle\int_{\mathbb{R}^d}(v\otimes Q^{-1}(vM))dv \quad \text{and} \quad -\Delta_x\phi = \rho. \end{cases}$$

6. The limit boundary condition. In this section, we want to pass to the limit in the boundary condition and prove that $\rho = \rho_b$ on $\partial\omega$. First notice that from the fact that $\nabla_x\sqrt{\rho}+\frac{1}{2}\sqrt{\rho}\,\nabla_x(\phi+\tilde{\phi}_b)$ is bounded in $L^2((0,T)\times\omega)$, we deduce that $\nabla_x\sqrt{\rho}$ is bounded in $L^1((0,T)\times\omega)$ and hence the trace of $\sqrt{\rho}$ makes sense on $\partial\omega$.

For each sequence $(g_\varepsilon)_\varepsilon$, $\overline{g_\varepsilon}$ will denote the weak limit of $(g_\varepsilon)_\varepsilon$ when ε goes to zero, extracting a subsequence if necessary. In particular, $\overline{f^{\varepsilon}_{|\Gamma}}$ denotes the weak limit of $f^{\varepsilon}_{|\Gamma}$ in $L^1((0,T)\times\Gamma^+, |v.n(x)|dtd\sigma_x dv)$. We recall that $\beta_\delta(f^{\varepsilon})$ is a weak solution of the renormalized semiconductor Boltzmann equation

$$\varepsilon\partial_t\beta_\delta(f^{\varepsilon}) + v\,.\nabla_x\beta_\delta(f^{\varepsilon}) - \nabla_x(\phi^{\varepsilon}+\tilde{\phi}_b).\nabla_v\beta_\delta(f^{\varepsilon}) = \frac{Q(f^{\varepsilon})}{\varepsilon}\beta'_\delta(f^{\varepsilon}) \tag{33}$$

with the following boundary condition and initial data

$$\beta_\delta(f^{\varepsilon})_{|\Gamma^-} = \beta_\delta(\rho_b M) \quad \text{and} \quad \beta_\delta(f^{\varepsilon})_{|t=0} = \beta_\delta(f^{\varepsilon}_0).$$

Passing to the limit in (33), we infer that

$$v\,.\nabla_x\beta_\delta(\rho M) - \nabla_x(\phi+\tilde{\phi}_b).\nabla_v\beta_\delta(\rho M) = 2\sqrt{\rho}\,Q(r\,M)\beta'_\delta(\rho M). \tag{34}$$

On one hand, by using $\xi(t,x,v) \in \mathcal{C}^\infty([0,T]\times\bar{\Omega})$ as a test function in (34) we get

$$-\int_0^T\int_\Omega \beta_\delta(\rho M)v.\nabla_x\xi + \int_0^T\int_\Omega \beta_\delta(\rho M)\nabla_x(\phi+\tilde{\phi}_b).\nabla_v\xi$$

$$-\int_0^T\int_\Omega 2\sqrt{\rho}\,Q(r\,M)\beta'_\delta(\rho M)\xi + \int_0^T\int_{\partial\Omega}\xi\beta_\delta(\rho M)(v.n(x)) = 0.$$

On the other hand, using $\xi(t,x,v)$ as a test function in (33) and passing to the limit, we deduce that

$$\beta_\delta(\rho M)_{|\partial\Omega} = \overline{\beta_\delta(f^{\varepsilon})_{|\partial\Omega}}. \tag{35}$$

From Corollary 3.2, we deduce that $f^{\varepsilon}_{|\partial\Omega} \in L^\infty(0,T;\ \mathrm{LlogL}(|v.n(x)|d\sigma_x dv))$ and hence $\beta_\delta(f^{\varepsilon})_{|\partial\Omega} = \beta_\delta(f^{\varepsilon}_{|\partial\Omega}) \in L^\infty(0,T;\ \mathrm{LlogL}(|v.n(x)|d\sigma_x dv))$ uniformly in ε,δ. Hence,

$\beta_\delta(\rho M)_{|\partial\Omega}$ is uniformly bounded in $L^\infty(0,T;\ \mathrm{LlogL}(|v.n(x)|d\sigma_x dv))$ and converges to $(\rho M)_{|\partial\Omega}$ when δ goes to 0.

Using $\psi(t,x) \in \mathcal{C}^\infty([0,T]\times\bar{\omega})$ as a test function in (33) and passing to the limit, we get

$$\int_0^T \int_\Omega \beta_\delta(\rho M) v.\nabla_x \psi + \int_0^T \int_\Omega 2\sqrt{\rho}\, Q(r\, M)\beta_\delta'(\rho M)\psi = \int_0^T \int_{\Gamma^+} \psi \left(\overline{\beta_\delta(f^\varepsilon)_{|\Gamma^+}} - \beta_\delta(\rho_b M) \right) (v.n(x)).$$

Sending δ to 0 and using that

$$\int_0^T \int_\Omega \rho M v.\nabla_x \psi + \int_0^T \int_\Omega 2\sqrt{\rho}\, Q(r\, M)\psi = 0,$$

we deduce that

$$\lim_{\delta\to 0} \int_0^T \int_{\Gamma^+} \left[\overline{\beta_\delta(f^\varepsilon)_{|\Gamma^+}} - \beta_\delta(\rho_b M) \right] (v.n(x))\psi = 0. \tag{36}$$

Using (35) and the fact that

$$\int_{\mathbb{R}^d \cap \{v.n(x)\geq 0\}} M(v.n(x))dv = \frac{1}{2\pi}$$

we infer that

$$\int_0^T \int_{\partial\omega} [\rho_{|\partial\omega} - \rho_b]\psi = 0 \tag{37}$$

and hence $\rho = \rho_b$ on $\partial\omega$.

7. Regularity of the density. In this section we shall prove that the limit $\rho \in L^\infty(0,T;\ L^2(\omega))$ and that $\sqrt{\rho} \in L^2(0,T;\ H^1(\omega))$.

We assume that $\rho_b(t,x)$ is defined in the whole domain ω. We can, for instance, extend $\sqrt{\rho_b(t,x)}$ in ω as a harmonic function. Hence, $\rho_b(t,x)$ satisfies

$$\rho_b \in L^\infty((0,T)\times\omega) \quad \text{and} \quad \sqrt{\rho_b} \in L^2(0,T;\ H^1(\omega)).$$

LEMMA 7.1. *Let ω be a regular and bounded open subset of $\mathbb{R}^d$ and ρ a positive function of $L^\infty(0,T;\ L^1(\omega))$ satisfying*

$$\begin{cases} \nabla_x\sqrt{\rho} + \frac{1}{2}\nabla_x\phi\sqrt{\rho} = G \in L^2(0,T;\ L^2(\omega)), \\ -\Delta_x\phi = \rho, \\ \nabla_x\phi \in L^\infty(0,T;\ L^2), \\ \rho = \rho_b \quad \textit{on } \partial\omega. \end{cases} \tag{38}$$

Then

$$\rho \in L^2(0,T;\ L^2(\omega)), \qquad \sqrt{\rho} \in L^2(0,T;\ H^1(\omega)),$$

and

$$\nabla\phi\sqrt{\rho} \in L^2(0,T;\ L^2(\omega)).$$

Proof. The first and third equations of (38) imply that $\nabla_x\sqrt{\rho} \in L^1$ and hence the boundary condition $\rho = \rho_b$ on $\partial\omega$ makes sense. Let us take β_δ as in the proof of Corollary 3.2. But since we will deal with possibly negative values, we take $\beta_\delta(s) = \frac{1}{\delta}\beta(\delta s)$, where β is a C^∞ function satisfying $\beta(s) = s$ for $-1 \le s \le 1$, $0 \le \beta'(s) \le 1$ for all s and $\beta(s) = 2$ for $|s| \ge 3$.

We denote $\psi = \sqrt{\rho} - \sqrt{\rho_b}$. Hence

$$\nabla_x\beta_\delta(\psi) = \nabla_x\psi\,\beta'_\delta(\psi).$$

Hence, after subtracting $\sqrt{\rho_b}$ from $\sqrt{\rho}$, we can renormalize the first equation appearing in (38), it gives

$$\nabla_x\beta_\delta(\psi) + \frac{1}{2}\nabla_x\phi\beta'_\delta(\psi)\psi = \tilde{G}\beta'_\delta(\psi), \tag{39}$$

where $\tilde{G} = G - \nabla_x\sqrt{\rho_b} - \frac{1}{2}\nabla_x\phi\sqrt{\rho_b}$ is also in $L^2(0,T;\,L^2(\omega))$. Then using that for fixed $\delta > 0$

$$|\nabla_x\phi\beta'_\delta(\psi)\psi| \le \frac{1}{2\delta}|\nabla\phi| \in L^2,$$

we deduce that $\nabla_x\beta_\delta(\psi) \in L^2$ for fixed δ. Taking the L^2 norm of (39), we get

$$\|\nabla_x\beta_\delta(\psi)\|^2_{L^2} + \frac{1}{4}\|\nabla_x\phi\beta'_\delta(\psi)\psi\|^2_{L^2} + \int_0^T\int_\omega \nabla_x\phi\nabla_x\beta_\delta(\psi)\beta'_\delta(\psi)\psi \le \|\tilde{G}\|^2_{L^2}. \tag{40}$$

Let $\tilde{\beta}$ be given by $\tilde{\beta}(s) = \int_0^s \tau\beta'(\tau)^2 d\tau$ and $\tilde{\beta}_\delta(s) = \frac{1}{\delta^2}\tilde{\beta}(\delta s)$. Hence, $\tilde{\beta}_\delta(s)$ goes to $\frac{s^2}{2}$ when δ goes to 0.

Computing the third term in (40), we get

$$\int_\omega \nabla_x\phi.\nabla_x\beta_\delta(\psi)\beta'_\delta(\psi)\psi \quad = \quad \int_\omega \nabla_x\phi.\nabla_x\tilde{\beta}_\delta(\psi) = \int_\omega \rho\tilde{\beta}_\delta(\psi).$$

Hence, we deduce that for all $\delta > 0$,

$$\|\nabla_x\beta_\delta(\psi)\|^2_{L^2} + \frac{1}{4}\|\nabla_x\phi\beta'_\delta(\psi)\psi\|^2_{L^2} + \int_0^T\int_\omega \rho\tilde{\beta}_\delta(\psi) \le \|\tilde{G}\|^2_{L^2}.$$

Letting δ go to zero, we get that

$$\|\nabla_x(\sqrt{\rho} - \sqrt{\rho_b})\|^2_{L^2} + \frac{1}{4}\|\nabla_x\phi(\sqrt{\rho} - \sqrt{\rho_b})\,\|^2_{L^2} + \frac{1}{2}\int_0^T\int_\omega \rho(\sqrt{\rho} - \sqrt{\rho_b})^2 \le \|\tilde{G}\|^2_{L^2}.$$

Using that $\sqrt{\rho_b}$ is bounded, we conclude the proof of the lemma. ∎

Now, using the lemma, we can see easily that we can rewrite the current

$$J(\rho,\phi) = 2\sqrt{\rho}\left[-\mathbf{D}\left(\nabla_x\sqrt{\rho} + \frac{1}{2}\sqrt{\rho}\,\nabla_x(\phi + \tilde{\phi_b})\right)\right] = -\mathbf{D}[\nabla_x\rho + \rho\nabla_x(\phi + \tilde{\phi_b})].$$

Finally, the regularity of ϕ can be easily deduced from that of ρ and this ends the proof of Theorem 1.4.

Appendix A. Existence of renormalized solution. We present here a proof of the existence of renormalized solution to (1–6) satisfying the conditions of Theorem

2.2. We refer to [25] for the existence of renormalized solution to the Vlasov–Poisson–Boltzmann system with a nonlinear Boltzmann kernel.

To simplify the notations we take $\varepsilon = 1$. We begin by regularizing the collision operator and both Boltzmann and Poisson equations. Let us define

$$Q_R(f) = \int_{\mathbb{R}^d} \sigma_R(v, v')(Mf' - M'f)dv', \tag{41}$$

where

$$\sigma_R(v, v') = \sigma(v, v')1_{|v|\le R}(v)1_{|v'|\le R}(v'). \tag{42}$$

The regularized semiconductor Boltzmann–Poisson system reads

$$(VBP)_{\alpha,R} \qquad \begin{cases} \partial_t f_{\alpha,R} + \left(v.\nabla_x f_{\alpha,R} - \nabla_x(\phi_{\alpha,R} + \tilde{\phi}_b).\nabla_v f_{\alpha,R}\right) = Q_R(f_{\alpha,R}), \\ -(1 - \alpha\,\Delta_x)^{2m}\Delta_x\phi_{\alpha,R} = \rho_{\alpha,R} = \displaystyle\int_{|v|\le R} f_{\alpha,R}dv, \\ f_{\alpha,R}(0, x, v) = f_0(x, v), \qquad (x, v) \in \Omega, \\ f_{\alpha,R}(t, x, v) = f_b(t, x, v), \qquad (x, v) \in \Gamma^-, \\ \phi_{\alpha,R} = \Delta_x\phi_{\alpha,R} = \cdots = \Delta_x^{2m}\phi_{\alpha,R} = 0, \quad x \in \partial\omega, \end{cases}$$

where α is a nonnegative parameter and $m \in \mathbb{N}^*$. We refer to [1, 5, 29] for further details about this approximation.

A simple computation gives the following H-theorem:

LEMMA A.1. *The collision operator Q_R is bounded in L^1 and L^∞ and satisfies*

$$\int_{|v|<R} Q_R(f)dv = \int_{\mathbb{R}^d} Q_R(f)dv = 0$$

and

$$\int_{|v|<R} Q_R(f)\log\frac{f}{M} = \int_{\mathbb{R}^d} Q_R(f)log\frac{f}{M} \le -\frac{\sigma_1}{2}\int\left(\sqrt{f} - \sqrt{M\int f dv}\right)^2.$$

As a consequence of these conservation properties, one can prove, by a fixed point procedure and by using the characteristic method, that the modified system $(VBP)_{\alpha,R}$ has a weak solution $(f_{\alpha,R}, \phi_{\alpha,R})$. More precisely, multiplying the semiconductor Boltzmann equation by $(1 + |v|^2/2 + \log f_{\alpha,R})$, and integrating with respect to $dtdxdv$, we get

$$\begin{aligned} &\left[\frac{1}{2}\int_\omega |(1-\alpha\Delta)^m\nabla_x\phi_{\alpha,R}|^2(s) + \int_\Omega f_{\alpha,R}\left(\log f_{\alpha,R} + \frac{|v|^2}{2} + \tilde{\phi}_b\right)(s)\right]_0^t \\ &\int_0^t\int_{\Gamma^+} f_{\alpha,R}\left(\log f_{\alpha,R} + \frac{|v|^2}{2} + \tilde{\phi}_b\right)|v.n(x)| + \frac{\sigma_1}{2}\int_0^t\int_\Omega\left(\sqrt{f_{\alpha,R}} - \sqrt{\rho_{\alpha,R}M}\right)^2 \\ &\le \int_0^t\int_{\Gamma^-} f_b\left(\log f_b + \frac{|v|^2}{2} + \phi_b\right)|v.n(x)|. \end{aligned}$$

For fixed α, the solution is weak. One can pass to the limit $R \to \infty$ and show that

there exists a weak solution $(f_\alpha,\ \rho_\alpha,\ \phi_\alpha)$ of

$$(VBP)_\alpha \qquad \begin{cases} \partial_t f_\alpha + v.\nabla_x f_\alpha - \nabla_x(\phi_\alpha + \tilde{\phi}_b).\nabla_v f_\alpha = Q(f_\alpha), \\ -(1-\alpha\Delta_x)^{2m}\Delta_x\phi_\alpha = \rho_\alpha = \displaystyle\int_{\mathbb{R}^d} f_\alpha dv, \\ f_\alpha(0,x,v) = f_0(x,v), \qquad (x,v)\in\Omega, \\ f_\alpha(t,x,v) = f_b(t,x,v) \qquad (x,v)\in\Gamma^-, \\ \phi_\alpha = \Delta_x\phi_\alpha = \cdots = \Delta_x^{2m}\phi_\alpha = 0, \qquad x\in\partial\omega. \end{cases}$$

Moreover, this weak solution satisfies (12), (13), and (14) and

$$\left[\frac{1}{2}\int_\omega |(1-\alpha\Delta)^m\nabla_x\phi_\alpha|^2(t) + \int_\Omega f_\alpha\left(\log f_\alpha + \frac{|v|^2}{2} + \tilde{\phi}_b\right)(s)\right]_0^t$$
$$\int_0^t\int_{\Gamma^+} f_\alpha\left(\log f_\alpha + \frac{|v|^2}{2} + \phi_b\right)|v.n(x)| + \frac{\sigma_1}{2}\int_0^t\int_\Omega\left(\sqrt{f_\alpha} - \sqrt{\rho_\alpha M}\right)^2$$
$$\le \int_0^t\int_{\Gamma^-} f_b\left(\log f_b + \frac{|v|^2}{2} + \phi_b\right)|v.n(x)| \le C_T.$$

As a consequence of this identity we get the following proposition.

PROPOSITION A.2.

1. *f_α and $f_{\alpha|_{\Gamma^+}}$ are, respectively, weakly, relatively compact in $L^1((0,T)\times\Omega)$ and $L^1((0,T)\times\Gamma^+, |v.n(x)|dtd\sigma_x dv)$.*
2. *$\|\nabla_x\phi_\alpha\|_{L^2} \le \|(1-\Delta_x)^m\nabla_x\phi_\alpha\|_{L^2} \le C_T$ and $\nabla_x\phi_\alpha$ is relatively compact in $L^p((0,T)\times\omega)$ for all $p<2$.*
3. *ρ_α is relatively compact in $L^1((0,T)\times\omega)$.*

The proof of this proposition follows the same lines as the proofs of compactness given in the paper (see the proofs of Corollary 3.2 and Propositions 3.3 and 4.1). Notice, however, that we do not immediately get the compactness of f_α as in (31).

Using this proposition, we can end the proof of Theorem 2.2.

Let $f,\ \rho,\ \phi$, and j be the weak limits of subsequences of $f_\alpha,\ \rho_\alpha,\ \phi_\alpha$, and $j^\alpha = \int v f_\alpha dv$. To prove that (f,ϕ) satisfies (12) and (13), we argue as in [21] (see also [25]). The method is based on a double renormalization. First, we write the equation satisfied by $\beta_\delta(f_\alpha)$, where β_δ was defined in section 3 and then weakly pass to the limit when α goes to zero. Then, we renormalize the resulting limit equation using the function β or the function $\sqrt{s+\lambda M}$. Finally, we let α go to zero and recover (12) and (13). The continuity equation (14) can be easily deduced from the continuity equation for ρ_α and the entropy inequality (15) can be deduced from a classical convexity argument (see also [25]). This ends the proof of Theorem 2.2.

Appendix B. Sketch of the proof of Lemma 4.2. Here we would like to prove Lemma 4.2.

Step 1: First, we rewrite the problem in the whole space in t and x. This step only uses the equi-integrability of h_ε. Indeed, for α small enough, we can find C^∞ cut-off functions $\chi_1(t)$ and $\chi_2(x)$ such that $\chi_1 = 1$ on $(\alpha, T-\alpha)$ has compact support in $(0,T)$, and $\chi_2 = 1$ on $\{x\in\omega, |, \mathrm{dist}(x,\partial\omega) > \alpha$ and has compact support in ω. Denoting $\chi(t,x) = \chi_1(t)\chi_2(x)$ and $\tilde{h}^\varepsilon = \chi(t,x)h^\varepsilon$, we get

$$\varepsilon\,\partial_t\tilde{h}^\varepsilon + v.\,\nabla_x\tilde{h}^\varepsilon = \chi h_0^\varepsilon + \nabla_v\,.\,(\chi h_1^\varepsilon) + (\varepsilon\partial_t + v.\,\nabla_x)\chi h^\varepsilon. \tag{43}$$

Moreover, due to the uniform bound of h^ε in L^2, we have

$$\left\| \int_{\mathbb{R}^d} \left(h^\varepsilon(t,x,v) - \tilde{h}^\varepsilon(t,x,v) \right) \psi(v) dv \right\|_{L^1((0,T)\times\omega)} \to 0$$

when α goes to zero, uniformly in ε.

Step 2: From step 1, we see that it is enough to prove the lemma when we replace $(0,T)\times\omega$ by $\mathbb{R}_t \times \mathbb{R}^d_x$ since the extra term on the right-hand side of (43) is in L^2. Notice also that since we localize in the v variable by integrating against $\psi(v)$, the L^2 norm controls the L^1 norm. More precisely, we can prove that under the conditions of the lemma, $\int_{\mathbb{R}^d} \psi(v) h^\varepsilon(t,x,v) dv \in L^{r,\infty}((0,T); B^{s,r}_{\infty,\infty})$, where $r = \frac{5}{3}$ and $s = \frac{1}{5}$ (see [12] for the definition of the Besov space built on the Lorentz space $L^{r,\infty}$). One way of proving the bound in the Besov space is to use the Littlewood–Paley decomposition and follow the proof of Theorem 3 of [12]. The only difference is that $\xi.v$ should be replaced by $\varepsilon\tau + \xi.v$ and that we take the Fourier transform in t and x (see also Theorem 1.8 of [8]).

Here, we would like to sketch a proof which follows the idea used in [19]. Adding λh^ε to both sides of (29), we get

$$\lambda h^\varepsilon + \varepsilon\, \partial_t h^\varepsilon + v.\nabla_x h^\varepsilon = h_0^\varepsilon + \nabla_v \,.\, h_1^\varepsilon + \lambda h^\varepsilon.$$

Hence,

$$\int_{\mathbb{R}^d} h^\varepsilon(t,x,v)\psi(v)dv = T_\lambda(h_0^\varepsilon + \nabla_v \,.\, h_1^\varepsilon + \lambda h^\varepsilon), \tag{44}$$

where

$$T_\lambda g(t,x) = \int_0^\infty \int_{\mathbb{R}^d} g(t-\varepsilon s, x - sv, v) e^{-\lambda s} \phi(v) dv\, ds. \tag{45}$$

Applying Proposition 3.1 of [19], we deduce that

$$\|T_\lambda(g)\|_{L^2_t H^{1/2}_x} \le C\lambda^{-1/2} \|g\|_{L^2(\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^d)}, \tag{46}$$

$$\|T_\lambda(g)\|_{\lambda^{-2}L^1_t W^{-1,1}_x + \lambda^{-1} L^1_t L^1_x} \le C \|g\|_{L^1(\mathbb{R}_t\times\mathbb{R}^d_x; W^{-1,1}(\mathbb{R}^d))}. \tag{47}$$

Hence,

$$\int_{\mathbb{R}^d} h^\varepsilon(t,x,v)\psi(v)dv = \rho = \rho^1 + \rho^2, \tag{48}$$

where

$$\|\rho^1\|_{L^2_t H^{1/2}_x} \le C\lambda^{1/2} \|h^\varepsilon\|_{L^2}, \tag{49}$$

$$\|\rho^2\|_{\lambda^{-2}L^1_t W^{-1,1}_x + \lambda^{-1} L^1_t L^1_x} \le C(\|h_0^\varepsilon\|_{L^1} + \|h_1^\varepsilon\|_{L^1}). \tag{50}$$

This can also be written as $\rho^2 = \rho_1^2 + \rho_2^2$, where

$$\|\rho_1^2\|_{L^1_t W^{-1,1}_x} \le C\lambda^{-2}(\|h_0^\varepsilon\|_{L^1} + \|h_1^\varepsilon\|_{L^1}), \tag{51}$$

$$\|\rho_2^2\|_{L^1_t L^1_x} \le C\lambda^{-1}(\|h_0^\varepsilon\|_{L^1} + \|h_1^\varepsilon\|_{L^1}). \tag{52}$$

We would like to deduce that $\rho \in [L^2H^{1/2}, L^1W^{-1,1}]_{(1/5,\infty)}$, the real interpolation of order $(1/5,\infty)$ of the couple $(L^2H^{1/2}, L^1W^{-1,1})$. For all $t \in \mathbb{R}_+$, we define the function

$$K(t) = \inf_{a_1+a_2=\rho} \|a_1\|_{L^2H^{1/2}} + t\|a_2\|_{L^1W^{-1,1}}. \tag{53}$$

To conclude, it is enough to prove that $K(t) \leq Ct^{1/5}$. For $t > 0$, we take λ such that $t = \lambda^{5/2}$.

If $0 < t < 1$, then ρ_2^2 also satisfies $\|\rho_2^2\|_{L_t^1 W_x^{-1,1}} \leq C\lambda^{-2}$, and hence taking $a_1 = \rho^1$ and $a_2 = \rho^2$, we deduce that $K(t) \leq Ct^{1/5}$.

If $t > 1$, then we write ρ_2^2 as the sum of two terms $\rho_2^2 = \rho_3^2 + \rho_4^2$ such that $\rho_3^2 \in \lambda^{-2}L^1W^{-1,1}$ and $\rho_4^2 \in \lambda^{1/2}L^1W^{3/2,1}$. Hence, if we define

$$K_1(t) = \inf_{a_1+a_2=\rho} \|a_1\|_{L^2H^{1/2}+L^1W^{3/2,1}} + t\|a_2\|_{L^1W^{-1,1}} \tag{54}$$

we get that $K_1(t) \leq Ct^{1/5}$ by taking $a_1 = \rho^1 + \rho_4^2$ and $a_2 = \rho_1^2 + \rho_3^2$.

This proves that $\rho \in [L^2H^{1/2} + L^1W^{3/2,1}, L^1W^{-1,1}]_{(1/5,\infty)}$. This is, of course, enough to get the compactness stated in the lemma.

Acknowledgments. The second author would like to thank N. Ben Abdallah for interesting discussions and encouragements. The authors would like to thank the referee for many constructive remarks.

REFERENCES

[1] N. Ben Abdallah, *Weak solutions of the initial-boundary value problem for the Vlasov-Poisson system*, Math. Methods Appl. Sci., 17 (1994), pp. 451–476.

[2] N. Ben Abdallah and P. Degond, *On hierarchy of macroscopic models for semiconductors*, J. Math. Phys., 37 (1996), pp. 3306–3333.

[3] N. Ben Abdallah and J. Dolbeault, *Relative entropies for kinetic equations in bounded domains (irreversibility, stationary solutions, uniqueness)*, Arch. Ration. Mech. Anal., 168 (2003), pp. 253–298.

[4] N. Ben Abdallah and M. L. Tayeb, *Asymptotique de diffusion pour le système de Boltzmann-Poisson uni-dimensionnel*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 735–740.

[5] N. Ben Abdallah and M. L. Tayeb, *Diffusion approximation for the one dimensional Boltzmann-Poisson system,* Discrete Contin. Dynamical Syst. Ser. B, 4 (2004), pp. 1129–1142.

[6] N. Ben Abdallah and M. L. Tayeb, *Diffusion limit of the Boltzman-Poisson system for elastic collison*, manuscript.

[7] C. Bardos, F. Golse, and C.-D. Levermore, *Fluid dynamic limits of kinetic equations.* II. *Convergence proofs for the Boltzmann equation*, Comm. Pure Appl. Math., 46 (1993), pp. 667–753.

[8] F. Bouchut, F. Golse, and M. Pulvirenti, *Kinetic Equations and Asymptotic Theory*, Series in Applied Mathematics 4, 2000.

[9] S. Chandrasekhar, *Radiative Transfer*, Dover Publications, New York, 1960.

[10] R. DeVore and G. Petrova, *The averaging lemma*, J. Amer. Math. Soc., 14 (2001), pp. 279–296.

[11] R. J. Diperna and P.-L. Lions, *On the Cauchy problem for Boltzmann equations: Global existence and weak stability*, Ann. of Math., 130 (1989), pp. 321–366.

[12] R. J. Diperna, P.-L. Lions, and Y. Meyer, *L^p regularity of velocity averages*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 271–287.

[13] N. Dunford and J. T. Schwartz, *Linear Operators*, Intersciences Publishers, New York, 1958.

[14] H. Gajewski, *On the uniqueness of solutions to the drift-diffusion-model of semiconductor devices*, Math. Models Methods Appl. Sci., 4 (1994), pp. 121–133.

[15] H. Gajewski and K. Gröger, *Semiconductor equations for variable mobilities based on Boltzmann statistics or Fermi-Dirac statistics*, Math. Nachr., 140 (1989), pp. 7–36.

[16] F. Golse, B. Perthame, and R. Sentis, *Un résultat de compacité pour les équations de transport et application au calcul de la limite de la valeur propre principale d'un opérateur de transport*, C. R. Acad. Sci. Paris Sér. I Math., 301 (1985), pp. 341–344.

[17] F. Golse, P.-L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.

[18] F. Golse and F. Poupaud, *Limite fluide des équation de Boltzmann des semi-conducteurs pour une statistique de Fermi-Dirac*, Asymptotic Anal., 6 (1992), pp. 135–160.

[19] P.-E. Jabin and L. Vega, *A real space method for averaging lemmas*, J. Math. Pures Appl., 9 (2004), pp. 1309–1351.

[20] C. Kittel, *Introduction to Solid State Physics*, John Wiley & Sons, New York, 1968.

[21] P.-L. Lions, *Compactness in Boltzmann's equation via Fourier integral operators and applications*. III, J. Math. Kyoto Univ., 34 (1994), pp. 539–584.

[22] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. Vol.* 2. *Compressible Models*, Oxford Lecture Series in Mathematics and its Applications 10. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1998.

[23] P.-L. Lions and N. Masmoudi, *From the Boltzmann equations to the equations of incompressible fluid mechanics*, I, II, Arch. Ration. Mech. Anal., 158 (2001), pp. 173–193, 195–211.

[24] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, New York, 1989.

[25] S. Mischler, *On the initial boundary value problem for the Vlasov-Poisson-Boltzmann system*, Comm. Math. Phys., 210 (2000), pp. 447–466.

[26] S. Mischler and B. Perthame, *Boltzmann equation with infinite energy: Renormalized solutions and distributional solutions for small initial data and initial data close to a Maxwellian*, SIAM J. Math. Anal., 28 (1997), pp. 1015–1027.

[27] F. Poupaud, *Diffusion approximation of the linear semiconductor Boltzmann equation: Analysis of boundary layers*, Asymptotic Anal., 4 (1991), pp. 293–317.

[28] F. Poupaud and J. Soler, *Parabolic limit and stability of the Vlasov-Fokker-Planck system*, Math. Models Methods Appl. Sci., 10 (2000), pp. 1027–1045.

[29] M.-L. Tayeb, *Etude mathématiques de quelques équations de Boltzmann des semiconducteurs*, Thèse de l'Ecole Nationale d'Ingénieurs de Tunis, Tunis, Tunisia, 2002.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1808–1824

ON THE CONVERGENCE OF THE NO RESPONSE TEST*

ROLAND POTTHAST†

Abstract. The no response test is a new scheme in inverse problems for partial differential equations which was recently proposed in [D. R. Luke and R. Potthast, *SIAM J. Appl. Math.*, 63 (2003), pp. 1292–1312] in the framework of inverse acoustic scattering problems. The main idea of the scheme is to construct special probing waves which are small on some test domain. Then the response for these waves is constructed. If the response is small, the unknown object is assumed to be a subset of the test domain. The response is constructed from *one*, *several*, or *many* particular solutions of the problem under consideration. In this paper, we investigate the *convergence* of the no response test for the reconstruction information about inclusions D from the Cauchy values of solutions to the Helmholtz equation on an outer surface $\partial\Omega$ with $\overline{D} \subset \Omega$. We show that the one-wave no response test provides a criterion to test the analytic extensibility of a field. In particular, we investigate the construction of approximations for the set of singular points $N(u)$ of the total fields u from one given pair of Cauchy data. Thus, the no response test solves a particular version of the classical *Cauchy problem*. Also, if an infinite number of fields is given, we prove that a *multifield version* of the no response test reconstructs the unknown inclusion D. This is the first convergence analysis which could be achieved for the no response test.

Key words. inverse problems, inverse scattering, shape reconstruction, sampling method, no response test

AMS subject classifications. 78A46, 81U40, 35J05, 31B20

DOI. 10.1137/S0036141004441003

1. Introduction.

We will study the no response test for the treatment of the following basic model problems from the theory of inverse boundary value problems.

Let Ω and D be bounded domains with boundary of class C^2, $\overline{D} \subset \Omega$ and $\Omega \setminus \overline{D}$ connected. For Ω we consider the simple situation where it is simply connected. We assume that the interior homogeneous Dirichlet problem for the Helmholtz equation in Ω does have only the trivial solution, i.e., κ is not an interior Dirichlet eigenvalue of Ω. In this case the interior Dirichlet problem for Ω can be solved with arbitrary continuous or L^2 boundary data and the solutions depend continuously on these data with respect to any C^μ-norm on compact subsets of Ω. Also, we assume that the same condition is satisfied for $\Omega \setminus \overline{D}$. In this case the following boundary value problem has a unique solution; cf. [3].

Boundary value problem. We consider a function $u \in C^2(\Omega \setminus \overline{D}) \cap C(\overline{\Omega} \setminus D)$ which satisfies the Helmholtz equation

$$\triangle u + \kappa^2 u = 0 \quad \text{in } \Omega \setminus \overline{D} \tag{1.1}$$

with boundary values

$$u = 0 \quad \text{on } \partial D. \tag{1.2}$$

For the *forward problem*, the function

$$u|_{\partial\Omega} = f \tag{1.3}$$

*Received by the editors February 12, 2004; accepted for publication (in revised form) June 9, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/sima/38-6/44100.html
†Institut for Numerical and Applied Mathematics, University of Göttingen, Germany (potthast@scienceatlas.de, http://www.scienceatlas.de/potthast/).

is given, and we look for the solution u to (1.1)–(1.3) and its normal derivative

$$\frac{\partial u}{\partial \nu} = g \tag{1.4}$$

on $\partial\Omega$. Here, the unit normal vector ν is assumed to point into the exterior of $\Omega \setminus \overline{D}$.

Inverse boundary value problem. For the inverse problem we assume that the *Cauchy data*

$$u|_{\partial\Omega} = f, \quad \left.\frac{\partial u}{\partial \nu}\right|_{\partial\Omega} = g \tag{1.5}$$

are given. The problem is to reconstruct D or to gain information about D by reconstructing a subset of D without knowledge of the boundary condition on ∂D.

The no response test. We assume that Cauchy data (1.5) of some solution u to (1.1)–(1.5) are given. For some test domain $G \subset \Omega$ the no response test calculates a *response* by probing the unknown scatterer D with special waves, for example, with single-layer potentials

$$v[\varphi](y) := \int_{\partial\Omega} \Phi(x,y)\varphi(x)ds(x), \quad y \in \Omega,\ \varphi \in L^2(\partial\Omega), \tag{1.6}$$

where the functions $v[\varphi]$ are chosen such that

$$\|v[\varphi]\|_{C^1(\overline{G})} \leq \epsilon \tag{1.7}$$

on G with some constant $\epsilon > 0$. We define

$$J(x) := \frac{1}{2}u(x) - \int_{\partial\Omega} \left(\frac{\partial u}{\partial \nu}(y)\Phi(x,y) - u(y)\frac{\partial \Phi(x,y)}{\partial \nu(y)} \right) ds(y), \quad x \in \partial\Omega, \tag{1.8}$$

and calculate the *response* of the domain D with respect to the probing function $v[\varphi]$ by

$$I(\varphi) := \int_{\partial\Omega} J(x)\varphi(x)ds(x). \tag{1.9}$$

The construction of the response is described in more detail in section 2.

The no response test uses a set $\mathcal{G}$ of sampling or test domains G, respectively. For each domain G the *maximal response* is defined as the supremum over all responses for special functions $v[\varphi]$ with (1.7). All test domains G with maximal response smaller than c_0 with some constant c_0 are called *positive* (= no response), otherwise *negative*. To obtain a reconstruction, we build the intersections

$$M_{rec} := \bigcap_{G\ positive} \overline{G} \tag{1.10}$$

of all positive test domains. The extension of the response functional to the multiwave case is described in section 2.

Convergence analysis for the no response test in different cases. For our convergence analysis we will work with admissible test domains.

DEFINITION 1.1 (admissible test domains). *Let $G \subset \Omega$ be some domain such that there is a homotopy $[0,1] \ni \lambda \mapsto G_\lambda$ with $G_0 = \Omega$ and $G_1 = G$ such that G_λ is a simply connected domain with boundary of class C^2 and $\overline{G_{\lambda_2}} \subset G_{\lambda_1}$ for $\lambda_2 > \lambda_1$. In this case G is called an* admissible *test domain.*

The results of our subsequent convergence analysis can be summarized as follows. Let G be some admissible test domain.

1. *One-wave case.* If the field u can be analytically continued into $\Omega \setminus G$, then for an arbitrary constant $c_0 > 0$ and sufficiently small $\epsilon > 0$ the domain G is positive. If u cannot be analytically continued into $\Omega \setminus \overline{G}$, then G is negative/not positive.
2. *Multiwave case.* The intersection of all positive test domains is exactly the unknown domain D.

For the *one-wave case* we will show that the no response functional is bounded from below by a constant times the derivatives of the field u in the exterior of $\overline{G}$. This is obtained by further developing techniques introduced in the framework of the *point-source method*; cf. [10], [12]. The point-source method reconstructs the total field u or its derivatives and then uses the boundary condition to find the unknown inclusion D from the knowledge of u and the boundary condition. However, the basic difference between the point-source method and the no response test is that the point-source method does calculate the field or one particular derivative, whereas the no response test estimates the behavior of the full set of Taylor coefficients.

The results show that the no response test and the *range test* [13] in principle reconstruct the same set, even though these methods use a different indicator function for the test domains. The idea of the range test is to solve the single-layer integral equation

$$\int_{\partial G} e^{-i\kappa \hat{x}\cdot y} \varphi(y) ds(y) = u^\infty(\hat{x}), \quad \hat{x} \in \mathbb{S}, \tag{1.11}$$

with $\varphi \in L^2(\partial G)$ to test whether the field u^∞ can be analytically extended up to ∂G. Both methods test for analytic extensibility.

For the convergence analysis in the *multiwave case* we will bound the no response functional by a factor times the indicator function of the *singular sources method* as introduced in [11] and [12] to obtain convergence. The singular sources scheme reconstructs the scattered field $\Psi^s(z,z)$ for an incident singular field $\Psi(\cdot,z)$ in its source point z and uses the behavior $|\Psi^s(z,z)| \to \infty$ for $z \to \partial D$.

The idea to use singular incident fields for *uniqueness results* has been employed in the work of Isakov [6] and Kirsch and Kress [7]. Corresponding *stability results* using singular sources have been derived in [11]. Also, these ideas have been transformed into a constructive *singular sources method* by the author [11], [12]. Independently, Ikehata proposed using singular waves for shape reconstruction by the *probe method*; see [5]. More recently, the probe method has been numerically realized by Erhard and Potthast [2], and the numerical behavior in the case of an impedance boundary condition was carefully investigated by Cheng, Liu, and Nakamura [1]. The convergence properties of different sampling methods have been compared and investigated by Honda et al. in [4].

2. The no response test for the boundary value problem. The goal of this section is to develop a formulation of the no response test for the inverse boundary value problem (1.1)–(1.5).

Preparations. By Green's formula the total field u can be calculated by

$$u(x) = \int_{\partial\Omega} \left(\frac{\partial u}{\partial \nu}(y)\Phi(x,y) - u(y)\frac{\partial \Phi(x,y)}{\partial \nu(y)} \right) ds(y) + \int_{\partial D} \frac{\partial u}{\partial \nu}(y)\Phi(x,y)ds(y), \quad x \in \Omega \setminus \overline{D}, \tag{2.1}$$

where we already used the boundary condition on ∂D to simplify the second integral over ∂D. Using the classical jump relations we can calculate the limit of $u(x)$ for $x \to \partial\Omega$. We obtain

$$u_-(x) = \frac{1}{2}u(x) + \int_{\partial\Omega} \left(\frac{\partial u}{\partial \nu}(y)\Phi(x,y) - u(y)\frac{\partial \Phi(x,y)}{\partial \nu(y)} \right) ds(y) + \int_{\partial D} \frac{\partial u}{\partial \nu}(y)\Phi(x,y)ds(y), \quad x \in \partial\Omega. \tag{2.2}$$

For the inverse boundary value problem the functions $u|_{\partial\Omega}$ and $\frac{\partial u}{\partial \nu}|_{\partial\Omega}$ are known. Thus, the function J defined in (1.8) can be calculated from the given data. In particular, from (2.2) we derive

$$\int_{\partial D} \frac{\partial u}{\partial \nu}(y)\Phi(x,y)ds(y) = J(x), \quad x \in \partial\Omega. \tag{2.3}$$

Now, consider the single-layer potential (1.6) with density $\varphi \in L^2(\partial\Omega)$ or $C(\partial\Omega)$. We multiply (2.3) by φ, integrate over $\partial\Omega$, and exchange the order of integration to obtain

$$\begin{aligned} \int_{\partial\Omega} \varphi(x)J(x)ds(x) &= \int_{\partial\Omega} \varphi(x) \left(\int_{\partial D} \frac{\partial u}{\partial \nu}(y)\Phi(x,y)ds(y) \right) ds(x) \\ &= \int_{\partial D} \frac{\partial u}{\partial \nu}(y) \left(\int_{\partial\Omega} \Phi(x,y)\varphi(x)ds(x) \right) ds(y) \\ &= \int_{\partial D} \frac{\partial u}{\partial \nu}(y)v(y)ds(y). \end{aligned} \tag{2.4}$$

Thus, the value of the last integral can be calculated from the given data f and g as defined in (1.5).

No response test, one-wave case. First, we consider the case where only one pair of Cauchy data $f = u|_{\partial\Omega}$, $g = \frac{\partial u}{\partial \nu}|_{\partial\Omega}$ is given on $\partial\Omega$. Let $G \subset \Omega$ be a test domain. The idea of the no response test is to construct $v[\varphi]$ given in (1.6) such that it is small on G and large outside of G.

DEFINITION 2.1. *Given some $\epsilon > 0$ let $\mathcal{M}(G,\epsilon)$ be the set of all $\varphi \in L^2(\partial\Omega)$ such that the* test functions *$v[\varphi]$ satisfy the* impulse size condition

$$\|v[\varphi]|_G\|_{C^1(\overline{G})} \le \epsilon. \tag{2.5}$$

We call this testing with impulses of maximal size ϵ on G.

In detail, the following test is carried out.

DEFINITION 2.2 (no response test, one-wave case). *Choose $\epsilon > 0$ and some constant $c_0 > 0$, and consider admissible test domains $G \subset \Omega$. For all $\varphi \in \mathcal{M}(G,\epsilon)$ calculate the functional*

$$I(\varphi) := \int_{\partial\Omega} J(x)\varphi(x)ds(x) \tag{2.6}$$

and

$$I_\epsilon(G) := \sup_{\varphi \in \mathcal{M}(G,\epsilon)} I(\varphi). \tag{2.7}$$

If $I_\epsilon(G) < c_0$, we call the domain G positive *and we say that we have* no response. *Otherwise, G is denoted as* negative, *since the response is larger than c_0.*

We define

$$I_0(G) := \lim_{\epsilon \to 0} I_\epsilon(G). \tag{2.8}$$

We will show that if u can be analytically extended into $\Omega \setminus G$, then $I_0(G) = 0$, and, conversely, the condition $I_0(G) = 0$ implies analytic extensibility into $\Omega \setminus \overline{G}$. The closure $\overline{G}$ of each test domain G with $I_0(G) = 0$ does contain the singular points $N(u)$ of the field u. Taking the intersections of the sets $\overline{G}$ for all possible test domains G with $I_0(G) = 0$ we obtain some upper estimate for this set of singular points $N(u)$ of u. The next sections will give a precise analysis of this situation. For practical reconstructions we will choose $\epsilon > 0$ and $c_0 > 0$ and obtain an approximation for $N(u)$ by the intersection of all positive test domains.

No response test, multiwave case. Now, we assume that more than one pair of Cauchy data f, g is given. Let $f(\xi), g(\xi)$ for $\xi = 1, \dots, N$ be a finite (for $N \in \mathbb{N}$) or infinite (for $N = \infty$) number of linearly independent Cauchy data. We assume that the set

$$\left\{ f(\xi) := u(\cdot, \xi)\Big|_{\partial\Omega} : \ \xi \in \mathbb{N} \right\} \tag{2.9}$$

is dense (and thus complete) in $L^2(\partial\Omega)$. Now, the functional J depends on $\xi = 1, 2, \dots, N$; i.e., we have

$$J(x,\xi) := \frac{1}{2} u(x,\xi) - \int_{\partial\Omega} \left(\frac{\partial u(y,\xi)}{\partial \nu(y)} \Phi(x,y) - u(y,\xi) \frac{\partial \Phi(x,y)}{\partial \nu(y)} \right) ds(y) \tag{2.10}$$

for $x \in \partial\Omega$ and $\xi = 1, \dots, N$. Then we formulate the following no response test for the multiwave case.

Definition 2.3 (no response test, multiwave case). *Let G be some test domain. Choose $\epsilon_1 > 0$ and some constant $c_0 > 0$. For all $\varphi \in \mathcal{M}(G, \epsilon_1)$ calculate the boundary values $v[\varphi]|_{\partial\Omega}$. With the given data $f(\xi)$, $\xi = 1, \dots, N$, find an approximation*

$$q[\alpha] := \sum_{\xi=1}^{N} \alpha_\xi f(\xi) \tag{2.11}$$

to $v[\varphi]|_{\partial\Omega}$ with

$$\| v[\varphi]|_{\partial\Omega} - q[\alpha] \|_{L^2(\partial\Omega)} \le \epsilon_2. \tag{2.12}$$

Here, we use $\alpha = (\alpha_1, \alpha_2, \dots)$ and define $\mathcal{Q}(\epsilon_2, \varphi)$ as the set of α such that (2.12) is satisfied. The size of ϵ_2 depends on the number of boundary data which are available. For $N = \infty$ the value ϵ_2 can be chosen arbitrarily small. Then, with the functional

$$I(\varphi, \alpha) := \sum_{\xi=1}^{N} \alpha_\xi \int_{\partial\Omega} J(x,\xi) \varphi(x) ds(x), \tag{2.13}$$

we calculate

$$I_{\epsilon_1,\epsilon_2}(G) := \sup_{\varphi\in\mathcal{M}(\epsilon_1),\alpha\in\mathcal{Q}(\epsilon_2,\varphi)} I(\varphi,\alpha). \tag{2.14}$$

If $I_{\epsilon_1,\epsilon_2}(G) < c_0$*, we call the domain* G positive *and we say that we have* no response. *Otherwise,* G *is denoted as* negative*, since the response is (relatively) large.*

Later, we will carry out the convergence analysis for the case where $N = \infty$. Then, analogously to the one-dimensional case, we define

$$I_{0,0}(G) := \lim_{\epsilon_1\to 0}\lim_{\epsilon_2\to 0} I_{\epsilon_1,\epsilon_2}(G). \tag{2.15}$$

We will show that $I_{0,0}(G)$ is zero if and only if $D \subset G$. If $D \not\subset G$, we obtain $I_{0,0}(G) = \infty$. Then in the multiwave case the no response test will denote exactly those domains G as positive which contain the unknown scatterer D in its interior.

3. On the *extensibility* and *singular points* of a solution to the boundary value problem. Let u be a solution to the boundary value problem (1.1)–(1.5). If the domain D has analytic boundary, it is well known (see 5.7.1 in [9]) that the field u can be analytically extended into the interior of D. The convergence of the no response test is strongly linked to the extensibility of the field u. Thus, for a proper analysis of the convergence of the no response test we need to study the extensibility of u into the interior of D.

We would like to define some type of "minimal" set M such that u can be extended into the exterior $\Omega \setminus M$ of M but not into any $\Omega \setminus N$ for $N \subset M$, $N \neq M$. However, due to the possible existence of Riemannian surfaces and the possible nonuniqueness of this extension, in general, the set M will not be well defined as a single set.

A different idea would be to test for some kind of "singular" points on a Riemannian surface. However, there are a number of different types of "singular points." One possible definition would be to consider points to be singular where one of the derivatives of the field u becomes singular. However, even if all derivatives are bounded, there might not be an analytic extension into the particular point. Further, a point z might be singular in the sense that one extension of u does have a singular derivative and there might still be some analytic extension of u into the point z along some different and complicated path. Here, we have chosen to use some kind of minimal definition. We call a point singular only if there does not exist *any* analytic continuation into this point.

Definition 3.1 (singular points of u). *Consider an analytic function in* $\Omega \setminus \overline{D}$*. We call a point* $z \in \Omega$ *a* singular point *of the field* u *if there is no analytic continuation of the field* u *into a neighborhood of* z*.*

Remark. The set of singular points might be empty. If it is not empty, it is a subset of the scattering domain $\overline{D}$, since by assumption u is analytic in $\Omega \setminus \overline{D}$. Due to the existence of different analytic extensions on Riemannian surfaces there might be many singularities of these extended fields in different points of Ω. But in singular points the field u is singular on any branch of the Riemannian surface or is not contained in these branches at all.

In this work, we will restrict our attention to the problem of testing extensibility into the exterior of some admissible test domain G. Then all points into which no analytic extension exists will be a subset of this test domain; i.e., we obtain some upper estimate for the set of singular points of the field u.

The rest of this section is used to formulate some preparations. One basic tool for the investigation of analytic extensions is Taylor's expansion of the field u.

LEMMA 3.2. *We assume that $A_e := \Omega \setminus \overline{A}$ is the open interior of the complement of some simply connected open set A with boundary of class C^2. If u is analytic in A_e and for some $\rho > 0$ the set*

$$\left\{ a_\mu(z) := \sup_{|h|=1} \rho^\mu \tfrac{|(h\cdot\nabla)^\mu u(z)|}{\mu!} : \quad \mu \in \mathbb{N} \right\} \tag{3.1}$$

of Taylor coefficients of u is uniformly bounded by a constant C for all $z \in A_e$ with $d(z, \partial\Omega) > \rho$, then u is extensible into an open neighborhood of $\overline{A_e}$; i.e., there is a set A' with $\overline{A'} \subset A$ such that u is extensible into $A'_e = \Omega \setminus \overline{A'}$.

Proof. Consider some boundary point z_0. Due to our assumption for some $\rho > 0$ the modulus of $a_\mu(z)$ is bounded for z in a neighborhood of z_0 by a constant C. Then the Taylor series

$$\sum_{\mu=0}^{\infty} \frac{1}{\mu!}((x-z)\cdot\nabla)^\mu u(z) \tag{3.2}$$

can be estimated by

$$\Big| \sum_{\mu=0}^{\infty} \frac{1}{\mu!}((x-z)\cdot\nabla)^\mu u(z) \Big| \;\le\; \sum_{\mu=0}^{\infty} a_\mu \left(\frac{|x-z|}{\rho} \right)^\mu . \tag{3.3}$$

For $|x - z| < \rho$, the series (3.3) is absolutely convergent. In this case we apply Lemma 7.2 to see that u can be analytically extended into the ball $B_\rho(z)$ of z_0. This is the case for every boundary point z.

At this point, we still might obtain different extensions of u for different boundary points z_0. However, since ∂A is of class C^2 we can choose $\tau > 0$ sufficiently small such that for each two balls $B_\tau(z_1)$ and $B_\tau(z_2)$ with $z_1, z_2 \in \partial A$ with $B_\tau(z_1) \cap B_\tau(z_2) \neq \emptyset$ the intersection $B_\tau(z_1) \cap B_\tau(z_2) \cap A_e$ is not empty and thus the extensions into the simply connected set $B_\tau(z_1) \cap B_\tau(z_2)$ are identical. ∎

4. Convergence in the one-wave case. The goal of this section is to prove the first of our two main theorems—the convergence for the one-wave case.

THEOREM 4.1. *Let G with $\overline{G} \subset \Omega$ be an admissible test domain. We have $I_0(G) = 0$ if u can be analytically extended into $\Omega \setminus G$. If u cannot be analytically extended into $\Omega \setminus \overline{G}$, then we have $I_\epsilon(G) = \infty$ for all $\epsilon > 0$ and thus $I_0(G) = \infty$.*

Proof. First, we show that if $D \subset G$, i.e., if the true inclusion D is a subset of the test domain G, then we obtain $I_0(G) = 0$. For $D \subset G$ and $v[\varphi]$ defined by (1.6) we have

$$|v[\varphi](y)| \le \epsilon, \quad y \in \partial D \subset \overline{G}. \tag{4.1}$$

Thus, from (2.4) we obtain

$$\begin{aligned} |I(\varphi)| &\le \Big| \int_{\partial D} \frac{\partial u}{\partial \nu}(y) v[\varphi](y) ds(y) \Big| \\ &\le \epsilon \int_{\partial D} \Big| \frac{\partial u}{\partial \nu}(y) \Big| ds(y) \\ &\le \epsilon c \end{aligned} \tag{4.2}$$

with some constant c. Passing to the limit $\epsilon \to 0$ we obtain $I_0(G) = 0$.

In the next step, we assume that the field u can be analytically extended up to $\mathbb{R}^m \setminus G$; i.e., u and $\frac{\partial u}{\partial \nu}$ are well defined and continuous on ∂G (cf. Figure 4.1). If we are not in the first case, then

$$D^* := D \cap (\mathbb{R}^m \setminus G) \tag{4.3}$$

is not empty. We define

$$\Gamma := \partial(G \cap D).$$

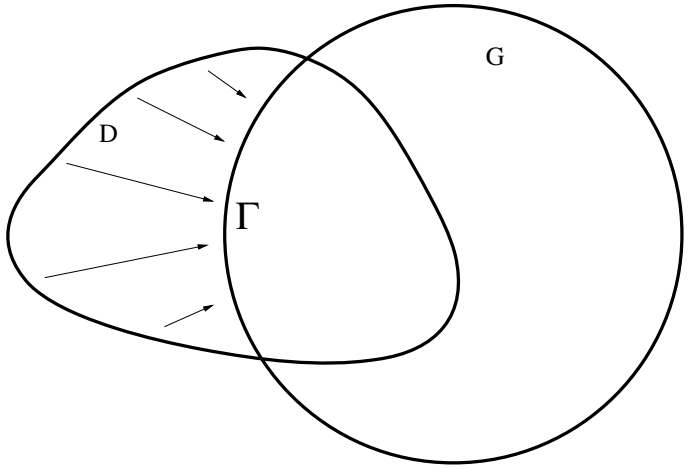


FIG. 4.1. *Domains D and G. The arrows indicate how the integration over $\partial D \setminus \overline{G}$ is transformed into an integration over Γ.*

In this case we can use $u|_{\partial D} = 0$ and apply Green's second theorem to v and u in D^* to calculate

$$\begin{aligned} &\int_{\partial D} \frac{\partial u}{\partial \nu}(y) v[\varphi](y) ds(y) \\ &= \int_{\partial D} \left(\frac{\partial u}{\partial \nu}(y) v[\varphi](y) - \frac{\partial v[\varphi]}{\partial \nu}(y) u(y) \right) ds(y) \\ &= \int_{\Gamma} \left(\frac{\partial u}{\partial \nu}(y) v[\varphi](y) - \frac{\partial v[\varphi]}{\partial \nu}(y) u(y) \right) ds(y). \end{aligned} \tag{4.4}$$

Since u and $\frac{\partial u}{\partial \nu}$ are well defined and bounded on Γ, and since

$$\|v[\varphi]\|_{C^1(\overline{G})} \to 0, \quad \epsilon \to 0, \tag{4.5}$$

the integral in (4.4) tends to zero for $\epsilon \to 0$. This proves that for u extensible into $\Omega \setminus G$ the no response test yields $I_0(G) = 0$.

In the third and final step we will show that if u cannot be analytically extended into $\Omega \setminus \overline{G}$, we have $I_0(G) = \infty$.

First, if u cannot be analytically extended into $\Omega \setminus \overline{G}$, then there must be some maximal parameter $\lambda_0 \in (0, 1)$ such that u can be analytically extended into $\Omega \setminus G_\lambda$ for all $\lambda < \lambda_0$, but not into any $\Omega \setminus G_\lambda$ for any $\lambda > \lambda_0$. Then, for all $\rho > 0$, the set (3.1) cannot be uniformly bounded for all $z \in \Omega \setminus G_{\lambda_0}$, since in this case we could construct

an analytic extension of u into the interior of G_{λ_0} and λ_0 would not be maximal. Thus, given some $\rho > 0$ there must be a point $z_0 \in \Omega \setminus G_{\lambda_0}$ (with $d(z_0, \partial\Omega) > \rho$) such that (3.1) is not uniformly bounded in any neighborhood of z_0. The point z_0 has a positive distance

$$\rho_0 := \inf_{y \in \overline{G}} |y - z_0| \tag{4.6}$$

to the set G. Given $h \in \mathbb{R}^m$ with $|h| = 1$ we define

$$\beta(z, \mu) := \sup_{y \in \overline{G}} \Big\{ (h \cdot \nabla_z)^\mu \Phi(y, z), \;\; \nabla_y (h \cdot \nabla_z)^\mu \Phi(y, z) \Big\} \tag{4.7}$$

for z in a neighborhood of z_0 and $\mu \in \mathbb{N}$. Then the $C^1(\overline{G})$-norm of the function

$$\Psi(\cdot, z) := \frac{\epsilon}{4\beta(z, \mu)} (h \cdot \nabla_z)^\mu \Phi(\cdot, z) \tag{4.8}$$

is bounded by $\epsilon/2$ for z in a neighborhood V of z_0. For later use we assume that $V \subset B_{\rho_0/2}(z_0)$.

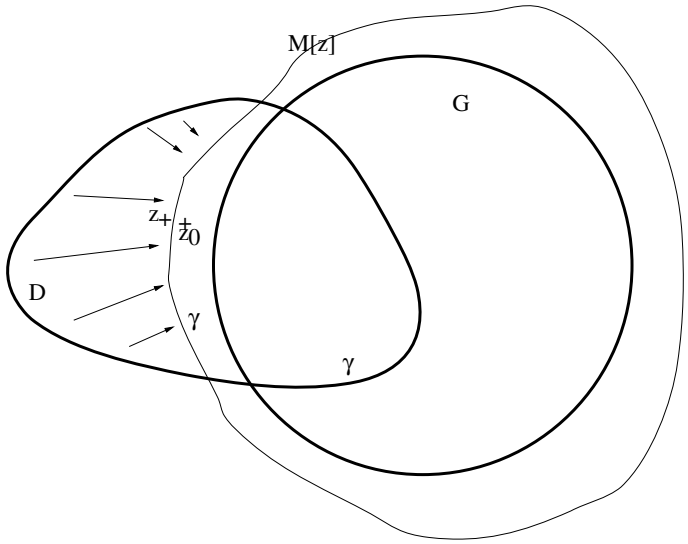


FIG. 4.2. *Domains D, G, and $M[z]$. The arrows indicate how the integration over $\partial D \setminus M[z]$ is transformed into an integration over γ.*

For $z \in V \setminus G_{\lambda_0}$ we can find a domain $M[z]$ with (1) boundary of class C^2 such that (2) the homogeneous interior Dirichlet problem for $M[z]$ has only the trivial solution, (3) u can be analytically extended into the exterior of $M[z]$, (4) $z \notin M[z]$, and (5) $M[z] \supset G$ (cf. Figure 4.2). Then the extension of u into the exterior of $M[z]$ implies that $u|_{\partial M[z]}$ and $\frac{\partial u}{\partial \nu}|_{\partial M[z]}$ are well defined.

Now, on $\overline{M[z]}$ we approximate the function $\Psi(\cdot, z)$ by a sequence $v_n[z] := v[\varphi_n]$ of single-layer potentials with density $\varphi_n \in L^2(\partial\Omega)$ as defined in (1.6), i.e., we have

$$\| v_n[z] - \Psi(\cdot, z) \|_{C^1(\overline{M[z]})} \leq \frac{1}{n}. \tag{4.9}$$

Since $G \subset M[z]$, we obtain the estimate

$$\begin{aligned} \|v_n[z]\|_{C^1(\overline{G})} &\leq \|v_n[z] - \Psi(\cdot, z)\|_{C^1(\overline{G})} + \|\Psi(\cdot, z)\|_{C^1(\overline{G})} \\ &\leq \frac{1}{n} + \frac{\epsilon}{2} \end{aligned} \tag{4.10}$$

for $z \in V \setminus N(u)$. For n sufficiently large, we have $\|v_n[z]\|_{C^1(\overline{G})} \leq \epsilon$. In this case we derive from Green's second theorem

$$\begin{aligned} &\int_{\partial D} \frac{\partial u}{\partial \nu}(y) v_n[z](y) ds(y) \\ &= \int_{\partial D} \left(\frac{\partial u}{\partial \nu}(y) v_n[z](y) - \frac{\partial v_n[z](y)}{\partial \nu} u(y) \right) ds(y) \\ &= \int_{\gamma} \left(\frac{\partial u}{\partial \nu}(y) v_n[z](y) - \frac{\partial v_n[z](y)}{\partial \nu} u(y) \right) ds(y), \end{aligned} \tag{4.11}$$

where we define $\gamma := \partial(M[z] \cap D)$ and where the normal on γ is oriented into the exterior of the closed curve. Since we have $\gamma \subset \overline{M[z]}$ in the limit $n \to \infty$ we obtain the convergence of

$$\begin{aligned} &\int_{\partial \Omega} J(x) \varphi_n(x) ds(x) \\ &= \int_{\partial D} \frac{\partial u}{\partial \nu}(y) v_n[z](y) ds(y) \\ &= \int_{\gamma} \left\{ \frac{\partial u}{\partial \nu}(y) v_n[z](y) - \frac{\partial v_n[z](y)}{\partial \nu} u(y) \right\} ds(y) \end{aligned}$$

towards the term

$$\begin{aligned} &\int_{\gamma} \left\{ \frac{\partial u}{\partial \nu}(y) \Psi(y, z) - \frac{\partial \Psi(y, z)}{\partial \nu(y)} u(y) \right\} ds(y) \\ &= \frac{\epsilon}{4\beta(z, \mu)} \int_{\gamma} \left\{ \frac{\partial u}{\partial \nu}(y) (h \cdot \nabla_z)^{\mu} \Phi(y, z) \right. \\ &\qquad \left. - \frac{\partial}{\partial \nu(y)} \left((h \cdot \nabla_z)^{\mu} \Phi(y, z) \right) u(y) \right\} ds(y) \end{aligned} \tag{4.12}$$

$$= \frac{\epsilon}{4\beta(z, \mu)} (h \cdot \nabla_z)^{\mu} u(z) - \frac{\epsilon}{4\beta(z, \mu)} (h \cdot \nabla_z)^{\mu} u_{ext}(z) \tag{4.13}$$

with

$$u_{ext}(z) := \int_{\partial \Omega} \left\{ \frac{\partial u}{\partial \nu}(y) \Phi(y, z) - \frac{\partial \Phi(y, z)}{\partial \nu(y)} u(y) \right\} ds(y). \tag{4.14}$$

As shown above for any fixed ρ the μth derivatives of u multiplied by $\rho^{|\mu|}/\mu!$ cannot be uniformly bounded for all $\mu \in \mathbb{N}$ in the neighborhood V of z_0, i.e.,

$$\forall C > 0 \ \exists \mu \in \mathbb{N} \text{ and } z \in V \setminus G_{\lambda_0} \text{ such that } \left| \frac{\rho^{\mu}}{\mu!} \cdot (h \cdot \nabla_{z_0})^{\mu} u(z) \right| \geq C. \tag{4.15}$$

For an analytic function $u(y)$ in a set V we know that for each $z \in V$ there is a neighborhood $B_{\rho}(z)$ such that

$$\sum_{\mu=0}^{\infty} \frac{((y - z) \cdot \nabla_z)^{\mu}}{\mu!} f(z) \tag{4.16}$$

is absolutely convergent towards f on $B_\rho(z)$. This yields

$$\left|(h\cdot\nabla_z)^\mu f(z)\right| \le c\frac{\mu!}{|y-z|^\mu} \tag{4.17}$$

with $h := \frac{(y-z)}{|y-z|}$ for some constant c. So due to the analyticity of the fundamental solution in z uniformly for $y \in \overline{G}$, the constant $\beta = \beta(z,\mu)$ is seen to be bounded by

$$|\beta(z,\mu)| \le c\frac{\mu!}{\rho_1^\mu}, \quad z \in V \setminus G_{\lambda_0}, \quad \mu \ge 1, \tag{4.18}$$

with some constants ρ_1 and c. Further, we remark that the exterior term (4.14) is analytic, and for ρ_1 sufficiently small the second term of (4.13) can be estimated by a constant $\tilde{c}$. Altogether, we can find sequences $(\mu_l)_{l\in\mathbb{N}} \subset \mathbb{N}, (z_l)_{l\in\mathbb{N}} \subset V$ and densities $\varphi_l \in L^2(\partial\Omega)$, $l \in \mathbb{N}$, which satisfy the impulse size condition (2.5) such that the indicator functional $I(\varphi_n)$ can be estimated from below by

$$\begin{aligned} I(\varphi_l) &= \int_{\partial D} \frac{\partial u}{\partial \nu}(y) v[\varphi_l](y) ds(y) \\ &\ge \frac{1}{c}\Big|\frac{\epsilon}{4}\frac{\rho_1^{\mu_l}}{\mu_l!}(h\cdot\nabla_z)^{\mu_l} u(z_l)\Big| - \tilde{c} \\ &\to \infty, \quad l \to \infty. \end{aligned} \tag{4.19}$$

Thus, $I_\epsilon(G)$ defined in (2.7) is unbounded and we have shown

$$I_\epsilon(G) = \infty, \quad I_0(G) = \infty, \tag{4.20}$$

which completes the proof of the theorem. □

5. Decompositions of the field u into $u = w + v$. This section contains several preparations which are needed to prove the convergence of the multiwave version of the no response test in the following section.

I. For the analysis of the multiwave case of the no response test we need the following decomposition of the field u into the sum of an *incident* field v and some *scattered field* w. Given the boundary data $f \in C(\partial\Omega)$ we define $v \in C^2(\Omega) \cap C(\overline{\Omega})$ to be the solution of the interior Dirichlet problem to the Helmholtz equation with boundary values f. Now, given a solution u to the boundary value problem (1.1)–(1.5), i.e., $u|_{\partial\Omega} = f$ and $u|_{\partial D} = 0$, we define

$$w := u - v \tag{5.1}$$

with the function v defined by its Dirichlet boundary data $f := u|_{\partial\Omega}$. The function v is considered to be some type of *incident* field, and the function w satisfies

$$w|_{\partial\Omega} = 0, \qquad w|_{\partial D} = -v|_{\partial D} \tag{5.2}$$

and is called a *scattered field*.

II. Let G be some test domain with $D \subset G$ and $z \in \Omega\setminus\overline{G}$ and consider a sequence (v_n), $n \in \mathbb{N}$, of single-layer potentials,

$$v_n := v[\varphi_n], \tag{5.3}$$

with boundary data $f_n := v[\varphi_n]$, such that

$$\|v[\varphi_n] - \beta\Phi(\cdot, z)\|_{C^1(\overline{G})} \to 0, \quad n \to \infty. \tag{5.4}$$

In particular, we obtain the convergence

$$\|v[\varphi_n] - \beta\Phi(\cdot, z)\|_{C^1(\partial D)} \to 0, \quad n \to \infty. \tag{5.5}$$

Let u_n be the solution of (1.1)–(1.5) with boundary data $\tilde{f}_n$ such that

$$\|\tilde{f}_n - f_n\|_{L^2(\partial\Omega)} \to 0 \tag{5.6}$$

for $n \to \infty$. We define

$$w_n := u_n - v_n. \tag{5.7}$$

Then we have

$$w_n|_{\partial\Omega} \to 0, \quad n \to \infty \text{ in } L^2(\partial\Omega), \tag{5.8}$$

$$w_n|_{\partial D} \to -\beta\Phi(\cdot, z), \quad n \to \infty, \text{ in } C(\partial D), \tag{5.9}$$

where the second equation is obtained via

$$\begin{aligned} \|w_n + \beta\Phi(\cdot, z)\|_{C(\partial D)} &\le \|u_n\|_{C(\partial D)} + \| - v_n + \beta\Phi(\cdot, z)\|_{C(\partial D)} \\ &\le \|v_n - \beta\Phi(\cdot, z)\|_{C(\partial D)}. \end{aligned} \tag{5.10}$$

From (5.8) and (5.9) by the unique solvability of the interior Dirichlet problem in $\Omega \setminus D$ we obtain the convergence of w_n towards some function w with

$$w|_{\partial\Omega} = 0, \tag{5.11}$$

$$w|_{\partial D} = -\beta\Phi(\cdot, z). \tag{5.12}$$

We will denote this function by $w(\cdot, z)$ to indicate the dependence on the source point z.

III. As a preparation for the next section now consider the response functional (2.13) for the multiwave case. Here, we construct $\alpha_{\xi,n}$ in such a way that

$$\left\| \sum_{\xi} \alpha_{\xi,n} u(\cdot, \xi) - f_n \right\|_{L^2(\partial\Omega)} \to 0, \quad n \to \infty, \tag{5.13}$$

for the boundary data f_n defined via (5.4). In this case using (2.4) we have

$$\begin{aligned} &\sum_{\xi} \alpha_{\xi,n} \int_{\partial\Omega} J(x, \xi)\varphi_n(x) ds(x) \\ &= \int_{\partial D} v_n(y) \sum_{\xi} \alpha_{\xi,n} \frac{\partial u(y, \xi)}{\partial \nu} ds(y) \\ &= \int_{\partial D} \sum_{\xi} \alpha_{\xi,n} \left\{ v_n(y) \frac{\partial u(y, \xi)}{\partial \nu} - \frac{\partial v_n(y)}{\partial \nu} \underbrace{u(y, \xi)}_{=0} \right\} ds(y). \end{aligned} \tag{5.14}$$

Trivial identity yields

$$\int_{\partial D}\left\{v_n(y)\frac{\partial v_n(y)}{\partial \nu}-\frac{\partial v_n(y)}{\partial \nu}v_n(y)\right\}ds(y)=0. \tag{5.15}$$

We subtract (5.15) from (5.14) to obtain

$$\begin{aligned}\sum_{\xi}\alpha_{\xi,n}\int_{\partial\Omega}J(x,\xi)\varphi_n(x)ds(x)&\\ =\int_{\partial D}\left\{v_n(y)\left(\sum_{\xi}\alpha_{\xi,n}\frac{\partial u(y,\xi)}{\partial\nu}-\frac{\partial v_n(y)}{\partial\nu}\right)\right.&\\ \left.-\frac{\partial v_n(y)}{\partial\nu}\left(\sum_{\xi}\alpha_{\xi,n}u(y,\xi)-v_n(y)\right)\right\}ds(y).&\end{aligned} \tag{5.16}$$

We define

$$u_n:=\sum_{\xi}\alpha_{\xi,n}u(\cdot,\xi),\quad n=1,2,3,\ldots. \tag{5.17}$$

By (5.13) this function u_n with $\tilde{f}_n := u_n|_{\partial\Omega}$ satisfies the condition (5.6) with v_n defined by (5.3)–(5.4). Then

$$w_n=u_n-v_n\ =\sum_{\xi=1}^{N}\alpha_{\xi,n}u(\cdot,\xi)-v_n \tag{5.18}$$

converges towards $w(\cdot)$ on ∂D, which is the scattered field for an incident point source with source point z. We remark that the function u_n is zero on ∂D and v_n is analytic and converges in $C^1(\partial D)$ towards $\beta\Phi(\cdot,z)$. By standard integral equation methods we obtain convergence of the normal derivatives on ∂D; i.e., the term

$$\frac{\partial w_n}{\partial\nu}=\frac{\partial}{\partial\nu}\Big(u_n-v_n\Big)=\sum_{\xi=1}^{N}\alpha_{\xi,n}\frac{\partial u(y,\xi)}{\partial\nu}-\frac{\partial v_n(y)}{\partial\nu} \tag{5.19}$$

converges towards $\frac{\partial w}{\partial\nu}$ on ∂D. Thus, for the functional (5.16) we have the convergence

$$\begin{aligned}&\sum_{\xi=1}^{N}\alpha_{\xi,n}\int_{\partial\Omega}J(x,\xi)\varphi_n(x)ds(x)\\ &\to\beta\int_{\partial D}\left\{\Phi(y,z)\frac{\partial w(y,z)}{\partial\nu(y)}-\frac{\partial\Phi(y,z)}{\partial\nu(y)}w(y,z)\right\}ds(y)\end{aligned} \tag{5.20}$$

for $n\to\infty$.

6. Convergence in the multiwave case. We are now prepared to prove the second of our two main theorems.

THEOREM 6.1. *Consider the multiwave case of the no response test where the condition* (2.9) *is satisfied. For some test domain G we obtain $I_{0,0}(G)=0$ if $D\subset G$ and $I_{0,0}(G)=\infty$ if $D\not\subset G$. In particular, for $D\not\subset G$ we have $I_{\epsilon_1,\epsilon_2}(G)=\infty$ for all*

$\epsilon_1, \epsilon_2 > 0$. *This proves in the limit* $\epsilon_1, \epsilon_2 \to 0$ *the convergence of the no response test for the multiwave case towards the true inclusion* D.

Proof. First, consider the case $D \subset G$. In this case, as shown in (5.16) for $\varphi \in \mathcal{M}(\epsilon_1)$ and $\alpha \in \mathcal{Q}(\epsilon_2, \varphi)$, we obtain

$$\begin{aligned}(6.1)\quad &\sum_{\xi=1}^{N} \alpha_\xi \int_{\partial\Omega} J(x,\xi)\varphi(x)ds(x) \\ &= \int_{\partial D} \left\{ v[\varphi](y) \left(\sum_{\xi} \alpha_\xi \frac{\partial u(y,\xi)}{\partial \nu} - \frac{\partial v[\varphi](y)}{\partial \nu} \right) \right. \\ &\qquad \left. - \frac{\partial v[\varphi](y)}{\partial \nu} \left(\sum_{\xi} \alpha_\xi u(y,\xi) - v[\varphi](y) \right) \right\} ds(y).\end{aligned}$$

For the multiwave version of the no response test according to (2.11) and (2.12) we have chosen the coefficients α_ξ such that the L^2-norm $\|w[\varphi]\|_{L^2(\partial\Omega)}$ of the scattered field

$$(6.2)\qquad w[\varphi] \;:=\; \sum_{\xi=1}^{N} \alpha_\xi u(\cdot,\xi) - v[\varphi]$$

is bounded by ϵ_2. By construction of the set $\mathcal{M}$ in (2.5) of Definition 2.1 and the Dirichlet boundary condition for u we further have

$$(6.3)\qquad \|w[\varphi]\|_{C(\partial D)} \le \epsilon_1, \quad \|v[\varphi]\|_{C(\partial D)} \le \epsilon_1, \quad \left\| \frac{\partial v[\varphi]}{\partial \nu} \right\|_{C(\partial D)} \le \epsilon_1.$$

By the continuity of the interior Dirichlet-to-Neumann map on ∂D in combination with classical regularity results for the normal derivative on a C^2-boundary ∂D, we obtain the estimate

$$(6.4)\qquad \left\| \frac{\partial w[\varphi]}{\partial \nu} \right\| = \left\| \sum_{\xi} \alpha_\xi \frac{\partial u(y,\xi)}{\partial \nu} - \frac{\partial v[\varphi](y)}{\partial \nu} \right\|_{C(\partial D)} \le c_1 \cdot (\epsilon_1 + \epsilon_2)$$

with some constant c_1. Thus, the functional (6.1) is bounded by

$$(6.5)\qquad c \cdot \epsilon_1 \cdot (\epsilon_1 + \epsilon_2)$$

with some constant c. We have proven that the functional $I_{\epsilon_1,\epsilon_2}(G)$ of the multiwave no response test (defined in (2.14)) is bounded. In the limit $\epsilon_1, \epsilon_2 \to 0$ we obtain convergence towards zero. This yields

$$(6.6)\qquad I_{0,0}(G) = 0.$$

Next, we assume that $D \not\subset G$. Then there is a point $z_0 \in \partial D \setminus \overline{G}$ located in the component of $\Omega \setminus (D \cup G)$ which is connected to $\partial\Omega$ and which satisfies

$$(6.7)\qquad \rho := \inf_{y \in \overline{G}} |y - z_0| > 0.$$

In this case for $z \in B_{\rho/2}(z_0) \cap (\Omega \setminus \overline{D})$ we use the construction worked out in section 5, parts II and III, but now for some test domain $M[z]$ which satisfies the conditions (1)

to (4) of section 4 and (5) $\overline{D \cup G} \subset M[z]$. This construction generates a sequence of incident fields v_n which approximate $\beta\Phi(\cdot, z)$ in $C^1(\overline{D} \cup \overline{G})$. The scattered field w_n converges towards the function defined in (5.11) and (5.12). Thus, using (5.20) we obtain

$$\begin{aligned} &\left| \sum_{\xi=1}^{N} \alpha_{\xi,n} \int_{\partial\Omega} J(x,\xi)\varphi_n(x) ds(x) \right| \\ &\to \beta \left| \int_{\partial D} \left\{ \Phi(y,z) \frac{\partial w(y,z)}{\partial \nu(y)} - \frac{\partial \Phi(y,z)}{\partial \nu(y)} w(y,z) \right\} ds(y) \right| \\ &= \beta \left| w(z,z) - \int_{\partial\Omega} \Phi(y,z) \frac{\partial w(y,z)}{\partial \nu(y)} ds(y) \right|. \end{aligned} \tag{6.8}$$

We remark that by Theorem 7.1 $w(z,z)$ is unbounded for $z \to z_0 \in \partial D$, but $\frac{\partial w(\cdot,z)}{\partial \nu}$ is bounded on $\partial\Omega$. We estimate the integral for $z \in B_{\rho/2}(z_0)$ by some constant $\tilde{c}$. For sufficiently large n this yields

$$\begin{aligned} &\left| \sum_{\xi=1}^{N} \alpha_{\xi,n} \int_{\partial\Omega} J(x,\xi)\varphi_n(x) ds(x) \right| \\ &\geq \frac{\beta}{2} \Big(\Big| w(z,z) \Big| - \tilde{c} \Big) \\ &\to \infty, \quad z \to z_0. \end{aligned} \tag{6.9}$$

Thus, we have constructed a sequence $\varphi_n \in \mathcal{M}(\epsilon_1)$ and $\alpha_n \in \mathcal{Q}(\epsilon_2, \varphi_n)$ such that $I(\varphi_n, \alpha_n)$ defined in (2.13) is unbounded. For the indicator function (2.14) this proves

$$I_{\epsilon_1,\epsilon_2}(G) = \infty \tag{6.10}$$

for all $\epsilon_1, \epsilon_2 > 0$ and thus $I_{0,0}(G) = \infty$. Hence, the proof is complete. ∎

7. Appendix. Here, our first goal is to analyze the behavior of a solution to the boundary value problem (5.11), (5.12) in the limit $z \to \partial D$.

THEOREM 7.1. *The solution $w(\cdot, z)$ of the Helmholtz equation in $\Omega \setminus \overline{D}$ with boundary values given by* (5.11), (5.12) *has the behavior*

$$|w(z,z)| \to \infty, \quad z \to \partial D. \tag{7.1}$$

Let M be a domain with $\overline{M} \subset \Omega \setminus \overline{D}$. Then we have

$$\left\| \frac{\partial w(\cdot, z)}{\partial \nu} \right\|_{C(\partial\Omega)} \leq C \tag{7.2}$$

uniformly for $z \in M$ with some constant C.

Proof. The proof is analogous to the case of obstacle scattering as worked out in detail in [12, Theorem 2.1.15 and Lemma 2.1.10]. ∎

Second, for the convenience of the reader we add some material about Taylor's series and analyticity of multidimensional functions.

LEMMA 7.2. *Consider the directional Taylor expansion*

$$u(x) = \sum_{\mu=0}^{\infty} \frac{1}{\mu!} ((x-z) \cdot \nabla)^{\mu} u(z). \tag{7.3}$$

If, for $|x-z|<\rho$, the series (3.3) *is absolutely convergent, then the function u is analytic in $B_\rho(z)$.*

Proof. First, recall the general form of the multidimensional Taylor series of u,

$$u(x)=\sum_{\xi=0}^{\infty} b_\xi (x-z)^\xi, \tag{7.4}$$

with the multi-index $\xi=(\xi_1,\dots,\xi_n)$, the coefficients

$$b_\xi=\frac{1}{\xi_1!\cdots\xi_n!}\cdot\frac{\partial^{\xi_1+\cdots+\xi_n}u(z)}{\partial z_1^{\xi_1}\cdots\partial z_n^{\xi_n}}, \tag{7.5}$$

and the multipower and multisum

$$z^\xi=z_1^{\xi_1}\cdot\ldots\cdot z_n^{\xi_n},\qquad \sum_{\xi=0}^{m}a_\xi:=\sum_{\xi_1=0}^{m_1}\cdots\sum_{\xi_n=0}^{m_n}a_{(\xi_1,\dots,\xi_n)} \tag{7.6}$$

for $m=(m_1,\dots,m_n)$. The form (7.4) is also the general local representation of an analytic function. We employ the standard formula

$$\begin{aligned}&\begin{pmatrix}\mu\\ k-1\end{pmatrix}+\begin{pmatrix}\mu\\ k\end{pmatrix}=\frac{\mu!}{(k-1)!(\mu-k+1)!}+\frac{\mu!}{k!(\mu-k)!}\\ &=\frac{\mu!}{k!(\mu-k+1)!}\{k+(\mu-k+1)\}=\frac{\mu!}{k!(\mu-k+1)!}(\mu+1)\\ &=\frac{(\mu+1)!}{k!(\mu+1-k)!}=\begin{pmatrix}\mu+1\\ k\end{pmatrix}.\end{aligned} \tag{7.7}$$

In the case of absolute convergence we are able to reorder the terms in (7.3) via the binomial formula, which in the two-dimensional case is given by

$$\Big((x_1-z_1)\partial_1+(x_2-z_2)\partial_2\Big)^\mu=\sum_{k=0}^{\mu}\begin{pmatrix}\mu\\ k\end{pmatrix}(x_1-z_1)^k(x_2-z_2)^{\mu-k}\partial_1^k\partial_2^{\mu-k}. \tag{7.8}$$

We obtain

$$\begin{aligned}&\sum_{\mu=0}^{\infty}\frac{1}{\mu!}((x-z)\cdot\nabla)^\mu u(z)\\ &=\sum_{\mu=0}^{\infty}\frac{1}{\mu!}\sum_{k=0}^{\mu}\begin{pmatrix}\mu\\ k\end{pmatrix}(x_1-z_1)^k(x_2-z_2)^{\mu-k}\partial_1^k\partial_2^{\mu-k}u(z)\\ &=\sum_{\xi_1=0}^{\infty}\sum_{\xi_2=0}^{\infty}\frac{1}{\xi_1!\xi_2!}(x_1-z_1)^{\xi_1}(x_2-z_2)^{\xi_2}\partial_1^{\xi_1}\partial_2^{\xi_2}u(z)\\ &=\sum_{\xi=0}^{\infty}b_\xi(x-z)^\xi,\end{aligned} \tag{7.9}$$

where we replaced $k=\xi_1$ and $\mu-k=\xi_2$ and observe that when $\mu=0$ to ∞ and $k=0$ to μ the variables ξ_1 and ξ_2 run through all numbers between 0 and ∞ such that each point $(\xi_1,\xi_2)\in\mathbb{N}_0\times\mathbb{N}_0$ is hit once. This proves analyticity of the function u. A proof for higher n is carried out analogously. ∎

Acknowledgments. For the completion of the proofs, discussions with Prof. Dr. Gen Nakamura and Dr. Mourad Sini, Hokkaido University, Japan, have been very important. The collaboration and interaction was a pleasure. Further, I would like to thank Prof. Dr. Simon Chandler-Wilde, Reading University, UK, and Prof. Dr. Rainer Kreß, University of Göttingen, for helpful discussions and remarks and Dr. Klaus Erhard and Dipl.-Phys. Jochen Schulz, Göttingen, for carefully reading the manuscript.

REFERENCES

[1] J. Cheng, J. Liu, and G. Nakamura, *The numerical realization of the probe method for the inverse scattering problems from the near-field data*, Inverse Problems, 21 (2005), pp. 839–855.

[2] K. Erhard and R. Potthast, *A numerical study of the probe method*, SIAM J. Sci. Comput., 28 (2006), pp. 1597–1612.

[3] K. Erhard and R. Potthast, *The point source method for reconstructing an inclusion from boundary measurements in electrical impedance tomography and acoustic scattering*, Inverse Problems, 19 (2003), pp. 1139–1157.

[4] N. Honda, G. Nakamura, M. Sini, and R. Potthast, *The no-response approach and its relation to non-iterative methods for the inverse scattering*, Ann. Mat. Pura Appl. (4), to appear.

[5] M. Ikehata, *Reconstruction of the shape of the inclusion by boundary measurements*, Comm. Partial Differential Equations, 23 (1998), pp. 1459–1474.

[6] V. Isakov, *On uniqueness in the inverse transmission scattering problem*, Comm. Partial Differential Equations, 15 (1990), pp. 1565–1587.

[7] A. Kirsch and R. Kress, *Uniqueness in inverse obstacle scattering*, Inverse Problems, 9 (1993), pp. 285–299.

[8] D. R. Luke and R. Potthast, *The no response test—a sampling method for inverse scattering problems*, SIAM J. Appl. Math., 63 (2003), pp. 1292–1312.

[9] C. B. Morrey, Jr., *Multiple Integrals in the Calculus of Variations*, Grundlehren Math. Wiss. 130, Springer-Verlag, New York, 1966.

[10] R. Potthast, *A fast new method to solve inverse scattering problems*, Inverse Problems, 12 (1996), pp. 731–742.

[11] R. Potthast, *Stability estimates and reconstructions in inverse acoustic scattering using singular sources*, J. Comput. Appl. Math., 114 (2000), pp. 247–274.

[12] R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, Chapman & Hall/CRC Res. Notes Math. 427, Chapman & Hall/CRC, Boca Raton, FL, 2001.

[13] R. Potthast, J. Sylvester, and S. Kusiak, *A "range test" for determining scatterers with unknown physical properties*, Inverse Problems, 19 (2003), pp. 533–547.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1825–1846

GLOBAL WEAK SOLUTIONS TO EQUATIONS OF COMPRESSIBLE MISCIBLE FLOWS IN POROUS MEDIA*

Y. AMIRAT† AND V. SHELUKHIN‡

Abstract. We study the one-dimensional equations governing compressible flows of m miscible components in a porous medium. The equations are reduced to a quasi-linear parabolic system for the discharge function P and the concentrations c_i. The equations of this system are strongly coupled since the parabolic equation for c_i contains both the second derivative c_{ixx} and the second derivative P_{xx}. We prove global weak solvability of an initial boundary-value problem both in the Eulerian and Lagrangian formulations.

Key words. porous media, compressible miscible flows, existence

AMS subject classifications. Primary, 35K55; Secondary, 35B65, 76S05, 76T05

DOI. 10.1137/050640321

1. Introduction. We address the question of global solvability of the one-dimensional equations which govern flows of m miscible components in porous media. For the reader's convenience, we restate the basic three-dimensional model [3], [10], [17].

The seepage velocity $\mathbf{u}(\mathbf{x}, t)$ obeys the Darcy law

$$\mathbf{u} = -\frac{k}{\mu}(\nabla p - \rho g \nabla z). \tag{1.1}$$

Here, p is the mixture pressure, ρ is the mixture density, z is the vertical component of the space-point vector $\mathbf{x}$ (we assume that z grows in the direction of the gravitation acceleration vector $\mathbf{g}$), μ is the viscosity, and k is the permeability.

When the fluid is slightly compressible, pressure and density are related by the state equation

$$\frac{d\rho}{\rho} = \nu dp,$$

where ν is the compressibility constant. Given some reference conditions, one can write the state equation as

$$\rho = \rho_r \exp(\nu(p - p_r)), \quad p_r = \text{const}, \quad \rho_r = \text{const} > 0,$$

or equivalently

$$p = \frac{1}{\nu} \ln \frac{\rho}{\kappa} \text{ with } \kappa = \rho_r e^{-\nu p_r}. \tag{1.2}$$

*Received by the editors September 14, 2005; accepted for publication (in revised form) September 5, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/sima/38-6/64032.html
†Laboratoire de Mathématiques, CNRS (UMR 6620), Université Blaise Pascal (Clermont-Ferrand 2), 63177 Aubière cedex, France (Youcef.Amirat@math.univ-bpclermont.fr).
‡Lavrentyev Institute of Hydrodynamics, Lavrentyev pr. 15, Novosibirsk, 630090, Russia (shelukhin@hydro.nsc.ru). The research of this author was partially supported by CNRS, UMR 6620, and by Russian Fund for Fundamental Research grant 05-01-00131.

The mixture mass conservation law is written as

$$\frac{\partial \Phi \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) = 0, \quad 0 \le \Phi \le 1, \tag{1.3}$$

where Φ is the porosity, the fraction of volume occupied by the fluid.

If chemical reactions do not occur, the change of volume concentration c_i of the ith component of the fluid mixture is described by the equation

$$\frac{\partial \Phi \rho c_i}{\partial t} + \operatorname{div} \mathbf{q}_i = 0, \quad (\mathbf{q}_i)_j \equiv \rho c_i u_j - \rho D_{jk} \frac{\partial c_i}{\partial x_k}, \quad i = 1, \dots, m, \quad j = 1, 2, 3. \tag{1.4}$$

The diffusion-dispersion tensor D is given by the formula

$$D_{ij} = \Phi d_m \delta_{ij} + \Phi |\mathbf{u}| \left(d_e \frac{u_i u_j}{|\mathbf{u}|^2} + d_t \left(1 - \frac{u_i u_j}{|\mathbf{u}|^2} \right) \right),$$

where d_m is the diffusion, and d_e and d_t are the dispersion coefficients.

Equations (1.1)–(1.3) cannot be solved independently of (1.4) because of the constitutive law

$$\mu = \mu(c_1, c_2, \dots, c_m). \tag{1.5}$$

The parameters $k, \nu, \Phi, d_m, d_e, d_t$ are normally assumed to be positive constants.

Due to the definition of c_i, equations (1.4) are coupled by the restrictions

$$0 \le c_i, \quad \sum_1^m c_j = 1. \tag{1.6}$$

Note that because of equality (1.6), the equation for c_m in (1.4) is a consequence of (1.3) and of the first $m-1$ equations in (1.4).

Equations (1.1)–(1.6) have many applications and serve as a basis for the computer modeling of petroleum reservoir flows, underground contamination problems, etc. [5, 6]. As for the incompressible flows, we refer the reader to the results in [4], [8], [7], [15].

One-dimensional vertical flows equations in the domain $|x| < 1$ reduce to the system

$$(\Phi \rho)_t + (\rho u)_x = 0, \quad u = -\lambda(p_x - g\rho), \quad \lambda \equiv \frac{k}{\mu}, \quad p = \frac{1}{\nu} \ln \frac{\rho}{\kappa}, \tag{1.7}$$

$$(\Phi \rho c_i)_t + (\rho c_i u)_x = (\rho D(u) c_{ix})_x, \quad D(u) \equiv \Phi(d_m + d_p |u|), \quad 1 \le i \le m-1, \tag{1.8}$$

with the restrictions

$$0 \le c_i, \quad \sum_1^{m-1} c_j \le 1. \tag{1.9}$$

Global existence of this system was studied in [2] under the Douglas–Roberts [5] assumption that the term $\rho_x c_i u$ in each equation of system (1.8) is negligible and the derivative $(\rho c_i u)_x$ can be substituted with the sum $\rho c_{ix} u + \rho c_i u_x$. Here, we perform the analysis of system (1.7)–(1.9) without any smallness hypothesis. All the mathematical difficulties to overcome stem from the fact that the vector

$$\mathbf{v} := (P, c_1, c_2, \dots, c_{m-1})^T,$$

with P standing for the fluid discharge, solves the quasi-linear parabolic system (2.1)–(2.2) in the form

$$\mathbf{v}_t = A(\mathbf{v}, \mathbf{v}_x)\mathbf{v}_{xx} + \mathbf{f}[x, \mathbf{v}, \mathbf{v}_x], \tag{1.10}$$

with a triangular $m \times m$-matrix A and with $\mathbf{f}$ being a nonlocal term containing the integral $\int_{-1}^{x} \exp(\nu P(y,t))dy$. Triangularity means that the parabolic equation for P does not contain the second derivative $\partial_{xx}c_i$, but the parabolic equation for c_i contains both the second derivatives $\partial_{xx}c_i$ and $\partial_{xx}P$. For system (1.10), a global existence theory is not yet developed. The theory of Ladyzhenskaya, Solonnikov, and Ural'ceva [13] can be applied to system (1.10) provided that A is diagonal and $A_{11} = A_{22} = \cdots = A_{mm}$. The results of Amann [1] concern the triangular matrix A, but they are valid under the assumption that some a priori estimates hold. Interestingly, triangular parabolic systems

$$\mathbf{v}_t = (B(\mathbf{v})\mathbf{v}_x)_x + (\mathbf{f}(\mathbf{v}))_x \tag{1.11}$$

arise also in the theory of three-phase capillary immiscible flows in porous media [9]. But the results obtained for system (1.11) are not applicable here, since the matrix B does not depend on $\mathbf{v}_x$.

We study system (1.7)–(1.9) in the domain

$$Q = \Omega \times (0,T), \quad \Omega = \{x : |x| < 1\}.$$

The initial and boundary-value conditions are

$$p|_{t=0} = p_0, \quad c_i|_{t=0} = c_i^0, \tag{1.12}$$

$$u|_{|x|=1} = 0, \quad \partial_x c_i|_{|x|=1} = 0, \quad 1 \le i \le m-1. \tag{1.13}$$

We denote $\rho_0 = \rho_r \exp(\nu(p_0 - p_r))$.

We assume that the mobility function $\lambda(\mathbf{c})$, $\mathbf{c} := (c_1, c_2, \dots, c_{m-1})^T$, is defined not only in the domain (1.9) but in the whole space $\mathbb{R}^{m-1}$ and

$$\lambda \in C^2(\mathbb{R}^{m-1}), \quad 0 < \lambda_0 \le \lambda(\mathbf{c}) \le \lambda_0^{-1}. \tag{1.14}$$

In what follows, we use the Sobolev spaces $W^{k,p}(\Omega)$, $W_p^{k,l}(Q)$, the Orlicz and Sobolev–Orlicz spaces $L_M(\Omega)$, $W_M^1(\Omega)$ (associated with an admissible convex function M), and the Hölder spaces $H^{k+\alpha,(k+\alpha)/2}(\bar{Q})$ [12], [11], [13]; the corresponding definitions will be restated below.

We define a *weak solution* $(u, p, \rho, \mathbf{c})$ of problem (1.7)–(1.9), (1.12), (1.13) as follows:

(i) $u \in L^\infty(0,T; L_M(\Omega)) \cap L^{3/2}(0,T; W^{1,3/2}(\Omega)) \cap L^3(Q)$;
$p, \rho \in L^\infty(0,T; W_M^1(\Omega)) \cap L^{3/2}(0,T; W^{2,3/2}(\Omega))$
$\cap L^3(0,T; W^{1,3}(\Omega)) \cap L^\infty(Q)$; $p_t, \rho_t \in L^{3/2}(Q)$;
$\mathbf{c} \in L^\infty(0,T; L^2(\Omega)) \cap L^2(0,T; W^{1,2}(\Omega))$
with $M(s) = |s|\ln(1+|s|) - |s| + \ln(1+|s|)$;

(ii) the equations

$$(\Phi\rho)_t + (\rho u)_x = 0, \quad u = -\lambda(\mathbf{c})(p_x - g\rho), \quad p = \frac{1}{\nu}\ln\frac{\rho}{\kappa},$$

hold a.e. in Q;

(iii) equations (1.8) hold weakly in the following sense:

$$\iint_{\bar{Q}} \rho\mathbf{c}(\Phi\psi_t + u\psi_x) - \rho D(u)\mathbf{c}_x\psi_x \, dxdt = -\int_\Omega \Phi\rho_0(x)\,\mathbf{c}_0(x)\psi(x,0)dx$$

for any $\psi \in C^1(\bar{Q})$ such that $\psi(x,T) = 0$.

We formulate the main result.

THEOREM 1.1. *Assume that* $p_0 \in W^{2,2}(\Omega)$, $\mathbf{c}_0 \in L^2(\Omega)$ *and that the function* $\lambda(\mathbf{c})$ *satisfies restrictions* (1.14). *Then, problem* (1.7)–(1.9), (1.12), (1.13) *has a weak solution such that* $\inf_Q \rho > 0$.

Observe that restrictions (1.9) are not included in the above definition. To justify the notion of weak solution we establish the following claim.

THEOREM 1.2. *If the weak solution* $(u, p, \rho, \mathbf{c})$ *is regular enough, it satisfies restrictions* (1.9), *provided that they are valid for the initial data* $\mathbf{c}_0$.

We comment on the proofs. The first step is the reduction of the basic equations (1.7)–(1.8) to the parabolic system

$$\Phi P_t - \lambda(\mathbf{c})|P_x|^2 = \left(\frac{\lambda(\mathbf{c})}{\nu} P_x\right) x, \tag{1.15}$$

$$\Phi \mathbf{c}_t - \lambda(\mathbf{c}) P_x \mathbf{c}_x = \Phi[(d_m + d_p|\lambda(\mathbf{c})P_x|)\mathbf{c}_x]_x + \Phi\Big(d_m + d_p|\lambda(\mathbf{c})P_x|\Big)\mathbf{c}_x \rho_x \rho^{-1}, \tag{1.16}$$

where P is the discharge function, related to the velocity u by the Darcy law $u = -\lambda(\mathbf{c})P_x$; the density ρ is given by the representation formula

$$\rho(x,t) = \kappa e^{\nu P(x,t)} \Big/ \Big(1 - \nu\kappa g \int_{-1}^{x} e^{\nu P(y,t)} dy\Big).$$

This reduction has a twofold meaning. On the one hand, it enables us to derive a number of a priori estimates like the maximum principles for P and c_i. Particularly, the claim of Theorem 1.2 follows immediately from (1.16). On the other hand, one can see that the basic equations are as difficult as the strongly coupled parabolic system (1.15)–(1.16), with (1.16) for $\mathbf{c}$ containing both the second derivative $\mathbf{c}_{xx}$ and the second derivative P_{xx}. Moreover, it becomes clear that it is impossible to obtain a regular solution without an a priori estimate

$$\|u\|_{C^0(\overline{Q})} \overset{?}{\leq} \text{const} \tag{1.17}$$

according to the existing theory of parabolic equations (see, for a example, [13]).

The a priori estimates obtained are those which correspond to item (i) of the definition of weak solution. They do not guarantee the bound (1.17); this is why our results concern only the weak solutions.

Starting from system (1.15)–(1.16), we construct approximate solutions and study their compactness.

As follows from a priori estimates, the last term in (1.16) belongs to only $L^1(Q)$. Thus, one cannot substitute this equation by Galerkin approximation equations. In this sense, (1.8) is better. But there is an obstacle in the study of the system

$$\Phi P_t - \lambda(\mathbf{c})|P_x|^2 = \left(\frac{\lambda(\mathbf{c})}{\nu} P_x\right) x, \quad (\Phi\rho\mathbf{c})_t + (\rho\mathbf{c}u)_x = (\rho D(u)\mathbf{c}_x)_x. \tag{1.18}$$

Whatever approximation $(P_n, \mathbf{c}_n)$, the functions $\mathbf{c}_n$ should converge a.e. in Q in order to pass to the limit in the equation $u_n = -\lambda(\mathbf{c}_n)P_{nx}$. With the estimate $\|\mathbf{c}_n\|_{L^2(0,T;W^{1,2}(\Omega))} \leq b$ at hand, one could prove the convergence of $\mathbf{c}_n$ a.e. in Q by the Aubin–Lions compactness theorem [14] if there were an estimate of $\mathbf{c}_{nt}$ in some weak norm. From (1.8), one can obtain such an estimate for the time derivative

$(\Phi\rho_n\mathbf{c}_n)_t$, but not for the time derivative $\mathbf{c}_{nt}$. Moreover, the presence of the density ρ makes (1.8) a degenerate parabolic equation for $\mathbf{c}$, and the problem of existence of Galerkin approximations also becomes more complicated.

To avoid these technical difficulties, we propose another approach. Its main ingredient is the passage to the Lagrangian variables $(x,t) \to (\xi,t)$ by the formulas (2.6) and (2.7) below. In the new variables, system (1.18) reads

$$\nu P_t = \Phi\rho(\lambda(\mathbf{c})\rho P_\xi)_\xi, \quad \mathbf{c}_t = \Phi(\rho^2 D\mathbf{c}_\xi)_\xi, \quad \rho = \kappa e^{\nu P + \nu g(1+\xi)/\Phi}. \tag{1.19}$$

We define approximate solutions $(P_n(\xi,t), \mathbf{c}_n(\xi,t))$, by the semi-Galerkin method, in the sense that P_n and $\mathbf{c}_n$ solve the first equation in (1.19) exactly and solve the second equation approximately (see section 4).

It is an advantage of system (1.19) that P_n and $\mathbf{c}_n$ enjoy the same a priori estimates (except (1.9)) as the solution $(P(\xi,t), \mathbf{c}(\xi,t))$ itself, including the estimates for the time derivatives of P_n and $\mathbf{c}_n$. One more advantage is the straightforward estimate $\inf_Q \rho > 0$.

Thus, we first prove global solvability of system (1.19) (Theorem 5.1), equipped with the corresponding initial and boundary conditions. Then, the same result for system (1.18) (Theorem 5.2) is obtained by the inverse change of variables $(\xi,t) \to (x,t)$. Theorem 1.1 is an easy consequence of Theorem 5.2.

Though any regular solution of (1.19) satisfies the restrictions (1.9), it is not true for the approximate concentrations $\mathbf{c}_n$ since they do not exactly solve the second equation in (1.19). Due to the technique developed in [16], there is a possibility of deriving properties (1.9) for the weak solutions using only the weak formulation (see item (iii) in the definition of weak solution). But to apply this technique one needs estimate (1.17). Thus, the validity of properties (1.9) for the weak solution is an open problem.

2. Different settings of the problem. First, we show that the basic flow equations can be formulated as a quasi-linear parabolic system. We introduce the fluid discharge function P as follows:

$$P := p - \frac{g(1+\xi)}{\Phi}, \quad \xi(x,t) := \Phi\int_{-1}^{x} \rho(y,t)dy - 1.$$

Here, ξ is the mass Lagrange variable. For simplicity we assume that

$$\Phi\int_{-1}^{1} \rho_0(x)dx = 2.$$

Due to the mass conservation law and the boundary condition $u|_{|x|=1} = 0$, we have

$$\Phi\int_{-1}^{1} \rho(x,t)dx = 2.$$

Hence, $\xi(\cdot,t)$ maps Ω onto itself, and $\xi(\pm1,t) = \pm1$.

By (1.7)–(1.8) and the formulas

$$u = -\lambda P_x, \quad \rho_x = \nu\rho p_x, \quad p_x = P_x + g\rho, \quad p_t = P_t + \Phi^{-1} g\,\lambda\rho P_x,$$

one can verify that P and $\mathbf{c}$ solve the parabolic system

$$\Phi P_t - \lambda(\mathbf{c})P_x^2 = \left(\frac{\lambda(\mathbf{c})}{\nu}P_x\right)x, \tag{2.1}$$

$$\Phi\mathbf{c}_t + u\mathbf{c}_x = (D(u)\mathbf{c}_x)_x + D(u)\mathbf{c}_x\rho_x\rho^{-1}, \tag{2.2}$$

with u, ρ, and D given by the representation formulas

$$u = -\lambda(\mathbf{c})P_x, \quad \rho(x,t) = \frac{\kappa e^{\nu P(x,t)}}{1 - \nu\kappa g \int_{-1}^{x} e^{\nu P(y,t)} dy}, \quad D = \Phi(d_m + d_p|u|). \tag{2.3}$$

The boundary and initial conditions are

$$\begin{aligned} &P_x|_{|x|=1} = 0, \quad \mathbf{c}_x|_{|x|=1} = 0, \\ &P|_{t=0} = p_0 - g\int_{-1}^{x} \rho_0(y)dy, \quad \mathbf{c}|_{t=0} = \mathbf{c}_0(x). \end{aligned} \tag{2.4}$$

All the formulas (2.1)–(2.4) can be obtained by simple calculations. We comment only on the derivation of the representation formula for the density which is the second one in (2.3). Since $\xi_x = \Phi\rho$, one can write the definition of P as

$$g\kappa\nu e^{\nu P} = \nu g\rho e^{-\nu g(\xi+1)/\Phi} = -e^{-\nu g/\Phi}\frac{\partial}{\partial x}e^{-\nu g\xi/\Phi}. \tag{2.5}$$

Then we integrate (2.5) with respect to the x variable to obtain

$$1 - g\kappa\nu \int_{-1}^{x} e^{\nu P(y,t)} dy = e^{-\nu g(\xi+1)/\Phi}.$$

Now the representation formula for density results from this relation and the first equality in (2.5).

In what follows we use a Lagrange formulation of the problem. The change of variables $(x,t) \to (\xi,t)$ maps the cylinder $Q = \Omega \times (0,T)$ onto itself and

$$\xi_x = \Phi\rho, \quad \xi_t = -\rho u, \quad \xi|_{x=\pm 1} = \pm 1. \tag{2.6}$$

Observe that the Jacobian of the transformation is equal to $\Phi\rho$. The inverse transformation $(\xi,t) \to (x,t)$ satisfies the equalities

$$\Phi\rho x_\xi = 1, \quad \Phi x_t = u, \quad x|_{\xi=\pm 1} = \pm 1. \tag{2.7}$$

Let $\tilde{f}(\xi,t)$ stand for the transformation of a given function $f(x,t)$:

$$\tilde{f}(\xi,t) = f(x(\xi,t),t).$$

By formulas (2.6) and (2.7), we have

$$\tilde{f_t} = f_t + uf_x/\Phi, \quad \tilde{f_\xi} = \frac{f_x}{\rho\Phi}, \quad x(\xi,t) = \int_{-1}^{\xi} \frac{1}{\tilde{\rho}(y,t)\Phi}\, dy - 1.$$

Now, omitting "˜" one can write the flow equations in the Lagrange variables as

$$\begin{aligned} &(\rho^{-1})_t = u_\xi, \quad u = -\Phi\lambda(\mathbf{c})\rho P_\xi, \quad P = p - \frac{g(1+\xi)}{\Phi}, \quad p = \frac{1}{\nu}\ln\frac{\rho}{\kappa}, \\ &\mathbf{c}_t = \Phi(\rho^2 D\mathbf{c}_\xi)_\xi, \quad D = \Phi(d_m + d_p|u|), \quad u|_{|\xi|=1} = 0. \end{aligned}$$

The parabolic system (2.1), (2.2) is written as

$$\nu P_t = \Phi\rho(\lambda(\mathbf{c})\rho P_\xi)_\xi, \quad \mathbf{c}_t = \Phi(\rho^2 D(u)\mathbf{c}_\xi)_\xi, \tag{2.8}$$

with

$$\rho = \kappa e^{\nu P + \nu g(1+\xi)/\Phi}, \quad D(u) = \Phi(d_m + d_p|u|), \quad u = -\Phi\lambda(\mathbf{c})\rho P_\xi. \tag{2.9}$$

The boundary and initial conditions are

$$\begin{aligned} &P_\xi|_{|\xi|=1} = 0, \quad \mathbf{c}_\xi|_{|\xi|=1} = 0, \\ &P|_{t=0} = P_0(x(\xi,0)) \equiv \tilde{P}_0(\xi), \quad \mathbf{c}|_{t=0} = \mathbf{c}_0(x(\xi,0)) \equiv \tilde{\mathbf{c}}_0(\xi), \end{aligned} \tag{2.10}$$

where the function $x(\xi, 0)$ can be calculated via the equality

$$\xi = \Phi \int_{-1}^{x(\xi,0)} \rho_0(y)dy - 1.$$

Observe that system (2.8) can be written in a conservative form:

$$v_t = \Phi\left(\frac{\lambda(\mathbf{c})v_\xi}{\nu\, v^2}\right)\xi + \left(\frac{g\lambda(\mathbf{c})}{v}\right)\xi, \quad \mathbf{c}_t = \Phi\left(\frac{D\mathbf{c}_\xi}{v^2}\right)\xi \tag{2.11}$$

by passing to the variables v and $\mathbf{c}$, where $v = 1/\rho$ is the specific volume.

Thus, starting from the flow equations, we have obtained three parabolic systems. We claim that each of these systems is equivalent to the basic flow equations. Let us verify that claim for system (2.1), (2.2). The verification for the other systems is similar.

It is enough to show that, given a solution $(P, \mathbf{c})$ of (2.1), (2.2), the functions ρ and u, defined by (2.3), solve the equations

$$\Phi\rho_t + (\rho u)_x = 0, \quad u = -\lambda(p_x - g\rho), \quad p = \frac{1}{\nu}\ln\frac{\rho}{\kappa}. \tag{2.12}$$

By definition (2.3) of u, it follows from (2.1) that

$$\Phi P_t + uP_x + u_x/\nu = 0. \tag{2.13}$$

Hence,

$$\Phi(e^{\nu P})_t + (ue^{\nu P})_x = 0.$$

Denoting $R = \int_{-1}^{x} e^{\nu P(y,t)}dy$, we see that R solves the transport equation

$$\Phi R_t + uR_x = 0 \tag{2.14}$$

and is linked with ρ by the equality

$$\rho = -\frac{1}{g\nu}\frac{\partial}{\partial x}\ln(1 - \kappa\nu g R).$$

Integrating the latter equation with respect to the x variable, we obtain that

$$1 - \kappa\nu g R = e^{-\nu g z}, \quad z := \int_{-1}^{x} \rho(y,t)dy.$$

Thus, we have proved that

$$P = \frac{1}{\nu}\ln\frac{\rho}{\kappa} - g\int_{-1}^{x}\rho(y,t)dy \quad \text{or} \quad \rho = \kappa e^{\nu P + \nu g z}.$$

We calculate

$$e^{\nu g z}\Big(\Phi\rho_t + (\rho u)_x\Big) = \kappa\nu e^{\nu g z}(\Phi P_t + uP_x + u_x/\nu) + \nu g z e^{\nu g z}(\Phi z_t + u z_x).$$

Observe that z, being a function of R, solves the transport equation (2.14). Hence, the first equation in (2.12) results from (2.13).

3. A priori estimates. In what follows, the norms in $L^q(\Omega)$ and $L^r(0,T;L^q(\Omega))$ are denoted by $\|\cdot\|_q$ and $\|\cdot\|_{r,q,}$, respectively.

We assume that the solutions to the basic flow equations are regular enough. It permits us to use any of the parabolic systems described above. If the change of variables $(x,t) \to (\xi,t)$ is regular, the solvability of the problem in the Euler setting (system (2.1)–(2.4)) implies the solvability of the problem in the Lagrange setting (system (2.8)–(2.10)), and vice versa. But it does not mean that the study of one system is as easy as the study of the other. Since the change of variables depends on the solution, systems (2.1)–(2.4) and (2.8)–(2.10) are completely different nonlinear systems. In fact, we do not tackle system (2.1)–(2.4) directly. First, we prove the solvability of system (2.8)–(2.10) and then we do it for system (2.1)–(2.4) via the change of variables. The main reason is that approximate solutions of the Eulerian system do not enjoy as many estimates as those of the Lagrangian system.

First, we consider the problem in the Lagrange setting.

LEMMA 3.1. *The discharge function satisfies the maximum principle*

$$\underline{P} \equiv \inf_{\Omega} \tilde{P}_0(\xi) \le P(\xi,t) \le \sup_{\Omega} \tilde{P}_0(\xi) \equiv \overline{P} \quad \forall (\xi,t) \in Q. \tag{3.1}$$

Proof. Given a function $F(P)$, $F \in C^2_{loc}(\mathbb{R})$, we multiply the first equation in (2.8) by $vF'(P)$, $v := \rho^{-1}$, and integrate the result over the domain $\Omega \times (0,\tau)$, $0 < \tau \le T$, to arrive at the equality

$$\int_\Omega v(\xi,\tau)F(P(\xi,\tau))d\xi + \frac{\Phi}{\nu}\int_0^\tau \int_\Omega \lambda\rho P_\xi^2(F'' - \nu F')d\xi dt = \int_\Omega v_0 F(P_0)d\xi. \tag{3.2}$$

Here, we have used the equation $v_t = u_\xi$. The continuous function

$$F_1(P) = \begin{cases} 0, & P < \overline{P} + \varepsilon, \\ \big(e^{\nu(P-\overline{P}-\varepsilon)} - 1\big)/\nu, & P \ge \overline{P} + \varepsilon, \end{cases}$$

solves the equation $F''(P) - \nu F'(P) = 0$ everywhere except at the point $P = \overline{P} + \varepsilon$.

For any $\delta > 0$, let $\theta_\delta(r)$ be a standard mollifier:

$$\theta_\delta(r) = \theta(r/\delta)/\delta, \quad \theta \in C^\infty(\mathbb{R}), \quad \theta \ge 0, \quad \operatorname{supp}\theta \subset [-1,1], \quad \int_{\mathbb{R}} \theta(r)dr = 1.$$

Clearly, the function $F_{1\delta} = F_1 * \theta_\delta$ enjoys the properties

$$F_{1\delta} \in C^\infty(\mathbb{R}),$$
$$F_{1\delta} \to F_1 \quad \text{in} \quad C_{\text{loc}}(\mathbb{R}), \quad \text{and} \quad F''_{1\delta}(P) - \nu F'_{1\delta}(P) \to 0 \quad \text{if} \quad P \ne \overline{P} + \varepsilon.$$

Let us take $F(P) = F_{1\delta}(P)$ in (3.2) and send δ to 0. Observe that $P_\xi(\xi,t) = 0$ a.e. at the set where $P(\xi,t) = \overline{P} + \varepsilon$. Hence, by the Lebesgue theorem,

$$\int_0^\tau \int_\Omega \lambda\rho P_\xi^2 \Big(F''_{1\delta}(P) - \nu F'_{1\delta}(P)\Big)d\xi dt \to 0 \quad \text{as} \quad \delta \to 0.$$

On the other hand, by the choice of F_1,

$$\int_\Omega v(\xi,\tau)F_{1\delta}(P(\xi,\tau))d\xi \to \int_\Omega v(\xi,\tau)F_1(P(x,\tau))d\xi, \quad \int_\Omega v_0 F_{1\delta}(P_0)d\xi \to 0,$$

as $\delta \to 0$. Thus, $\int_\Omega v(\xi,\tau)F_1(P(\xi,\tau))d\xi = 0$ for any $\tau \geq 0$. This implies that $P \leq \overline{P} + \varepsilon$. Since ε is arbitrary, the estimate $P \leq \overline{P}$ follows.

Next, we take the function $F(P)$ in (3.2) equal to $F_2(P)$, with F_2 such that

$$F_2 \in C^2_{\rm loc}(\mathbb{R}), \quad F_2'' \geq 0, \quad F_2' \leq 0, \quad F_2(P) = 0 \quad \text{if} \quad P \leq \underline{P}.$$

Due to this choice, we obtain $P \geq \underline{P}$. Estimates (3.1) are established. ∎

Consequence. It follows from the density representation formula (2.9) that

$$\underline{\rho} \equiv \kappa e^{\nu \underline{P}} \leq \rho \leq \kappa e^{\nu \overline{P} + 2\nu g/\Phi} \equiv \overline{\rho}.$$

LEMMA 3.2. *The function P_ξ satisfies the estimate*

$$\sup_{t\geq 0} \int_\Omega |P_\xi| d\xi \leq \int_\Omega |\tilde{P}_{0\xi}| d\xi. \tag{3.3}$$

Proof. Denoting $z = P_\xi$, we obtain from (2.8) the following equation for z:

$$\nu z_t = (\Phi \rho(\lambda\rho z)_\xi)_\xi.$$

Given $\delta > 0$, we multiply this equation by $M_1'(\lambda\rho z)$, where $M_1(s) = (s^2+\delta^2)^{1/2}$ is a smooth approximation of $|s|$, and integrate the result over $\Omega \times (0,t)$. We obtain

$$\int_0^t \int_\Omega \frac{\nu z_t \lambda\rho z}{((\lambda\rho z)^2+\delta^2)^{1/2}} + \frac{\delta^2\, \Phi\rho |(\lambda\rho z)_\xi|^2}{((\lambda\rho z)^2+\delta^2)^{3/2})}\, d\xi d\tau = 0.$$

In the limit as $\delta \to 0$ we have

$$\int_0^t \int_\Omega z_t \,\mathrm{sign}(\lambda\rho z)\, d\xi d\tau = \int_0^t \int_\Omega z_t \,\mathrm{sign}(z)\, d\xi d\tau = \int_0^t \frac{d}{d\tau} \int_\Omega |z|\, d\xi d\tau \leq 0,$$

from which follows (3.3). ∎

LEMMA 3.3. *The function* $\mathbf{c}$ *satisfies the estimate*

$$\sup_{t\geq 0} \int_\Omega |\mathbf{c}|^2 d\xi + \Phi \int_0^t \int_\Omega \rho^2 D |\mathbf{c}_\xi|^2 d\xi d\tau \leq \int_\Omega |\mathbf{c}_0|^2 d\xi.$$

Proof. It follows immediately if one multiplies the second equation in (2.8) by $\mathbf{c}$ and integrate the result over $\Omega \times (0,t)$. ∎

Let us now introduce the convex function

$$M(U) = |U| \ln(1+|U|) - |U| + \ln(1+|U|). \tag{3.4}$$

One can easily verify that

$$M'(U) = \ln(1+|U|)\,\mathrm{sign} U, \quad M''(U) = (1+|U|)^{-1}.$$

Below, b (with or without a subscript) indicates a generic constant, depending only on some bounds of the initial data, which will probably take on different values in different occurrences.

We have the following result.

LEMMA 3.4. *The function*

$$U = -\Phi\rho P_\xi$$

satisfies the estimates

$$\|U\|_{L^\infty(0,T;L_M(\Omega))} \le b, \quad \|U\|_{3,3} \le b,$$
$$\|U_\xi\|_{3/2,3/2} \le b, \quad \|U\|_{2,\infty} \le b,$$

where $L_M(\Omega)$ denotes the Orlicz space associated with the function M defined by (3.4)*, and b is a constant depending only on Φ, ν, g, $\underline{\rho}$, $\overline{\rho}$, and the function λ.*

Proof. By differentiation with respect to ξ, it follows from the first equation in (2.8) that the function U solves the equation

$$U_t + \rho U(\lambda U)_\xi = \Phi\nu^{-1}\rho\,(\rho(\lambda U)_\xi)_\xi. \tag{3.5}$$

We multiply (3.5) by $vM'(U)$, where $v = 1/\rho$, and integrate the result over Ω using the equalities $v_t = u_\xi$ and $u = \lambda U$. This gives

$$\frac{d}{dt}\int_\Omega vM(U)\,d\xi = \int_\Omega (\lambda U)_\xi\,(M(U) - UM'(U))\,d\xi \\ -\Phi\nu^{-1}\int_\Omega M''(U)U_\xi\rho(\lambda U)_\xi\,d\xi.$$

The last equality can be written in the form

$$\frac{d}{dt}\int_\Omega vM(U)d\xi + \int_\Omega \Phi\nu^{-1}\rho\lambda U_\xi^2 M''(U)\,d\xi \\ = -\int_\Omega \Phi\nu^{-1}\rho\lambda_\xi U U_\xi M''(U)\,d\xi + \int_\Omega \lambda U^2 U_\xi M''(U)\,d\xi. \tag{3.6}$$

We denote

$$J_1 = \int_\Omega \Phi\nu^{-1}\rho\lambda U_\xi^2 M''(U)\,d\xi, \quad J_2 = -\int_\Omega \Phi\nu^{-1}\rho\lambda_\xi U U_\xi M''(U)\,d\xi,$$
$$J_3 = \int_\Omega \lambda U^2 U_\xi M''(U)\,d\xi.$$

By the Young inequality, we have

$$|J_2| \le J_1/2 + b_1\int_\Omega \frac{U^2|\mathbf{c}_\xi|^2}{1+|U|}d\xi \le J_1/2 + b_2\int_\Omega \rho^2 D|\mathbf{c}_\xi|^2 d\xi. \tag{3.7}$$

To estimate J_3 we write it in the form

$$J_3 = \int_\Omega \lambda F(U)_\xi d\xi = -\int_\Omega \lambda_\xi F(U)d\xi,$$

where the function F is chosen in such a way that $F(0) = 0$, $F'(U) = M''(U)U^2$, and $|F(U)| \le U^2/2$. Then, by the Young inequality,

$$|J_3| \le \delta\int_\Omega |U|^3 d\xi + b_{1\delta}\int_\Omega |U||\lambda_\xi|^2\,d\xi,$$
$$\le \delta\int_\Omega |U|^3 d\xi + b_{2\delta}\int_\Omega |U||c_\xi|^2\,d\xi$$
$$\le \delta\int_\Omega |U|^3 d\xi + b_{3\delta}\int_\Omega \rho^2 D|\mathbf{c}_\xi|^2 d\xi \text{ for any } \delta > 0.$$

On the other hand,

$$|U(\xi,t)|^3 = |U(\xi,t)|\,|\int_{-1}^{\xi} U_\xi(y,t)dy\,|^2,$$

and then

$$\int_\Omega |U(\xi,t)|^3\,d\xi \le \int_\Omega \frac{U_\xi^2}{1+|U|}\,d\xi\Big(\int_\Omega |U(\xi,t)|\,d\xi\Big)\Big(\int_\Omega (1+|U|)\,d\xi\Big). \tag{3.8}$$

Thus,

$$\begin{aligned}|J_3| \le \delta \int_\Omega \frac{U_\xi^2}{1+|U|}\,d\xi\Big(\int_\Omega |U(\xi,t)|\,d\xi\Big)\Big(\int_\Omega (1+|U|)\,d\xi\Big)\\ + b_{3\delta}\int_\Omega \rho^2 D|\mathbf{c}_\xi|^2 d\xi \text{ for any } \delta>0.\end{aligned} \tag{3.9}$$

Choosing δ small enough and applying Lemma 3.2 we infer from (3.6), (3.7), and (3.9) that

$$\|U\|_{L^\infty(0,T;L_M(\Omega))} \le b, \quad \iint_{\mathrm{Q}} \frac{U_\xi^2}{1+|U|}\,d\xi dt \le b.$$

From (3.8) and Lemma 3.2, we deduce that

$$\iint_{\mathrm{Q}} |U(\xi,t)|^3\,d\xi dt \le b.$$

Then, by the Young inequality we have

$$\begin{aligned}\iint_{\mathrm{Q}} |U_\xi|^{3/2}\,d\xi dt &= \iint_{\mathrm{Q}} \frac{|U_\xi|^{3/2}(1+|U|)^{3/4}}{(1+|U|)^{3/4}}\,d\xi dt\\ &\le \left(\iint_{\mathrm{Q}} \frac{|U_\xi|^2}{(1+|U|)} d\xi dt\right)^{3/4}\left(\iint_{\mathrm{Q}} (1+|U|)^3\,d\xi dt\right)^{1/4} \le b.\end{aligned}$$

By the same token,

$$U(\xi,t)^2 = 2\int_{-1}^{\xi} \frac{UU_\xi(1+|U|)^{1/2}}{(1+|U|)^{1/2}}\,dy.$$

Hence,

$$\sup_{\xi\in\Omega}\int_0^T |U(\xi,t|^2\,dt \le \iint_{\mathrm{Q}}\left(|U|^2+|U|^3+\frac{|U_\xi|^2}{(1+|U|)}\right)d\xi dt \le b.$$

Lemma 3.4 is proved. □

Let us summarize the above estimates. From the above lemmas one may conclude that any solution satisfies the estimates

$$\begin{aligned}&b^{-1} \le P,\rho,v \le b, \quad \||U|^{1/2}\mathbf{c}_\xi\|_{2,2} \le b,\\ &\|\mathbf{c}\|_{\infty,2} \le b, \quad \|\mathbf{c}\|_{L^2(0,T;W^{1,2}(\Omega))} \le b,\\ &\|U\|_{L^\infty(0,T;L_M(\Omega))} \le b, \quad \|U\|_{3,3} \le b,\end{aligned} \tag{3.10}$$

$$\|U_\xi\|_{3/2,3/2} \le b, \quad \|U\|_{2,\infty} \le b, \tag{3.11}$$

where the constant b depends only on $\|\mathbf{c}_0\|_2$, $\underline{\rho}$, $\overline{\rho}$, and $\|\rho_{0\xi}\|_{L_M(\Omega)}$.

Finally, we remark that

$$0 \le c_i, \quad \sum_1^{m-1} c_j \le 1 \tag{3.12}$$

if the functions $\mathbf{c}$ is regular and the initial data $\mathbf{c}_0$ satisfy the same conditions. Estimates (3.12) are easily derived from the second equation in (2.8) by the maximum principle, and they imply that the functions c_i $(1 \le i \le m-1)$ stand for volume fractions.

Because of the estimates $b^{-1} \le \rho \le b$, one can conclude that estimates (3.10) and (3.11) remain valid in the Euler setting for the solutions $(P(x,t), \mathbf{c}(x,t))$ of problem (2.1)–(2.4) in the following sense:

$$\begin{aligned}
&b^{-1} \le P, \rho \le b, \quad \||U|^{1/2}\mathbf{c}_x\|_{2,2} \le b,\\
&\|\mathbf{c}\|_{\infty,2} \le b, \quad \|\mathbf{c}\|_{L^2(0,T;W^{1,2}(\Omega))} \le b,\\
&\|U\|_{L^\infty(0,T;L_M(\Omega))} \le b, \quad \|U\|_{3,3} \le b,\\
&\|U_x\|_{3/2,3/2} \le b, \quad \|U\|_{2,\infty} \le b,
\end{aligned}$$

where $U = -P_x$.

4. Approximate solutions. We study problem (2.8)–(2.10), applying the semi-Galerkin method. Given a smooth basis $(w_i)_{i\ge1}$ in $L^2(\Omega)$, we define an approximate solution $(P_n(\xi,t), \mathbf{c}_n(\xi,t))$ by the following scheme. We look for each component c_{in}, $(i = 1, \dots, m-1)$ of the concentration vector $\mathbf{c}_n$ in the form

$$c_{in}(\xi,t) = \sum_{j=1}^{n} \alpha_j^{in}(t) w_j(\xi), \quad \alpha_{in} = (\alpha_1^{in}, \dots, \alpha_n^{in}),$$

and we find $\mathbf{c}_n$, P_n, ρ_n, and U_n from the equations

$$\frac{d}{dt}\int_\Omega c_{in} w_j \, d\xi + \int_\Omega \rho_n^2 D_n c_{in\xi} w_{j\xi} \, d\xi = 0, \quad j = 1, 2, \dots, n, \tag{4.1}$$

$$c_{in}|_{t=0} = c_{in}^{(0)}(\xi) \equiv \sum_{j=1}^{n} \alpha_{0j}^{in} w_j(\xi), \tag{4.2}$$

$$D_n = \Phi(d_m + d_p|u_n|), \quad \rho_n = \kappa e^{\nu P_n + \nu g(1+\xi)/\Phi}, \quad u_n = -\Phi\lambda(\mathbf{c}_n)\rho_n P_{n\xi}, \tag{4.3}$$

$$\nu P_{nt} = \Phi\rho_n(\lambda(\mathbf{c}_n)\rho_n P_{n\xi})_\xi, \quad P_{n\xi}|_{|\xi|=1} = 0, \quad P_n|_{t=0} = \tilde{P}_0(\xi). \tag{4.4}$$

Here, $\mathbf{c}_n^{(0)} \to \tilde{\mathbf{c}}_0$ in $L^2(\Omega)$ as $n \to \infty$.

To simplify notation in this section, we omit the index n and we assume that the function c stands for any component c_i of $\mathbf{c}_n$ and that the vector $\alpha = (\alpha_1, \dots, \alpha_n)$ stands for $(\alpha_1^{in}, \dots, \alpha_n^{in})$.

First, we prove that the approximate solution $(P(\xi,t), \mathbf{c}(\xi,t))$ exists locally in time. To this end we consider the ball

$$B(T) = \left\{ \alpha(t) \in C([0,T]; \mathbb{R}^n) : \max_{0<t<T} |\alpha(t) - \alpha_0| \le 1 \right\}$$

in $C([0,T]; \mathbb{R}^n)$, and we introduce the operator

$$\mathcal{A} : B(T) \to C([0,T]; \mathbb{R}^n), \quad \mathcal{A}\hat{\alpha} = \alpha,$$

by the following scheme. Given $\hat{\alpha} \in B(T)$, we first solve the parabolic problem (4.4) for $\hat{P}$, with the components c_{in} of the vector $\mathbf{c}_n$ substituted by

$$\hat{c} = \sum_{j=1}^{n} \hat{\alpha}_j(t) w_j, \quad \hat{\lambda} := \lambda(\hat{\mathbf{c}}),$$

and with $\hat{\rho}$ given by the second formula in (4.3). Then we take $\hat{u} = -\Phi \hat{\lambda} \hat{\rho} \hat{P}_\xi$ and find α from the ODE which is (4.1) and (4.2) for $c = \sum_{j=1}^{n} \alpha_j(t) w_j$, with ρ and u substituted by $\hat{\rho}$ and $\hat{u}$.

We denote by $W_M^1(\Omega)$ the Sobolev–Orlicz space

$$W_M^1(\Omega) = \{f(x) : f \in L_M(\Omega), \quad \partial_x f \in L_M(\Omega)\}$$

associated with an admissible convex function $M : \mathbb{R} \to \mathbb{R}$.

The Hölder space $H^{2k+\beta, k+\beta/2}(\overline{Q})$ consists of continuous functions $f(x,t)$ such that $\partial^{2l} f(x,t)/\partial x^{2l}, \partial^l f(x,t)/\partial t^l \in C(\overline{Q})$, $0 \le l \le k$, and

$$\sup_{|x_i| \le 1,\, 0 \le t \le T} \frac{|\partial^{2k} f/\partial x^{2k}(x_1,t) - \partial^{2k} f/\partial x^{2k}(x_2,t)|}{|x_1 - x_2|^\beta} < \infty,$$

$$\sup_{0 \le t_i \le T,\, |x| \le 1} \frac{|\partial^k f/\partial t^k(x,t_1) - \partial^k f/\partial t^k(x,t_2)|}{|t_1 - t_2|^{\beta/2}} < \infty.$$

For more details on the above spaces we refer the reader to [13].

To justify the scheme (4.1)–(4.4), we prove the following.

LEMMA 4.1. *Assume that*

$$\lambda_0 \le \hat{\lambda} \le \lambda_0^{-1}, \quad \hat{\lambda}_\xi \in L^2(0,T; W_0^{1,2}(\Omega)), \quad \tilde{P}_0 \in W_M^1(\Omega), \tag{4.5}$$

where $W_0^{1,2}(\Omega)$ is the closure of $\mathcal{D}(\Omega)$ in $W^{1,2}(\Omega)$ and the function M is given by (3.4). *Then, the problem*

$$\begin{aligned} &\nu P_t = \Phi \rho (\hat{\lambda} \rho P_\xi)_\xi, \quad \rho = \kappa e^{\nu P + \nu g(1+\xi)/\Phi}, \\ &P_\xi|_{|\xi|=1} = 0, \quad P|_{t=0} = \tilde{P}_0(\xi), \end{aligned} \tag{4.6}$$

has a unique solution with the properties

$$\underline{P} \le P \le \overline{P}, \quad P_\xi \in L^\infty(0,T; L_M(\Omega)) \cap L^3(Q), \quad P_{\xi\xi}, P_t \in L^{3/2}(Q).$$

Moreover,

$$\|P_{\xi\xi}, P_t\|_{3/2,3/2} + \|P_\xi\|_{L^\infty(0,T;L_M(\Omega))} + \|P_\xi\|_{3,3} \le b,$$

where the constant b depends only on $\lambda_0, \underline{P}, \overline{P}$ and on the norms $\|\tilde{P}_0\|_{W_M^1(\Omega)}$ and $\|\hat{\lambda}_\xi\|_{L^2(0,T;W_0^{1,2}(\Omega))}$.

Proof. Let (λ_δ) be a sequence of functions from the set

$$\begin{aligned} &\lambda_0 \le \lambda \le \lambda_0^{-1}; \quad \lambda, \lambda_\xi, \lambda_t, \lambda_{t\xi} \in C(\overline{Q}); \\ &\lambda_\xi|_{|\xi|=1} = 0, \quad \lambda, \lambda_\xi \in H^{1+\beta,(1+\beta)/2}(\overline{Q}) \quad (0 < \beta < 1), \end{aligned} \tag{4.7}$$

and such that

$$\lambda_\delta \to \hat{\lambda} \text{ in } L^2(0,T;W^{2,2}(\Omega)) \text{ as } \delta \to 0.$$

Let $P_{\delta 0}(\xi)$ be a sequence from $H^{3+\beta}(\Omega)$ such that

$$P_{\delta 0} \to \tilde{P}_0 \text{ in } W^1_M(\Omega) \text{ as } \delta \to 0.$$

We also assume that the compatibility conditions

$$\frac{\partial}{\partial \xi}\left(\rho_{\delta 0}\frac{\partial}{\partial \xi}\left(\lambda_{\delta 0}\rho_{\delta 0}\frac{\partial}{\partial \xi}P_{\delta 0}\right)\right)|_{|\xi|=1} = 0 \tag{4.8}$$

are fulfilled.

Let us consider problem $(4.6)_\delta$, which is (4.6) with λ and $\tilde{P}_0$ substituted by λ_δ and $P_{\delta 0}$, respectively. By Theorem 7.4 of [13, Chapter V], there is a unique solution $P_\delta \in H^{3+\beta,(3+\beta)/2}(\overline{Q})$ of $(4.6)_\delta$. By the arguments applied in Lemmas 3.1 and 3.2, one obtains the estimates

$$\underline{P} \le P_\delta \le \overline{P}, \quad \underline{\rho} \le \rho_\delta \le \overline{\rho}, \quad \sup_{t\ge 0}\int_\Omega |P_{\delta\xi}|\,d\xi \le \int_\Omega |P_{\delta 0\xi}|\,d\xi.$$

It follows from Lemma 3.4 (see equality (3.6)) that the function $U_\delta = -\Phi\rho_\delta P_{\delta\xi}$ satisfies the equality

$$\begin{aligned}&\frac{d}{dt}\int_\Omega v_\delta M(U_\delta)\,d\xi + \int_\Omega \Phi\nu^{-1}\rho_\delta\lambda_\delta U^2_{\delta\xi}M''(U_\delta)\,d\xi\\ &= \int_\Omega \Phi\nu^{-1}\rho_\delta\lambda_{\delta\xi}U_\delta U_{\delta\xi}M''(U_\delta)\,d\xi + \int_\Omega (\lambda_\delta U_\delta)^2 U_{\delta\xi}M''(U_\delta)\,d\xi.\end{aligned} \tag{4.9}$$

We denote

$$J^\delta_1 = \int_\Omega \Phi\nu^{-1}\rho_\delta\lambda_\delta U^2_{\delta\xi}M''(U_\delta)\,d\xi, \quad J^\delta_2 = \int_\Omega \Phi\nu^{-1}\rho_\delta\lambda_{\delta\xi}U_\delta U_{\delta\xi}M''(U_\delta)\,d\xi,$$
$$J^\delta_3 = \int_\Omega (\lambda_\delta U_\delta)^2 U_{\delta\xi}M''(U_\delta).$$

We have

$$|J^\delta_2| \le J^\delta_1/2 + b_1\int_\Omega \lambda^2_{\delta\xi}|U_\delta|d\xi,$$

$$|J^\delta_3| \le \eta\int_\Omega |U_\delta|^3 d\xi + b_{1\eta}\int_\Omega |U_\delta||\lambda^2_{\delta\xi}\,d\xi \text{ for any } \eta > 0.$$

Observe that

$$\int_\Omega \lambda^2_{\delta\xi}|U_\delta|d\xi \le 2\int_\Omega \lambda^2_{\delta\xi\xi}d\xi\int_\Omega |U_\delta|d\xi.$$

Then, as in the proof of Lemma 3.4, we obtain

$$\|P_{\delta\xi}\|_{L^\infty(0,T;L_M(\Omega))} + \|P_{_\delta\xi\xi}\|_{3/2,3/2} + \|P_{_\delta\xi}\|_{3,3} + \|P_{_\delta\xi}\|_{2,\infty} \le b, \tag{4.10}$$

uniformly in δ.

Observe also that

$$\int_\Omega |\lambda_\xi|^3 d\xi \le 2 \int_\Omega |\lambda_\xi| d\xi \int_\Omega |\lambda_{\xi\xi}|^2 d\xi, \tag{4.11}$$

for any function $\lambda(\xi)$ such that $\lambda_\xi|_{|\xi|=1} = 0$. Hence, it follows from $(4.6)_\delta$ that the time derivative $P_{\delta t}$ satisfies the estimate

$$\|P_{\delta t}\|_{3/2,3/2} \le b,$$

uniformly in δ.

By the compact imbedding of $W^{1,3/2}(Q)$ into $L^{3/2}(Q)$, there is a subsequence (P_{δ_k}) and a function $P \in W^{1,3/2}(Q)$ such that $P_{\delta_k} \to P$ in $L^{3/2}(Q)$. Since P_{δ_k} is bounded in $L^\infty(Q)$, $P_{\delta_k} \to P$ in $L^q(Q)$ for any $1 \le q < \infty$, by interpolation. Clearly, $\rho_{\delta_k} \to \rho$ in $L^q(Q)$ as well.

Let us consider the function $V_\delta := v_{\delta\xi}$ which solves (see (2.11)) the equation

$$V_{\delta t} = \Phi\nu^{-1}(\rho_\delta^2 \lambda_\delta V_\delta)_{\xi\xi} + (\rho_\delta g \lambda_\delta)_{\xi\xi}.$$

From the above estimates it follows that

$$\|V_\delta\|_{L^{3/2}(0,T;W^{1,3/2}(\Omega))} + \|V_{\delta t}\|_{L^3(0,T;W^{-2,3}(\Omega))} \le b.$$

Hence, by the Aubin–Lions compactness theorem [14] one may assume that $P_{\delta_k\xi} \to P_\xi$ in $L^{3/2}(Q)$. Moreover, by estimates (4.10) and interpolation, $P_{\delta_k\xi} \to P_\xi$, $\rho_{\delta_k\xi} \to \rho_\xi$ in $L^q(Q)$, for any $1 \le q < 3$.

Let us show that the limit function P solves problem (4.6). Equation $(4.6)_\delta$ is written as

$$\nu P_{\delta t} = \Phi\rho_\delta\Big(\lambda_{\delta\xi}\rho_\delta P_{\delta\xi} + \lambda_\delta \rho_{\delta\xi} P_{\delta\xi} + \lambda_\delta \rho_\delta P_{\delta\xi\xi}\Big). \tag{4.12}$$

We claim that each term converges in $\mathcal{D}'(Q)$. Let us consider the last term. Denote for simplicity $P_\delta = P_{\delta_k}$. We write the decomposition equality

$$\lambda_\delta \rho_\delta^2 P_{\delta\xi\xi} = \lambda\rho^2 P_{\delta\xi\xi} + (\rho_\delta - \rho)(\rho_\delta + \rho)\lambda_\delta P_{\delta\xi\xi} + \rho^2(\lambda_\delta - \lambda)P_{\delta\xi\xi} \equiv \sum_{i=1}^3 J_i^\delta.$$

By (4.10), $P_{\delta\xi\xi} \to P_{\xi\xi}$ weakly in $L^{3/2}(Q)$. On the other hand, due to (4.11), $\lambda_{\delta\xi} \to \lambda_\xi$ in $L^3(Q)$. Hence,

$$\int_Q J_1^\delta \varphi \, d\xi \to \int_Q \lambda\rho^2 P_{\xi\xi}\varphi \, d\xi, \quad \int_Q J_2^\delta \varphi \, d\xi \to 0, \quad \int_Q J_3^\delta \varphi \, d\xi \to 0$$

for any $\varphi \in \mathcal{D}(Q)$.

One can study the convergence of the other terms of (4.12) similarly. Thus, the limit function P solves problem (4.6). Observe that equality (4.6) is valid also as equality of functions in $L^{3/2}(Q)$; hence the function v ($v = 1/\rho$) calculated via formula $(4.6)_2$ solves (2.11) in the same sense.

Let us prove uniqueness. Given two solutions v_1 and v_2 of equation (2.11), with the boundary conditions $v_{i\xi}|_{|\xi|=1} = 0$, we denote $v = v_1 - v_2$, $\rho = \rho_1 - \rho_2$. The function v solves the parabolic problem

$$v_t = \Phi\nu^{-1}(\rho_1^2 \hat\lambda v_\xi)_\xi + \Phi\nu^{-1}(\hat\lambda v_{2\xi}\rho(\rho_1 + \rho_2))_\xi + (g\hat\lambda\rho)_\xi, \; v_\xi|_{|\xi|=1} = 0, \; v|_{t=0} = 0.$$

We multiply this equation by the function $M_1'(v) = v/(v^2+\delta^2)^{1/2}$, a smooth approximation of $\operatorname{sign}(v)$, and integrate the result over $\Omega \times (0,t)$. It follows that

$$\begin{aligned}\int_\Omega M_1(v(t))d\xi + \underbrace{\Phi\nu^{-1}\int_0^t\int_\Omega M_1''(v)\rho_1^2\hat\lambda v_\xi^2\,d\xi d\tau}\\ = -\int_0^t\int_\Omega M_1''(v)\rho\hat\lambda v_\xi\Big(\Phi\nu^{-1}v_{2\xi}(\rho_1+\rho_2)+g\Big)\,d\xi d\tau\\ \le \underbrace{\cdots}/2 + b\int_0^t\int_\Omega M_1''(v)\rho^2(1+v_{2\xi}^2)\,d\xi d\tau.\end{aligned}$$

Observe that $0 \le M_1''(v)v^2 \le \delta$, $v_{2\xi}^2 \in L^{3/2}(Q)$, $\rho = -\rho_1\rho_2 v$. Hence, in the limit $\delta \to 0$ we obtain that $\int_\Omega |v(t)|\,d\xi \le 0$. Lemma 4.1 is proved. □

By the previous lemma, the operator $\mathcal{A}$, defined by the right-hand side of the equality

$$\alpha_j = \alpha_{0j} + (M^{-1})_{ij}\int_0^t\int_\Omega \hat\rho^2\hat D\hat c_\xi w_{i\xi}d\xi d\tau \equiv (\mathcal{A}\hat\alpha)_j, \quad M_{ij} = \int_\Omega w_i w_j d\xi$$

is well defined. Moreover, it is compact since it maps $B(T)$ on a bounded set of $W^1(0,T)$. Observe that all the norms in the n-dimensional space spanned by the basis function w_j, $j = 1,\dots,n$, are equivalent. Clearly, $\mathcal{A}$ maps $B(T_n)$ into itself provided T_n is small enough. Consequently, one can use the Schauder theorem to conclude that there exists at least one fixed point α. Thus, problem (4.1)–(4.4) has a local solution.

Now, this procedure can be iterated as many times as necessary to reach $T_n = T$ as long as there is a bound on α independent of T_n. The existence of such a bound follows from the equality

$$\frac{d}{dt}\int_\Omega c_{in}^2 d\xi + \int_\Omega \rho_n^2 D_n c_{in\xi}^2 d\xi = 0,$$

which is valid for any fixed point α. Indeed, by integration we have

$$\sup_{t\ge0}\int_\Omega c_{in}^2(t)d\xi + \int_0^t\int_\Omega \rho_n^2 D_n c_{in\xi}^2 d\xi \le \limsup_{n\to\infty}\int_\Omega |c_{in}^{(0)}(\xi)|^2 d\xi.$$

Hence, an approximate solution defined by (4.1)–(4.4) exists globally.

Let us show that the regularity of $P_n(\xi,t)$ can be improved provided that the initial datum $\tilde P_0(\xi)$ is more regular than in Lemma 4.1.

LEMMA 4.2. *Assume that $\tilde P_0(\xi)$ belongs to $W^{2,2}(\Omega)$. Then*

$$P_n \in L^\infty(0,T;W^{2,2}(\Omega)) \cap L^2(0,T;W^{3,2}(\Omega)), \quad \frac{\partial^2 P_n}{\partial t\partial\xi} \in L^2(Q).$$

Moreover, the functions $U_n = -\Phi\rho_n P_{n\xi}$ and $V_n = v_{n\xi}$ solve the parabolic problems

$$\begin{aligned}&U_{nt} + \rho_n U_n(\lambda(\mathbf{c}_n)U_n)_\xi = \Phi\nu^{-1}\rho_n(\,\rho_n(\lambda(\mathbf{c}_n)U_n)_\xi\,)_\xi,\\ &U_n|_{|\xi|=1} = 0, \quad U_n|_{t=0} = -\Phi\tilde\rho_0(\xi)\tilde P_0(\xi),\end{aligned} \tag{4.13}$$

and

$$\begin{aligned}&V_{nt} = \Phi\nu^{-1}(\rho_n^2\lambda(\mathbf{c}_n)V_n)_{\xi\xi} + (\rho_n g\lambda(\mathbf{c}_n))_{\xi\xi}.\\ &V_n|_{|\xi|=1} = 0, \quad V_n|_{t=0} = \tilde v_{0\xi}(\xi).\end{aligned} \tag{4.14}$$

The differential equations (4.13) *and* (4.14) *hold as equalities of functions from* $L^2(Q)$.

Proof. Let (λ_δ) be a sequence satisfying conditions (4.7) and such that

$$\lambda_\delta \to \lambda(\mathbf{c}_n) \text{ in } W^1(0,T;C^2(\overline{\Omega})) \text{ as } \delta \to 0.$$

Let $P_{\delta 0}(\xi)$ be a sequence from $H^{3+\beta}(\Omega)$ such that

$$P_{\delta 0} \to \tilde{P}_0 \text{ in } W^{2,2}(\Omega) \text{ as } \delta \to 0.$$

Assume that the compatibility conditions (4.8) are fulfilled. We consider the parabolic problem $(4.4)_\delta$, which is problem (4.4) with $\lambda(\mathbf{c}_n)$ and $\tilde{P}_0$ substituted by λ_δ and $P_{\delta 0}$, respectively. By Theorem 7.4 of [13, Chapter V], there is a unique solution $P_\delta \in H^{3+\beta,(3+\beta)/2}(\overline{Q})$ of $(4.4)_\delta$.

The function $U_\delta = -\Phi\rho_\delta P_{\delta\xi}$ solves the parabolic equation

$$U_{\delta t} + \rho_\delta U_\delta(\lambda_\delta U_\delta)_\xi = \Phi\nu^{-1}\rho_\delta(\rho_\delta(\lambda_\delta U_\delta)_\xi)_\xi. \tag{4.15}$$

Due to the choice of λ_δ, one can obtain, by the same arguments as in Lemma 4.1, the following estimates:

$$\underline{\rho} \le \rho_\delta \le \overline{\rho}, \quad \|P_{\delta t}\|_{3/2,3/2} \le b, \tag{4.16}$$

$$\|U_\delta\|_{L^\infty(0,T;L_M(\Omega))} + \|U_{\delta\xi}\|_{3/2,3/2} + \|U_\delta\|_{3,3} + \|U_\delta\|_{2,\infty} \le b \tag{4.17}$$

uniformly in δ.

The function $V_\delta := v_{\delta\xi}$ solves the equation

$$V_{\delta t} = \Phi\nu^{-1}(\rho_\delta^2\lambda_\delta V_\delta)_{\xi\xi} + (\rho_\delta g\lambda_\delta)_{\xi\xi}. \tag{4.18}$$

We multiply it by V_δ^3 and integrate the result over Ω to obtain

$$\frac{1}{4}\frac{d}{dt}\int_\Omega V_\delta^4\,d\xi + \overbrace{3\Phi\nu^{-1}\int_\Omega \rho_\delta^2\lambda_\delta V_\delta^2 V_{\delta\xi}^2\,d\xi}^{\ddots} = 6\Phi\nu^{-1}\int_\Omega \rho_\delta^3\lambda_\delta V_\delta^4 V_{\delta\xi}\,d\xi$$
$$+ 3\int_\Omega \rho_\delta^2 V_\delta^3 V_{\delta\xi}(g\lambda_\delta - \Phi\nu^{-1}\lambda_{\delta\xi})\,d\xi - 3g\int_\Omega \rho_\delta\lambda_{\delta\xi}V_\delta^2 V_{\delta\xi}\,d\xi \equiv \sum_{i=1}^3 J_i.$$

We estimate

$$J_1 + J_2 \le \overbrace{\cdots}/4 + b_1(\|U_\delta\|_\infty^2 + 1)\int_\Omega V_\delta^4\,d\xi, \quad J_3 \le \overbrace{\cdots}/4 + b_2\|U_\delta\|_2^2.$$

Due to estimates (4.17), one concludes that

$$\frac{d}{dt}\int_\Omega V_\delta^4\,d\xi \le f_1(t)\int_\Omega V_\delta^4\,d\xi + f_2(t),$$

with the functions f_i bounded in $L^1(0,T)$ uniformly in δ. By Gronwall's lemma it follows that

$$\|V_\delta, U_\delta\|_{\infty,4} \le b, \quad \|V_\delta V_{\delta\xi}, U_\delta U_{\delta\xi}\|_{2,2} \le b \tag{4.19}$$

uniformly in δ. Observe that some constants b in this proof may depend on n.

With such estimates at hand, we multiply (4.15) by $U_{\delta\xi\xi}$ and integrate the result over Ω to obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_\Omega U_{\delta\xi}^2 d\xi + \underbrace{\Phi\nu^{-1}\int_\Omega \rho_\delta^2\lambda_\delta U_{\delta\xi\xi}^2 d\xi}\\
&= -\int_\Omega \rho_\delta U_{\delta\xi}U_{\delta\xi\xi}\Big([\Phi\nu^{-1}(\rho_\delta\lambda_\delta)_\xi + \lambda_\delta U_\delta] + [\Phi\nu^{-1}(\rho_{\delta\xi}\lambda_{\delta\xi} + \rho_\delta\lambda_{\delta\xi\xi}) + \lambda_{\delta\xi}U_\delta]\Big)d\xi\\
&\le \underbrace{\cdots}/2 + b\int_\Omega (1 + U_\delta^4 + U_{\delta\xi}^2U_\delta^2)\,d\xi.
\end{aligned}$$

Hence,

$$\|U_{\delta\xi}, V_{\delta\xi}\|_{\infty,2} \le b, \quad \|U_{\delta\xi\xi}, V_{\delta\xi\xi}, U_{\delta t}, V_{\delta t}\|_{2,2} \le b \tag{4.20}$$

uniformly in δ.

Due to estimates (4.16), (4.17), (4.19), and (4.20), and the uniqueness of solutions of problem (4.4), one can prove by the arguments of Lemma 4.1 that $P_\delta \to P_n$ in the same sense as in Lemma 4.1. Moreover,

$$U_{\delta\xi} \to U_{n\xi}, \quad U_{\delta\xi\xi} \to U_{n\xi\xi}, \quad U_{\delta t} \to U_{nt} \text{ weakly in } L^2(Q),$$

and

$$U_\delta \to U_n \text{ strongly in } L^q(Q) \text{ for any } 1 \le q < 4.$$

The similar convergences are valid for the sequence V_δ:

$$V_{\delta\xi} \to V_{n\xi}, \quad V_{\delta\xi\xi} \to V_{n\xi\xi}, \quad V_{\delta t} \to V_{nt} \text{ weakly in } L^2(Q),$$

and

$$V_\delta \to V_n \text{ strongly in } L^q(Q) \text{ for any } 1 \le q < 4.$$

The above claims of convergence enable us to pass to the limit, as $\delta \to 0$, in (4.15) and (4.18). Lemma 4.2 is proved. □

Consequence. It follows easily from the above lemma that the function $z_n = P_{n\xi}$ solves the equation

$$\nu z_{nt} = (\Phi\rho_n(\lambda(\mathbf{c}_n)\rho_n z_n)_\xi)_\xi \tag{4.21}$$

in $L^2(Q)$.

5. Global solvability. We argue by a compactness method. To this end we use estimates independent of n for the approximate solutions $(P_n, \mathbf{c}_n)$, defined by (4.1)–(4.4). It is an advantage of the semi-Galerkin method that all the a priori estimates obtained in section 3 are also valid for $(P_n, \mathbf{c}_n)$:

$$b^{-1} \le P_n, \rho_n \le b, \quad \||u_n|^{1/2}\mathbf{c}_{n\xi}\|_{2,2} \le b,$$

$$\|\mathbf{c}_n\|_{\infty,2} \le b, \quad \|\mathbf{c}_n\|_{L^2(0,T;W^{1,2}(\Omega))} \le b, \tag{5.1}$$

$$\|P_{n\xi}\|_{L^\infty(0,T;L_M(\Omega))} \le b, \quad \|P_{n\xi}\|_{3,3} \le b,$$

$$\|P_{n\xi}\|_{2,\infty} \le b, \quad \|P_{n\xi\xi}, P_{nt}\|_{3/2,3/2} \le b, \tag{5.2}$$

uniformly in n. By the Young inequality, one can derive from (5.1), (5.2) that

$$\|D_n \mathbf{c}_{n\xi}\|_{3/2,3/2} \le b,$$

uniformly in n. It follows from (4.21) that

$$\left\| \frac{\partial}{\partial t} P_{n\xi} \right\|_{L^{3/2}(0,T;\, W^{-1,3/2}(\Omega))} \le b, \tag{5.3}$$

uniformly in n.

To derive the estimate

$$\|\mathbf{c}_{nt}\|_{L^{3/2}(0,T;\,(W^{2,2}(\Omega))^*)} \le b, \tag{5.4}$$

we choose the basis $(w_j)_{j\ge 1}$, defined as

$$w_{j\xi\xi} = -\lambda_j w_j, \quad w_{j\xi}|_{|\xi|=1} = 0.$$

Clearly, $w_j = \cos(\pi(j-1)\xi)$ and $\lambda_j = \pi^2(j-1)^2$.

Denote by $\mathcal{P}_n$ the projection of $L^2(\Omega)$ onto X_n, the subspace of $L^2(\Omega)$, spanned by $w_1, \dots, w_n$. Observe that

$$\|(\mathcal{P}_n w)_{\xi\xi}\|_2^2 = \sum_{j=1}^{n} \lambda_j^2 \beta_j^2 \le \sum_{j=1}^{\infty} \lambda_j^2 \beta_j^2 = \|w_{\xi\xi}\|_2^2, \quad w = \sum_{j=1}^{\infty} \beta_j w_j,$$

and the norm of w in $W^{2,2}(\Omega)$ is equivalent to $(\|w\|_2^2 + \|w_{\xi\xi}\|_2^2)^{1/2}$.

One can write equality (4.1) as

$$\frac{d}{dt} \int_\Omega c_{in}(\mathcal{P}_n w)\, d\xi + \int_\Omega \rho_n^2 D_n c_{in\xi} (\mathcal{P}_n w)_\xi \, d\xi = 0 \quad \forall w \in W^{2,2}(\Omega).$$

Now it follows that

$$|\langle \mathbf{c}_{nt}, w\rangle| \le \overline{\rho}^2 \|D_n \mathbf{c}_{n\xi}\|_{3/2} \|(\mathcal{P}_n w)_\xi\|_3.$$

Observe that

$$\|(\mathcal{P}_n w)_\xi\|_3 \le b_1 \|(\mathcal{P}_n w)_\xi\|_{W^{1,2}(\Omega)} \le b_2 \|\mathcal{P}_n w\|_{W^{2,2}(\Omega)} \le b_3 (\|\mathcal{P}_n w\|_2^2 + \|(\mathcal{P}_n w)_{\xi\xi}\|_2^2)^{1/2}$$

$$\le b_3 (\|w\|_2^2 + \|w_{\xi\xi}\|_2^2)^{1/2} \le b_4 \|w\|_{W^{2,2}(\Omega)}.$$

Hence,

$$\|\mathbf{c}_{nt}(t)\|_{(W^{2,2}(\Omega))^*} \le b \|D_n(t) \mathbf{c}_{n\xi}(t)\|_{3/2}.$$

Thus, estimate (5.4) is established.

THEOREM 5.1. *Assume that $\tilde{P}_0 \in W^{2,2}(\Omega)$, $\tilde{\mathbf{c}}_0 \in L^2(\Omega)$, and function $\lambda(\mathbf{c})$ satisfies restrictions* (1.14). *Then, problem* (2.8)–(2.10) *has a solution $(P(\xi,t), \mathbf{c}(\xi,\mathbf{t}))$ with the properties*

$$\underline{\rho} \le \rho \le \overline{\rho},\ P_t \in L^{3/2}(Q),$$
$$P \in L^\infty(0,T; W_M^1(\Omega)) \cap L^{3/2}(0,T; W^{2,3/2}(\Omega)) \cap L^2(0,T; W^{1,\infty}(\Omega))$$
$$\cap\ L^3(0,T; W^{1,3}(\Omega)) \cap L^\infty(Q), \quad \mathbf{c} \in L^\infty(0,T; L^2(\Omega)) \cap L^2(0,T;\, W^{1,2}(\Omega)),$$

where the function M is given by (3.4). *The first equation in* (2.8) *holds as an equality of functions in* $L^{3/2}(Q)$. *The second equation in* (2.8) *holds weakly:*

$$\iint_{\mathrm{Q}} \left(\mathbf{c}\Psi_t - \Phi\rho^2 D\mathbf{c}_\xi \Psi_\xi\right) d\xi dt = -\int_\Omega \tilde{\mathbf{c}}_0 \Psi(\xi,0) d\xi \tag{5.5}$$

for any $\Psi \in C^1(\bar{Q})$ *such that* $\Psi(\xi,T)=0$.

Proof. The uniform bounds $(5.1)_4$ and (5.4) imply by the Aubin–Lions theorem that there are a subsequence, still denoted $(\mathbf{c})_n$, and a function $\mathbf{c}$, such that $\mathbf{c}_n \to \mathbf{c}$ a.e. in Q, and $\mathbf{c}_n \to \mathbf{c}$ strongly in $L^2(Q)$ and weakly in $L^2(0,T;W^{1,2}(\Omega))$. Hence $\lambda(\mathbf{c}_n) \to \lambda(\mathbf{c})$ a.e. in Q and strongly in $L^q(Q)$ for any $q \in [1,\infty)$.

The uniform bound $\|P_n\|_{W^{1,3/2}(Q)} \le b$ and the compact imbedding of $W^{1,3/2}(Q)$ into $L^{3/2}(Q)$ imply that, up to a subsequence, there is a function $P \in W^{1,3/2}(Q)$ such that $P_n \to P$, $\rho_n \to \rho$ strongly in $L^q(Q)$, for any $q \in [1,3/2)$. Since $\underline{P} \le P \le \overline{P}$, one can also obtain by an interpolation argument that $P_n \to P$, $\rho_n \to \rho$ strongly in $L^q(Q)$ for any $q \in [1,\infty)$.

It follows from (5.2) that

$$P_{n\xi\xi} \to P_{\xi\xi}, \quad P_{nt} \to P_t, \text{ weakly in } L^{3/2}(Q).$$

Due to estimates (5.3) and $(5.2)_4$, one concludes that $P_{n\xi} \to P_\xi$ in $L^{3/2}(Q)$. Since $\|P_{n\xi}\|_{3,3} \le b$, one has by interpolation that $P_{n\xi} \to P_\xi$ in $L^q(Q)$, for any $q \in [1,3)$.

As for the function $u_n = -\Phi\rho_n\lambda(\mathbf{c}_n)P_{n\xi}$, one has by the above token that $u_n \to u$ in $L^q(Q)$, for any $q \in [1,3)$. Hence, $|u_n|\,\mathbf{c}_{n\xi} \to |u|\,\mathbf{c}_\xi$ in $\mathcal{D}'(Q)$.

The convergences described above enable us to pass to the limit, as $n \to \infty$, in (4.1)–(4.4). The proof of Theorem 5.1 is complete. ∎

The next result concerns the problem in the Euler setting.

THEOREM 5.2. *Assume that* $P_0 \in W^{2,2}(\Omega)$, $\mathbf{c}_0 \in L^2(\Omega)$, *and the function* $\lambda(\mathbf{c})$ *satisfies restrictions* (1.14). *Then, problem* (2.1)–(2.4) *has a solution* $(P(x,t),\mathbf{c}(x,t))$ *with the properties*

$$\begin{aligned}
&\underline{\rho} \le \rho \le \overline{\rho}, P_t \in L^{3/2}(Q),\\
&P \in L^\infty(0,T;W^1_M(\Omega)) \cap L^{3/2}(0,T;W^{2,3/2}(\Omega)) \cap L^2(0,T;W^{1,\infty}(\Omega))\\
&\quad \cap\, L^3(0,T;W^{1,3}(\Omega)) \cap L^\infty(Q), \quad \mathbf{c} \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;W^{1,2}(\Omega)),
\end{aligned}$$

where the function M is given by (3.4). *Equation* (2.1) *holds as an equality of functions from* $L^{3/2}(Q)$. *Equation* (2.2) *holds weakly:*

$$\iint_{\mathrm{Q}} \rho\mathbf{c}(\Phi\psi_t + u\psi_x) - \rho D\mathbf{c}_x\psi_x \, dxdt = -\int_\Omega \Phi\rho_0(x)\mathbf{c}_0(x)\psi(x,0)dx, \tag{5.6}$$

for any $\psi \in C^1(\bar{Q})$ *such that* $\psi(x,T)=0$.

Proof. Let $(\tilde{P}(\xi,t),\tilde{\mathbf{c}}(\xi,t))$ be a solution in the Lagrange setting given by Theorem 5.1. The function $\tilde{\rho}(\xi,t)$ enables us to define the Euler variables by the formulae

$$x(\xi,t) = \Phi^{-1}\int_{-1}^{\xi} \tilde{v}(y,t)dy - 1, \quad x_\xi = \Phi^{-1}\tilde{v}, \quad x_t = \Phi^{-1}\tilde{u}.$$

Clearly,

$$x \in C(\bar{Q}) \cap L^2(0,T;W^{2,3/2}(\Omega)), \quad x_t \in L^3(Q) \cap L^{3/2}(0,T;W^{1,3/2}(\Omega)).$$

The inverse function $\xi(x,t)$ enjoys the same regularity as a function of the variables x and t.

One can verify that the functions

$$P(x,t) = \tilde{P}(\xi,t)|_{\xi=\xi(x,t)}, \quad \mathbf{c}(x,t) = \tilde{\mathbf{c}}(\xi,t)|_{\xi=\xi(x,t)}$$

solve the equations

$$\Phi P_t + uP_x = \left(\frac{\lambda(\mathbf{c})}{\nu} P_x\right) x, \quad u = -\lambda(\mathbf{c})P_x$$

a.e. in Q.

Let us verify that the functions $P(x,t)$ and $\mathbf{c}(x,t)$ solve (5.6). First, we observe that, by a continuity argument, the functions $\tilde{P}(\xi,t)$ and $\tilde{\mathbf{c}}(\xi,t)$ satisfy (5.5) for all $\Psi \in L^2(Q)$ such that $\Psi_t \in L^2(Q)$, $\Psi_\xi \in L^3(Q)$, and $\Psi(\xi,T) = 0$. Observe that because of the inclusion $\Psi_t \in L^2(Q)$, the function Ψ belongs to $C([0,T];L^2(\Omega))$ and Ψ has traces at $t = \text{const}$.

Given an arbitrary function $\psi(x,t)$ such that $\psi \in C^1(\bar{Q})$, $\psi(x,T) = 0$, let us consider the sum of integrals

$$J = \iint_{\mathrm{Q}} \Big(\rho\mathbf{c}(\Phi\psi_t + u\psi_x) - \rho D\mathbf{c}_x\psi_x\Big)\, dxdt + \int_\Omega \mathbf{c}_0(x)\psi(x,0)\, dx.$$

Denoting $\Psi(\xi,t) = \psi(x,t)|_{x=x(\xi,t)}$, and making the change of variables $(x,t) \to (\xi,t)$, we have

$$J = \iint_{\mathrm{Q}} \Big(\tilde{\mathbf{c}}\Psi_t - \Phi\tilde{\rho}^2\tilde{D}\tilde{\mathbf{c}}_\xi\Psi_\xi\Big)\, d\xi dt + \int_\Omega \tilde{\mathbf{c}}_0\Psi(\xi,0)\, d\xi,$$

where

$$\Psi_t = \Phi^{-1}\psi_x(x(\xi,t),t)\tilde{u}(\xi,t) + \psi_t(x(\xi,t),t) \in L^2(Q),$$

$$\Psi_\xi = \Phi^{-1}\psi_x(x(\xi,t),t)\tilde{v}(\xi,t) \in L^3(Q),$$

and $\Psi(\xi,0) = \psi(x(\xi,0),0) \in L^2(\Omega)$. Clearly, Ψ belongs to $C([0,T];L^2(\Omega))$. Thus, $J = 0$ and the theorem is proved. □

REFERENCES

[1] H. Amann, *Dynamic theory of quasi-linear parabolic systems* III. *Global existence*, Math. Z., 202 (1989), pp. 219–250.

[2] Y. Amirat and A. Ziani, *Global weak solutions for a parabolic system modelling a one-dimensional miscible flow in porous media*, J. Math. Anal. Appl., 220 (1998), pp. 697–718.

[3] G. Chavent and J. Jaffré, *Mathematical Methods and Finite Elements for Reservoir Simulation*, Stud. Math. Appl. 17, North–Holland, Amsterdam, 1986.

[4] Z. Chen and R. Ewing, *Mathematical analysis for reservoir models*, SIAM J. Math. Anal., 30 (1999), pp. 431–453.

[5] J. Douglas and J. E. Roberts, *Numerical methods for a model of compressible miscible displacement in porous media*, Math. Comp., 41 (1983), pp. 441–459.

[6] R. E. Ewing, ed., *The Mathematics of Reservoir Simulation*, Frontiers in Appl. Math. 1, SIAM, Philadelphia, 1984.

[7] P. Fabrie and M. Langlais, *Mathematical analysis of miscible displacement in porous medium*, SIAM J. Math. Anal., 23 (1992), pp. 1375–1392.

[8] X. Feng, *On existence and uniqueness results for a coupled system modeling miscible displacement in porous media*, J. Math. Anal. Appl., 194 (1995), pp. 883–910.

[9] H. Frid and V. Shelukhin, *Initial boundary value problems for a quasi-linear parabolic system in three-phase capillary flow in porous media*, SIAM J. Math. Anal., 36 (2005), pp. 1407–1425.
[10] R. Helmig, *Multiphase Flow and Transport Processes in the Subsurface*, Springer-Verlag, Berlin, 1997.
[11] M. A. Krasnosel'skii and Ya. B. Rutitskii, *Convex Functions and Orlicz Spaces*, Noordhoff, Leiden, 1961.
[12] A. Kufner, S. Fucik, and O. John, *Functions Spaces*, Academia, Prague, 1977.
[13] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[14] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod; Gauthier-Villars, Paris, 1969.
[15] A. Mikelić, *Mathematical theory of stationary miscible filtration*, J. Differential Equations, 90 (1991), pp. 186–202.
[16] F. Otto, *L^1-contraction and uniqueness for quasi-linear elliptic-parabolic equations*, J. Differential Equations, 131 (1996), pp. 20–38.
[17] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, Amsterdam, 1977.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1847–1859

A LOSING ESTIMATE FOR THE IDEAL MHD EQUATIONS WITH APPLICATION TO BLOW-UP CRITERION*

MARCO CANNONE†, QIONGLEI CHEN‡, AND CHANGXING MIAO‡

Abstract. In this paper we study the blow-up criterion of smooth solution to the ideal MHD equations in $\mathbb{R}^n$. By means of the Fourier frequency localization and Bony paraproduct decomposition, we show a losing estimate for the ideal MHD equations and apply it to establish an improved blow-up criterion of smooth solutions. As a special case, we recover a previous result of Planchon for the incompressible Euler equations.

Key words. ideal MHD equations, blow-up, Littlewood–Paley decomposition, Besov space

AMS subject classifications. 76W05, 35B65

DOI. 10.1137/060652002

1. Introduction.

We are concerned with the blow-up phenomena of smooth solutions to the ideal MHD equations in $\mathbb{R}^n$:

$$\text{(IMHD)}\quad \begin{cases} u_t + u\cdot\nabla u = -\nabla p - \frac{1}{2}\nabla b^2 + b\cdot\nabla b, \\ b_t + u\cdot\nabla b = b\cdot\nabla u, \\ \nabla\cdot u = \nabla\cdot b = 0, \\ u(0,x) = u_0(x), \quad b(0,x) = b_0(x), \end{cases} \tag{1.1}$$

where $x\in\mathbb{R}^n, t\geq 0$, u, b describes the flow velocity vector and the magnetic field vector, respectively, p is a scalar pressure, while u_0 and b_0 are the given initial velocity and initial magnetic field with $\nabla\cdot u_0 = \nabla\cdot b_0 = 0$, $b^2 = |b|^2$.

By the standard energy method [12], it is known that for $(u_0, b_0)\in H^s(\mathbb{R}^3)$, $s\geq 3$, there exists $T>0$ such that the Cauchy problem (1.1) has a unique smooth solution $(u(t,x), b(t,x))$ on $[0,T)$ satisfying

$$(u,b)\in C([0,T); H^s)\cap C^1([0,T); H^{s-1}).$$

However, in contrast to the Euler equation (when $b=0$), the global existence of smooth solutions to the inviscid MHD equations is not known even in two dimensions. Later on, Caflisch, Klapper and Steele [3] extended the well-known result of Beale–Kato–Majda [2] for the incompressible Euler equations to the three dimensional IMHD equations. They showed precisely that if the smooth solution (u,b) satisfies the condition

$$\int_0^T \|\nabla\times u\|_{L^\infty} + \|\nabla\times b\|_{L^\infty} dt < \infty, \tag{1.2}$$

*Received by the editors February 13, 2006; accepted for publication (in revised form) September 22, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/sima/38-6/65200.html

†Laboratoire d'Analyse et de Mathématiques Appliquées, Université de Marne-la-Vallée, Cité Descartes-5, bd Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2 France (marco.cannone@univ-mlv.fr).

‡Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, People's Republic of China (qleichen@yahoo.com.cn, miao_changxing@iapcm.ac.cn). These authors were partly supported by the National Natural Science Foundation (NNSF) of China, No. 10571016.

then the solution (u,b) can be extended beyond $t=T$, namely, for some $T<\tilde{T}$, $(u,b)\in C([0,\tilde{T});H^s)\cap C^1([0,\tilde{T});H^{s-1})$. In other words, let $[0,T)$ be the maximal time interval to the smooth solution (u,b) for (IMHD); then (u,b) blows up at T iff

$$\lim_{\varepsilon\to 0}\int_{\varepsilon_0}^{T-\varepsilon}\|\nabla\times u\|_{L^\infty}+\|\nabla\times b\|_{L^\infty}dt=\infty\qquad\forall\ 0\le\varepsilon_0<T. \tag{1.3}$$

Recently, Zhang and Liu [17] refined the blow-up condition (1.3) to

$$\lim_{\varepsilon\to 0}\int_{\varepsilon_0}^{T-\varepsilon}\|\nabla\times u\|_{\dot{B}^0_{\infty,\infty}}+\|\nabla\times b\|_{\dot{B}^0_{\infty,\infty}}dt=\infty\quad\forall\ 0\le\varepsilon_0<T.$$

Here and thereafter, $\dot{B}^s_{p,q}$ denotes the homogenous Besov space, whose definition will be given in section 2.

Recently, for the incompressible Euler equations,

$$\text{(IE)}\qquad\begin{cases}u_t+u\cdot\nabla u+\nabla p=0,\\ \nabla\cdot u=0,\\ u(0,x)=u_0(x),\end{cases} \tag{1.4}$$

Planchon [14] was able to relax previous blow-up conditions such as

$$\lim_{\varepsilon\to 0}\int_{\varepsilon_0}^{T-\varepsilon}\|\omega(t)\|_\infty dt=\infty\qquad\forall\ 0\le\varepsilon_0<T\qquad\text{(see, e.g., [2])},$$
$$\lim_{\varepsilon\to 0}\int_{\varepsilon_0}^{T-\varepsilon}\|\omega(t)\|_{BMO}dt=\infty\qquad\forall\ 0\le\varepsilon_0<T\qquad\text{(see, e.g., [10])},$$
$$\lim_{\varepsilon\to 0}\int_{\varepsilon_0}^{T-\varepsilon}\|\omega(t)\|_{\dot{B}^0_{\infty,\infty}}dt=\infty\qquad\forall\ 0\le\varepsilon_0<T\qquad\text{(see, e.g., [11])},$$

and he established an improved blow-up criterion in the framework of mixed time-space Besov spaces. His result is the following: there exists a positive constant M_0 such that if

$$\lim_{\varepsilon\to 0}\sup_{j\in\mathbb{Z}}\int_{T-\varepsilon}^{T}\|\Delta_j\omega(t)\|_\infty dt\ge M_0, \tag{1.5}$$

then u cannot be continued beyond $t=T$. Here $\omega=\nabla\times u$ is the vorticity and Δ_j denotes a frequency localization operator at $|\xi|\approx 2^j$; see section 2 for the definition.

Inspired by [14], we want to obtain the corresponding result for the IMHD equations. Unfortunately, the method of [16, 14] does not apply directly, since there is not enough evidence to show if the solution u remains smooth past time T from the deduced inequality (see the last inequality in [14]). More precisely, we cannot infer that $f(T)$ can be dominated by $f(T-\varepsilon)$ from the following inequality:

$$f(T)\le f(T-\varepsilon)+\delta(\varepsilon)f(T)\log(1+f(T)),$$

where $\delta(\varepsilon)$ is a function such that $\delta(\varepsilon)\to 0$ when ε tends to zero.

To overcome this difficulty, we first obtain a losing estimate for the IMHD equations (1.1), which is motivated by [7, 8], and further establishes a blow-up criterion of smooth solution for the IMHD equations (1.1). Let us point out that losing estimates

might be of independent interest, as they can be used to establish improved blow-up criteria of smooth solutions for the incompressible Navier–Stokes equations

$$\text{(NS)} \quad \begin{cases} u_t - \nu\Delta u + u\cdot\nabla u + \nabla p = 0, \\ \nabla\cdot u = 0, \\ u(0,x) = u_0(x), \end{cases} \tag{1.6}$$

and MHD equations with dissipation and resistance

$$\text{(MHD)} \quad \begin{cases} u_t - \nu\Delta u + u\cdot\nabla u = -\nabla p - \frac{1}{2}\nabla b^2 + b\cdot\nabla b, \\ b_t - \eta\Delta u + u\cdot\nabla b = b\cdot\nabla u, \\ \nabla\cdot u = \nabla\cdot b = 0, \\ u(0,x) = u_0(x), \quad b(0,x) = b_0(x). \end{cases} \tag{1.7}$$

We will prove the losing estimate for the IMHD equation (1.1) is the following theorem.

THEOREM 1.1. *Let (u,b) be a smooth solution to* (1.1) *with $(u_0,b_0)\in B^s_{p,q}$, $s>\frac{n}{p}+1$, $1\le p,q<\infty$. Then there exists a positive constant $C_0>0$ such that for $\lambda\ge C_0$,*

$$\mathcal{M}(u)+\mathcal{M}(b)\lesssim \|u_0\|_{\dot B^s_{p,q}}+\|b_0\|_{\dot B^s_{p,q}}, \tag{1.8}$$

where

$$\mathcal{M}(f)\triangleq \Big\|\sup_{t\in[0,T)} 2^{js-\Psi_\lambda(t)}\|\Delta_j f(t)\|_p\Big\|_{\ell^q(\mathbb{Z})} \quad \textit{with}$$

$$\Psi_\lambda(t',t)\triangleq \lambda\int_{t'}^{t}\|\nabla u(t'')\|_\infty+\|\nabla b(t'')\|_\infty dt'' \quad \textit{and} \quad \Psi_\lambda(t)=\Psi_\lambda(0,t).$$

As an application to Theorem 1.1, and inspired by [14], we obtain a similar blow-up criterion for the IMHD equations (1.1).

THEOREM 1.2. *Let $(u_0,b_0)\in B^s_{p,q}$, $s>\frac{n}{p}+1$, $1\le p,q<\infty$. Suppose that $(u,b)\in C([0,T);B^s_{p,q})\cap C^1([0,T);B^{s-1}_{p,q})$ is the smooth solution to* (1.1). *There exists an absolute constant $M>0$ such that*

(i) *If*

$$\limsup_{\varepsilon\to 0}\sum_{j\in\mathbb{Z}}\int_{T-\varepsilon}^{T}(\|\Delta_j(\nabla\times u)\|_\infty+\|\Delta_j(\nabla\times b)\|_\infty)dt=\delta<M, \tag{1.9}$$

then $\delta=0$, and the solution (u,b) can be extended past time $t=T$.

(ii) *If*

$$\limsup_{\varepsilon\to 0}\sum_{j\in\mathbb{Z}}\int_{T-\varepsilon}^{T}(\|\Delta_j(\nabla\times u)\|_\infty+\|\Delta_j(\nabla\times b)\|_\infty)dt\ge M, \tag{1.10}$$

then the solution blows up at $t=T$.

Remark 1.1. In Theorem 1.2, M can be viewed as a threshold of blow-up. Planchon gave an exact explanation for the appearance of M; see [14] for details.

Remark 1.2. Note that $B^{s-1}_{p,q}$ is a Banach algebra for $s>\frac{n}{p}+1$. One can easily prove that there exists a unique smooth solution $(u,b)\in C([0,T);B^s_{p,q})\cap C^1([0,T);B^{s-1}_{p,q})$ to (1.1) by standard method; see [16] for details.

Remark 1.3. (i) In the case of $b = 0$, (IMHD) can be read as the incompressible Euler equations (IE), and what is proved in [14] is a straightforward consequence of Theorem 1.2.

(ii) Making use of losing estimate techniques, we can establish an improved blow-up criterion of smooth solution in the framework of mixed time-space Besov space for the incompressible Navier–Stokes equations (NS) and MHD equations (MHD).

Notation. Throughout the paper, C stands for a "harmless" constant, and we will use the notation $A \lesssim B$ as an equivalent to $A \leq CB$, $A \approx B$ as $A \lesssim B$ and $B \lesssim A$.

2. Preliminaries. Let us recall the Littlewood–Paley decomposition (see also [1, 15]) and then establish a certain kind of logarithmic Sobolev inequalities in the framework of mixed time-space Besov space. Let $\mathcal{S}(\mathbb{R}^3)$ be the Schwartz class of rapidly decreasing functions. Given $f \in \mathcal{S}(\mathbb{R}^n)$, its Fourier transform $\mathcal{F}f = \hat{f}$ is defined by

$$\hat{f}(\xi) = (2\pi)^{-\frac{n}{2}} \int_{\mathbb{R}^n} e^{-ix\cdot\xi} f(x)dx.$$

Choose two nonnegative radial functions χ, $\varphi \in \mathcal{S}(\mathbb{R}^n)$, supported in $B = \{\xi \in \mathbb{R}^n, |\xi| \leq \frac{4}{3}\}$ and $\mathcal{C} = \{\xi \in \mathbb{R}^n, \frac{3}{4} \leq |\xi| \leq \frac{8}{3}\}$, respectively, such that

$$\chi(\xi) + \sum_{j\geq 0} \varphi(2^{-j}\xi) = 1, \quad \xi \in \mathbb{R}^n,$$
$$\sum_{j\in\mathbb{Z}} \varphi(2^{-j}\xi) = 1, \quad \xi \in \mathbb{R}^n\backslash\{0\}.$$

Setting $\varphi_j(\xi) = \varphi(2^{-j}\xi)$, one easily verifies that $\operatorname{supp}\varphi_j \cap \operatorname{supp}\varphi_{j'} = \emptyset$ if $|j - j'| \geq 2$ and $\operatorname{supp}\chi \cap \operatorname{supp}\varphi_j = \emptyset$ if $j \geq 1$. Let $h = \mathcal{F}^{-1}\varphi$ and $\tilde{h} = \mathcal{F}^{-1}\chi$; we define the dyadic blocks as follows:

$$\Delta_j f = \varphi(2^{-j}D)f = 2^{jn} \int_{\mathbb{R}^n} h(2^j y)f(x - y)dy,$$
$$S_j f = \sum_{k\leq j-1} \Delta_k f = \chi(2^{-j}D)f = 2^{jn} \int_{\mathbb{R}^n} \tilde{h}(2^j y)f(x - y)dy.$$

Informally, $\Delta_j = S_j - S_{j-1}$ is a frequency projection to the annulus $\{|\xi| \approx 2^j\}$, while S_j is a frequency projection to the ball $\{|\xi| \lesssim 2^j\}$.

Making use of Littlewood–Paley's decomposition, we give the definition of the homogenous Besov space. Let $s \in \mathbb{R}, 1 \leq p, q \leq \infty$; the homogenous Besov space $\dot{B}^s_{p,q}$ is defined by

$$\dot{B}^s_{p,q} = \{f \in \mathcal{Z}'(\mathbb{R}^n); \|f\|_{\dot{B}^s_{p,q}} < \infty\},$$

where

$$\|f\|_{\dot{B}^s_{p,q}} = \begin{cases} \Big(\displaystyle\sum_{j=-\infty}^{\infty} 2^{jsq} \|\Delta_j f\|_p^q \Big)^{\frac{1}{q}} & \text{for} \quad q < \infty, \\ \displaystyle\sup_{j\in\mathbb{Z}} 2^{js} \|\Delta_j f\|_p & \text{for} \quad q = \infty, \end{cases}$$

and $\mathcal{Z}'(\mathbb{R}^n)$ denotes the dual space of $\mathcal{Z}(\mathbb{R}^n) = \{f \in \mathcal{S}(\mathbb{R}^n); \partial^\alpha \hat{f}(0) = 0; \forall \alpha \in \mathbb{N}^n \text{ multi-index}\}$ and can be identified by the quotient space of $\mathcal{S}'/\mathcal{P}$ with the polynomials space $\mathcal{P}$.

For $s > 0$, we define the inhomogeneous Besov space as

$$B^s_{p,q} = \{f \in \mathcal{S}'(\mathbb{R}^n);\ \|f\|_{B^s_{p,q}} < \infty\},$$

where

$$\|f\|_{B^s_{p,q}} = \|f\|_p + \|f\|_{\dot{B}^s_{p,q}}.$$

For details, we refer to [1, 15].

In addition to normal time-space Besov space such as $L^r(I; B^s_{p,q})$ or $L^r(I; \dot{B}^s_{p,q})$, we also use another kind of mixed time-space Besov space, used previously in [6]; see also [4]. Let $s \in \mathbb{R}, 1 \le r, p, q \le \infty$, $0 < T \le +\infty$; the space $\widetilde{L}^r_T(\dot{B}^s_{p,q})$ is defined by

$$\widetilde{L}^r_T(\dot{B}^s_{p,q}) = \{f \in \mathcal{D}'((0,T); \mathcal{Z}'(\mathbb{R}^n));\ \|f\|_{\widetilde{L}^r_T(\dot{B}^s_{p,q})} < \infty\},$$

where

$$\|f\|_{\widetilde{L}^r_T(\dot{B}^s_{p,q})} = \left\| 2^{js} \left(\int_0^T \|\Delta_j f(t)\|_p^r dt \right)^{\frac{1}{r}} \right\|_{\ell^q}.$$

Before establishing logarithmic Sobolev inequality in the framework of mixed time-space Besov space, we introduce two well-known lemmas which will be used repeatedly in this paper.

LEMMA 2.1 (Bernstein's estimates). *Let $1 \le p \le q \le \infty$. Assume that $f \in L^p$; then there exist constants C independent of f, j such that*

$$\operatorname{supp}\hat{f} \subset \{|\xi| \lesssim 2^j\} \Rightarrow \|\partial^\alpha f\|_q \le C 2^{j|\alpha| + jn(\frac{1}{p} - \frac{1}{q})} \|f\|_p, \tag{2.1}$$

$$\operatorname{supp}\hat{f} \subset \{|\xi| \approx 2^j\} \Rightarrow \|f\|_p \le C 2^{-j|\alpha|} \sup_{|\beta| = |\alpha|} \|\partial^\beta f\|_p. \tag{2.2}$$

This lemma is classic and for a proof, see [13].

LEMMA 2.2 (commutator estimate). *Let $j \in \mathbb{Z}$, $1 \le p \le \infty$. Assume that $f \in L^p$, $\nabla g \in L^\infty$; then there exists a positive constant C independent of f, g, and j such that*

$$\|[\Delta_j,\, g] f\|_p \le C 2^{-j} \|\nabla g\|_\infty \|f\|_p. \tag{2.3}$$

The proof is rather standard and can be found in [5, 7, 14].

Now we are in the position to prove a certain kind of logarithmic Sobolev inequalities initiated in [14] in terms of mixed time-space Besov space, which will play an important role in the proof of Theorem 1.1 as well.

PROPOSITION 2.1. *Let $1 \le p < \infty$, $1 \le q \le \infty$, and $s > \frac{n}{p} + 1$. Assume that $f \in \widetilde{L}^1_T(\dot{B}^0_{\infty,\infty}) \cap L^\infty_T(B^{s-1}_{p,q})$. Then the following inequality holds:*

$$\int_0^T \|f(t)\|_\infty dt \le C \left(1 + \sup_j \int_0^T \|\Delta_j f\|_\infty dt (1 + \log^+(T \|f\|_{L^\infty_T(B^{s-1}_{p,q})})) \right), \tag{2.4}$$

where $\log^+ x = \log x$, for $x > 1$; $\log^+ x = 0$, for $x \le 1$; and C is an absolute constant independent of f, T.

Proof. Using the Littlewood–Paley decomposition, we decompose f as follows:

$$f = S_{-N} f + \sum_{j=-N}^{N} \Delta_j f + \sum_{j>N} \Delta_j f \triangleq f_1 + f_2 + f_3,$$

where N is a positive integer to be chosen later. We first estimate f_1. It follows from (2.1) that

$$\int_0^T \|f_1\|_\infty dt \le \int_0^T 2^{-N\frac{n}{p}} \|S_{-N} f\|_p dt \le CT 2^{-N\frac{n}{p}} \|f\|_p. \tag{2.5}$$

Second, it is obvious that

$$\int_0^T \|f_2\|_\infty dt \le (2N+1) \sup_j \int_0^T \|\Delta_j f\|_\infty dt. \tag{2.6}$$

Finally, we turn to f_3. In view of (2.1) together with the definition of Besov space, it follows that

$$\begin{aligned} \int_0^T \|f_3\|_\infty dt &\le \int_0^T \sum_{j>N} 2^{j(\frac{n}{p}-s+1)} 2^{j(s-1)} \|\Delta_j f\|_p dt \\ &\le CT 2^{-N(s-\frac{n}{p}-1)} \|f\|_{L^\infty_T(B^{s-1}_{p,q})}, \end{aligned} \tag{2.7}$$

where we have used the Hölder inequality in the last inequality of (2.7).

Combining (2.5)–(2.7) and setting $\alpha = \min(\frac{n}{p}, s - \frac{n}{p} - 1)$, one easily verifies that

$$\int_0^T \|f\|_\infty dt \le C \left(T 2^{-\alpha N} \|f\|_{L^\infty_T(B^{s-1}_{p,q})} + N \sup_j \int_0^T \|\Delta_j f\|_\infty dt \right). \tag{2.8}$$

Now we choose N such that $T 2^{-\alpha N} \|f\|_{L^\infty_T(B^{s-1}_{p,q})} \le 1$, i.e.,

$$N \ge \frac{\log(T \|f\|_{L^\infty_T(B^{s-1}_{p,q})})}{\alpha \log 2}.$$

Then our desired estimate (2.4) follows from the above inequality and (2.8). ∎

3. A losing a priori estimate. Let us begin with the proof of Theorem 1.1.

Proof of Theorem 1.1. We first consider the symmetrized IMHD equations rewritten in terms of the classical Elsasser variables [9]. Denote

$$\begin{cases} z^+ = u + b, \\ z^- = u - b. \end{cases}$$

Then (IMHD) can be reduced to

$$(\mathrm{IMHD})_* \quad \begin{cases} \partial_t z^+ + (z^- \cdot \nabla) z^+ = -\nabla \pi, \\ \partial_t z^- + (z^+ \cdot \nabla) z^- = -\nabla \pi, \\ \mathrm{div} z^+ = \mathrm{div} z^- = 0, \\ z^+(0) = z_0^+ = u_0 + b_0, \quad z^-(0) = z_0^- = u_0 - b_0, \end{cases} \tag{3.1}$$

where $\pi = p + \frac{1}{2} b^2$.

Let (z^+, z^-) be a smooth solution to $(\mathrm{IMHD})_*$ with $(z_0^+, z_0^-) \in \dot{B}^s_{p,q}$, $s > \frac{n}{p} + 1$, $1 \le p, q < \infty$. Then the proof of Theorem 1.1 can be reduced to establish the following estimate: for λ sufficient large, such as $\lambda > C_0$, there holds

$$\alpha_T^q + \beta_T^q \lesssim \|z_0^+\|^q_{\dot{B}^s_{p,q}} + \|z_0^-\|^q_{\dot{B}^s_{p,q}}, \tag{3.2}$$

where $\alpha_T = \|\alpha_{j,T}\|_{\ell^q}$, $\beta_T = \|\beta_{j,T}\|_{\ell^q}$ and

$$\alpha_{j,T} = \sup_{t\in[\,0,T)} 2^{js-\Phi_\lambda(t)}\|\Delta_j z^+(t)\|_p,$$
$$\beta_{j,T} = \sup_{t\in[\,0,T)} 2^{js-\Phi_\lambda(t)}\|\Delta_j z^-(t)\|_p \quad \text{with}$$
$$\Phi_\lambda(t',t) \triangleq \lambda\int_{t'}^{t} \|\nabla z^+(t'')\|_\infty + \|\nabla z^-(t'')\|_\infty dt'' \quad \text{and} \quad \Phi_\lambda(t) = \Phi_\lambda(0,t).$$

Denote $\mathbb{P}$ the matrix operator with entries $(\delta_{ij} - R_iR_j)$; here δ_{ij} is the Kronecker symbol and R_i, R_j are Riesz transforms. As such $\mathbb{P}$ is the well-known projection onto divergence-free vector fields. Applying $\mathbb{P}$ to two sides of (3.1) and taking the operation Δ_j on both sides of the resulting equations, we get

$$\begin{cases} \partial_t z_j^+ + \mathbb{P}\Delta_j((z^-\cdot\nabla)z^+) = 0, \\ \partial_t z_j^- + \mathbb{P}\Delta_j((z^+\cdot\nabla)z^-) = 0, \end{cases} \tag{3.3}$$

where

$$z_j = \Delta_j z.$$

Multiplying the first equation of (3.3) by $|z_j^+|^{p-2}z_j^+$ and the second one by $|z_j^-|^{p-2}z_j^-$, integrating the resulting equations in $\mathbb{R}^n$ with respect to the space variable gives

$$\begin{cases} \dfrac{1}{p}\dfrac{d}{dt}\|z_j^+\|_p^p = -\displaystyle\int_{\mathbb{R}^n} \mathbb{P}\Delta_j((z^-\cdot\nabla)z^+)\,|z_j^+|^{p-2}z_j^+ dx, \\ \dfrac{1}{p}\dfrac{d}{dt}\|z_j^-\|_p^p = -\displaystyle\int_{\mathbb{R}^n} \mathbb{P}\Delta_j((z^+\cdot\nabla)z^-)\,|z_j^-|^{p-2}z_j^- dx. \end{cases}$$

In what follows, we need to estimate the convection terms

$$\int_{\mathbb{R}^n} \mathbb{P}\Delta_j((z^-\cdot\nabla)z^+)\,|z_j^+|^{p-2}z_j^+ dx \quad \text{and} \quad \int_{\mathbb{R}^n} \mathbb{P}\Delta_j((z^+\cdot\nabla)z^-)\,|z_j^-|^{p-2}z_j^- dx. \tag{3.4}$$

We only deal with the first term appearing in (3.4), as the second one can be treated in the same way. Now we decompose $((z^-\cdot\nabla)z^+)$ as a paraproduct

$$\begin{aligned}(z^-\cdot\nabla)z^+ &= \sum_k (S_{k-1}z^-\cdot\nabla)z_k^+ + \sum_k (z_k^-\cdot\nabla)S_{k-1}z^+ + \sum_{|k-k'|\le 1}(z_{k'}^-\cdot\nabla)z_k^+ \\ &\triangleq I_1 + I_2 + I_3. \end{aligned} \tag{3.5}$$

It is easy to see that we only need to preserve the vector structure for I_1. For the remaining two terms one can forget about vectors and think of them as scalar.

• I_2 estimates. Applying the Hölder inequality in conjunction with the boundedness of $\mathbb{P}$ on L^p $(1 < p < \infty)$, we get

$$\left|\int_{\mathbb{R}^n} \mathbb{P}\Delta_j(I_2)|z_j^+|^{p-2}z_j^+ dx\right| \lesssim \sum_{k\sim j}\|\nabla S_{k-1}z^+\|_\infty\|z_k^-\|_p\|z_j^+\|_p^{p-1}, \tag{3.6}$$

where $k\sim j$ stands for $|k-j|\le 1$ owing to the support conditions.

• I_3 estimates. Noting that $\nabla\cdot z_j^{\pm}=0$, one easily shows that

$$\int_{\mathbb{R}^n}\mathbb{P}\Delta_j(I_3)|z_j^+|^{p-2}z_j^+dx=\sum_{1\le\ell\le n}\int_{\mathbb{R}^n}\Delta_j\Big(\sum_{j\lesssim k\sim k'}\mathbb{P}\nabla\cdot(z_{k'}^-\otimes z_k^+)\Big)^{(\ell)}|z_j^{+(\ell)}|^{p-2}z_j^{+(\ell)}dx.$$

If $p\ge 2$, integrating by parts together with the Hölder inequality yields

$$\Big|\int_{\mathbb{R}^n}\mathbb{P}\Delta_j(I_3)|z_j^+|^{p-2}z_j^+dx\Big|\lesssim(p-1)\sum_{j\lesssim k\sim k'}\|z_{k'}^-\|_p\|z_k^+\|_p\|z_j^+\|_p^{p-2}\|\nabla z_j^+\|_\infty. \tag{3.7}$$

If $p<2$, it follows by the Hölder inequality

$$\Big|\int_{\mathbb{R}^n}\mathbb{P}\Delta_j(I_3)|z_j^+|^{p-2}z_j^+dx\Big|\lesssim\sum_{j\lesssim k\sim k'}\|\nabla z_k^+\|_\infty\|z_{k'}^-\|_p\|z_j^+\|_p^{p-1}. \tag{3.8}$$

• I_1 estimates. Observe that

$$\begin{aligned}\mathbb{P}\Delta_j\sum_k(S_{k-1}z^-\cdot\nabla)z_k^+&=\mathbb{P}\Delta_j(S_{j-1}z^-\cdot\nabla)\sum_{k\sim j}z_k^++\mathbb{P}\Delta_j\sum_{k\sim k'\sim j}(z_{k'}^-\cdot\nabla)z_k^+\\&\triangleq II_1+II_2.\end{aligned}$$

We have by the Hölder inequality

$$\Big|\int_{\mathbb{R}^n}II_2|z_j^+|^{p-2}z_j^+dx\Big|\lesssim\sum_{k\sim k'\sim j}\|\nabla z_k^+\|_\infty\|z_{k'}^-\|_p\|z_j^+\|_p^{p-1}. \tag{3.9}$$

Furthermore, we can rewrite II_1 as

$$II_1=\sum_{1\le\ell\le n}C_\ell\partial_\ell\sum_{k\sim j}z_k^++(S_{j-1}z^-\cdot\nabla)\mathbb{P}\Delta_j\sum_{k\sim j}z_k^+\triangleq III_1+III_2,$$

where C_ℓ is the commutator denoted by

$$C_\ell=[\mathbb{P}\Delta_j,\,S_{j-1}z^{-(\ell)}Id].$$

It is easy to see $\sum_{k\sim j}\mathbb{P}\Delta_jz_k^+=z_j^+$. Making use of $\mathrm{div}z_j^+=0$ and integrating by parts, we have

$$\begin{aligned}\int_{\mathbb{R}^n}III_2|z_j^+|^{p-2}z_j^+dx&=\sum_{1\le\ell\le n}\int_{\mathbb{R}^n}S_{j-1}z^{-(\ell)}\partial_\ell z_j^+|z_j^+|^{p-2}z_j^+dx\\&=\sum_{1\le\ell\le n}\int_{\mathbb{R}^n}\partial_\ell(S_{j-1}z^{-(\ell)}z_j^+)|z_j^+|^{p-2}z_j^+dx\\&=-(p-1)\sum_{1\le\ell\le n}\int_{\mathbb{R}^n}S_{j-1}z^{-(\ell)}z_j^+\partial_\ell z_j^+|z_j^+|^{p-2}dx,\end{aligned}$$

which implies that

$$\int_{\mathbb{R}^n}III_2|z_j^+|^{p-2}z_j^+dx=0. \tag{3.10}$$

On the other hand, from the commutator estimate (2.3), we obtain

$$\left|\int_{\mathbb{R}^n} III_1|z_j^+|^{p-2}z_j^+dx\right| \lesssim \|\nabla S_{j-1}z^-\|_\infty\|z_j^+\|_p^p. \tag{3.11}$$

Combining (3.6)–(3.7) and (3.9)–(3.11), we infer that

$$\begin{aligned}\frac{1}{p}\frac{d}{dt}\|z_j^+\|_p^p &\lesssim \sum_{k\sim j}\Big(\|\nabla S_{j-1}z^+\|_\infty\|z_k^-\|_p\|z_j^+\|_p^{p-1}+\|\nabla S_{j-1}z^-\|_\infty\|z_k^+\|_p\|z_j^+\|_p^{p-1}\Big)\\ &\quad+\sum_{j\lesssim k'\sim k}\|z_k^+\|_p\|z_{k'}^-\|_p\|z_j^+\|_p^{p-2}\|\nabla z_j^+\|_\infty\\ &\lesssim \sum_{k\sim j}\Big(\|\nabla z^+\|_\infty\|z_k^-\|_p\|z_j^+\|_p^{p-1}+\|\nabla z^-\|_\infty\|z_k^+\|_p\|z_j^+\|_p^{p-1}\Big)\\ &\quad+\sum_{j\lesssim k'\sim k}\|z_k^+\|_p\|z_{k'}^-\|_p\|z_j^+\|_p^{p-2}\|\nabla z^+\|_\infty,\end{aligned}$$

where use has been made of the inequalities $\|\nabla S_{j-1}z^+\|_\infty\le\|\nabla z^+\|_\infty$ and $\|\nabla z_k^+\|_\infty\le\|\nabla z^+\|_\infty$ to derive the second estimate. Thus we deduce

$$\begin{aligned}\frac{d}{dt}\|z_j^+\|_p^2 &\lesssim \sum_{k\sim j}\Big(\|\nabla z^+\|_\infty\|z_k^-\|_p+\|\nabla z^-\|_\infty\|z_k^+\|_p\Big)\|z_j^+\|_p\\ &\quad+\sum_{j\lesssim k'\sim k}\|z_k^+\|_p\|z_{k'}^-\|_p\|\nabla z^+\|_\infty,\end{aligned}$$

which implies

$$\begin{aligned}\frac{d}{dt}\|z_j^+\|_p^q &\lesssim \frac{q}{2}\sum_{k\sim j}\Big(\|\nabla z^+\|_\infty\|z_k^-\|_p+\|\nabla z^-\|_\infty\|z_k^+\|_p\Big)\|z_j^+\|_p^{q-1}\\ &\quad+\frac{q}{2}\sum_{j\lesssim k'\sim k}\|z_k^+\|_p\|z_{k'}^-\|_p\|\nabla z^+\|_\infty\|z_j^+\|_p^{q-2}.\end{aligned}\tag{3.12}$$

In the same way as leading to (3.12), we have for z_j^-

$$\begin{aligned}\frac{d}{dt}\|z_j^-\|_p^q &\lesssim \frac{q}{2}\sum_{k\sim j}\Big(\|\nabla z^-\|_\infty\|z_k^+\|_p+\|\nabla z^+\|_\infty\|z_k^-\|_p\Big)\|z_j^-\|_p^{q-1}\\ &\quad+\frac{q}{2}\sum_{j\lesssim k'\sim k}\|z_k^-\|_p\|z_{k'}^+\|_p\|\nabla z^-\|_\infty\|z_j^-\|_p^{q-2}.\end{aligned}\tag{3.13}$$

Integrating (3.13) over $[0,t)$ with respect to time variable τ, and then multiplying by $2^{q(js-\Phi_\lambda(t))}$ both sides of the resulting inequality, we get

$$\begin{aligned}2^{q(js-\Phi_\lambda(t))}\|z_j^+\|_p^q(t) &\lesssim 2^{jqs}\|z_j^+\|_p^q(0)+\int_0^t 2^{-q\Phi_\lambda(\tau,t)}(\|\nabla z^+\|_\infty+\|\nabla z^-\|_\infty)(\tau)\\ &\quad\times\Big(\sum_{k\sim j}2^{q(js-\Phi_\lambda(\tau))}\|z_j^+(\tau)\|_p^{q-1}(\|z_k^-\|_p+\|z_k^+\|_p)(\tau)+\sum_{j\lesssim k'\sim k}2^{2(j-k)s}\\ &\quad\times 2^{ks-\Phi_\lambda(\tau)}\|z_k^+\|_p 2^{k's-\Phi_\lambda(\tau)}\|z_{k'}^-\|_p 2^{(q-2)(js-\Phi_\lambda(\tau))}\|z_j^+\|_p^{q-2}\Big)d\tau.\end{aligned}$$

Taking the supremum over $[0,T)$ on both sides of the above inequality, we deduce that

$$\alpha_{j,T}^q \lesssim \alpha_{j,0}^q + \sup_{t\in[0,T)} \int_0^t 2^{-q\Phi_\lambda(\tau,t)}(\|\nabla z^+\|_\infty + \|\nabla z^-\|_\infty)(\tau)d\tau \bigg(\alpha_{j,T}^{q-1} \sum_{k\sim j}(\beta_{k,T}+\alpha_{k,T})$$

$$+\alpha_{j,T}^{q-2} \sum_{j\lesssim k'\sim k} 2^{2(j-k)s}\alpha_{k,T}\beta_{k',T}\bigg). \tag{3.14}$$

Moreover, integrating by parts yields

$$\int_0^t 2^{-q\Phi_\lambda(\tau,t)}(\|\nabla z^+\|_\infty + \|\nabla z^-\|_\infty)(\tau)d\tau \le \frac{1}{q\lambda\log 2}. \tag{3.15}$$

• When $q \ge 2$. Taking the sum over j of (3.14); then using the above estimate and Hölder inequality leads to

$$\alpha_T^q \lesssim \|z_0^+\|_{\dot B^s_{p,q}}^q + \frac{1}{q\lambda\log 2}(\alpha_T^{q-1}\beta_T + \alpha_T^q), \tag{3.16}$$

where use has been made of the $\ell^1 \to \ell^{\frac{q}{2}}$ convolution to the second sum of the right side of (3.14). Let us go back to (3.13); arguing similarly as in deriving (3.16), we obtain

$$\beta_T^q \lesssim \|z_0^-\|_{\dot B^s_{p,q}}^q + \frac{1}{q\lambda\log 2}(\beta_T^{q-1}\alpha_T + \beta_T^q). \tag{3.17}$$

Summing up (3.16) and (3.17), together with the Hölder inequality, yields

$$\alpha_T^q + \beta_T^q \lesssim \|z_0^+\|_{\dot B^s_{p,q}}^q + \|z_0^-\|_{\dot B^s_{p,q}}^q + \frac{1}{q\lambda\log 2}(\alpha_T^q + \beta_T^q). \tag{3.18}$$

Now if we choose λ large enough such that

$$\lambda \gtrsim \frac{2}{q\log 2}, \tag{3.19}$$

then

$$\alpha_T^q + \beta_T^q \lesssim \|z_0^+\|_{\dot B^s_{p,q}}^q + \|z_0^-\|_{\dot B^s_{p,q}}^q. \tag{3.20}$$

• When $q < 2$. One may substitute (3.7) by (3.8) in the estimate of I_3. It is not hard to verify that

$$\alpha_{j,T}^q \lesssim \alpha_{j,0}^q + \sup_{t\in[0,T)} \int_0^t 2^{-q\Phi_\lambda(\tau,t)}(\|\nabla z^+\|_\infty + \|\nabla z^-\|_\infty)(\tau)d\tau \bigg(\alpha_{j,T}^{q-1} \sum_{k\sim j}(\beta_{k,T}+\alpha_{k,T})$$

$$+\alpha_{j,T}^{q-1} \sum_{j\lesssim k'\sim k} 2^{(j-k)s}\beta_{k',T}\bigg).$$

By exactly the same procedure as the case of $q \ge 2$, we can get the required estimate (3.2). This completes the proof of Theorem 1.1. □

4. The blow-up criterion. In this section, we want to prove Theorem 1.2.

Proof of Theorem 1.2. Assume that

$$\limsup_{\varepsilon\to 0}\sup_{j}\int_{T-\varepsilon}^{T}(\|\Delta_j\omega\|_\infty+\|\Delta_j J\|_\infty)dt=\delta<M. \tag{4.1}$$

Then the solution (u,b) of (IMHD) can be extended past time $t=T$, where $\omega=\nabla\times u$ and $J=\nabla\times b$ are the vorticity with respect to u and b, respectively. This fact is equivalent to the blow-up criterion (1.9) in Theorem 1.2.

Let $(u,b)\in C_T(B^s_{p,q})$, $s>\frac{n}{p}+1$, $1\le p,q<\infty$ be the smooth solution to (IMHD); then (z^+,z^-) is the smooth solution to $(\mathrm{IMHD})_*$ and (4.1) can be read as

$$\limsup_{\varepsilon\to 0}\sup_{j}\int_{T-\varepsilon}^{T}(\|\Delta_j(\nabla\times z^+)\|_\infty+\|\Delta_j(\nabla\times z^-)\|_\infty)dt=\delta<2M. \tag{4.2}$$

Thanks to the Biot–Savard law,

$$\nabla u=(-\Delta)^{-1}\nabla\nabla\times\omega,\quad \nabla b=(-\Delta)^{-1}\nabla\nabla\times J,$$

and $\|R_k\Delta_j f\|_p\le C_0\|\Delta_j f\|_p$, $(1\le p\le\infty)$; it follows that (4.2) is equivalent to

$$\limsup_{\varepsilon\to 0}\sup_{j}\int_{T-\varepsilon}^{T}(\|\Delta_j\nabla z^+\|_\infty+\|\Delta_j\nabla z^-\|_\infty)dt=\delta_0<2C_0M, \tag{4.3}$$

where C_0 is a constant from the boundedness of Riesz transform R_k, $1\le k\le n$.

We are now in position to set the blow-up criterion. In order to do this, let us estimate

$$\|z^+(t)\|_p\le\|z_0^+\|_p+\int_0^t\|\nabla z^+(\tau)\|_\infty\|z^-(\tau)\|_p d\tau.$$

Multiplying by $2^{-\Phi_\lambda(t)}$, both sides of the above inequality yield

$$2^{-\Phi_\lambda(t)}\|z^+(t)\|_p\lesssim\|z_0^+\|_p+\int_0^t 2^{-\Phi_\lambda(\tau,t)}\|\nabla z^+(\tau)\|_\infty 2^{-\Phi_\lambda(\tau)}\|z^-(\tau)\|_p d\tau.$$

This along with (3.15) implies that

$$\sup_{t\in[0,T)}2^{-\Phi_\lambda(t)}\|z^+(t)\|_p\lesssim\|z_0^+\|_p+\frac{1}{\lambda\log 2}\sup_{[0,T)}2^{-\Phi_\lambda(t)}\|z^-(t)\|_p.$$

An analogous argument leading to the above estimate allows us to get

$$\sup_{t\in[0,T)}2^{-\Phi_\lambda(t)}\|z^-(t)\|_p\lesssim\|z_0^-\|_p+\frac{1}{\lambda\log 2}\sup_{[0,T)}2^{-\Phi_\lambda(t)}\|z^+(t)\|_p.$$

If $\lambda\gtrsim\frac{2}{\log 2}$, then

$$\sup_{t\in[0,T)}2^{-\Phi_\lambda(t)}(\|z^+(t)\|_p+\|z^-(t)\|_p)\lesssim\|z_0^+\|_p+\|z_0^-\|_p. \tag{4.4}$$

Let us define

$$\gamma_T\triangleq\sup\left((\alpha_T^q+\beta_T^q)^{\frac{1}{q}},\ \sup_{[0,T)}2^{-\Phi_\lambda(t)}(\|z^+(t)\|_p+\|z^-(t)\|_p)\right).$$

This together with (3.2) and (4.4) implies that

$$\gamma_T \lesssim \|z_0^+\|_{B^s_{p,q}} + \|z_0^-\|_{B^s_{p,q}}.$$

In particular, we have

$$\|z^+(t)\|_{B^s_{p,q}} + \|z^-(t)\|_{B^s_{p,q}} \lesssim 2^{\Phi_\lambda(t)}(\|z_0^+\|_{B^s_{p,q}} + \|z_0^-\|_{B^s_{p,q}}) \quad \forall\, t \in [0,T), \tag{4.5}$$

where

$$\Phi_\lambda(t) = \lambda \int_0^t (\|\nabla z^+\|_\infty + \|\nabla z^-\|_\infty)(\tau) d\tau.$$

Applying Proposition 2.1 with $f(t) = \nabla z^+$ and $f(t) = \nabla z^-$, respectively, and collecting the resulting inequalities, we have

$$\begin{aligned}\int_0^t (\|\nabla z^+\|_\infty + \|\nabla z^-\|_\infty)(\tau) d\tau \lesssim 1 + \sup_j \int_0^t (\|\Delta_j \nabla z^+\|_\infty + \|\Delta_j \nabla z^-\|_\infty) d\tau \\ \times \bigg(1 + \log^+(t\|z^+\|_{L^\infty_t(B^s_{p,q})} + t\|z^-\|_{L^\infty_t(B^s_{p,q})})\bigg).\end{aligned} \tag{4.6}$$

For the sake of convenience, we denote that

$$\zeta(T) \triangleq \sup_{[0,T)} \big\|z^+(t)\big\|_{B^s_{p,q}} + \sup_{[0,T)} \big\|z^-(t)\big\|_{B^s_{p,q}}.$$

Putting (4.6) into (4.5), then taking supremum over $[0,T)$ with respect to t, we have

$$\zeta(T) \lesssim 2^{\lambda\bigg(1 + \sup_j \int_0^T (\|\Delta_j \nabla z^+\|_\infty + \|\Delta_j \nabla z^-\|_\infty) d\tau (1 + \log^+(T\zeta(T)))\bigg)} \zeta(0).$$

We should point out that the above inequality still holds if the time interval $[0,T)$ is replaced by $[T-\varepsilon, T)$. Thanks to condition (4.3), we deduce that

$$\zeta(T) \lesssim 2^{\lambda \sup_j \int_{T-\varepsilon}^T (\|\Delta_j \nabla z^+\|_\infty + \|\Delta_j \nabla z^-\|_\infty) d\tau \log^+(\varepsilon\zeta(T))} \zeta(T-\varepsilon).$$

Setting $Z(T) \triangleq \log(e + \zeta(T))$, we finally have

$$Z(T) \le C\lambda \sup_j \int_{T-\varepsilon}^T (\|\Delta_j \nabla z^+\|_\infty + \|\Delta_j \nabla z^-\|_\infty) d\tau Z(T) + Z(T-\varepsilon). \tag{4.7}$$

If we choose $M = \frac{1}{2CC_0\lambda}$, the condition (4.3) ensures the term $\lambda \sup_j \int_{T-\varepsilon}^T (\|\Delta_j \nabla z^+\|_\infty + \|\Delta_j \nabla z^-\|_\infty) d\tau < \frac{1}{C\lambda}$ when $\varepsilon \to 0$, which implies that

$$Z(T) \lesssim Z(T-\varepsilon).$$

Hence we have the $B^s_{p,q}$ regularity for the solution at $t = T$ and the solution can be continued after $t = T$. This completes the proof of Theorem 1.2. ∎

Acknowledgments. Changxing Miao would like to thank Professor Chemin for giving him the paper [7] and his helpful suggestions. The authors are deeply grateful to the referees for their valuable advice.

REFERENCES

[1] J. Bergh and J. Löfstrom, *Interpolation Spaces, An Introduction*, Springer-Verlag, Berlin-New York, 1976.

[2] J. T. Beale, T. Kato, and A. J. Majda, *Remarks on the breakdown of smooth solutions for the* 3*-D Euler equations*, Comm. Math. Phys., 94 (1984), pp. 61–66.

[3] R. E. Caflisch, I. Klapper, and G. Steele, *Remarks on singularities, dimension and energy dissipation for ideal hydrodynamics and MHD*, Comm. Math. Phys., 184 (1997), pp. 443–455.

[4] M. Cannone and F. Planchon, *More Lyapunov functions for the Navier-Stokes equations*, in Navier-Stokes Equations: Theory and Numerical Methods, Lecture Notes in Pure and Appl. Math. 223, Marcel-Dekker, New York, 2002.

[5] J.-Y. Chemin, *Perfect Incompressibe Fluids*, in Oxford Lecture Series in Mathematics and its Applications 14, Oxford University Press, New York, 1998.

[6] J.-Y. Chemin and N. Lerner, *Flot de champs de vecteurs non lipschitziens et equations de Navier-Stokes*, J. Differential Equations, 121 (1995), pp. 314–328.

[7] J.-Y. Chemin and N. Masmoudi, *About lifespan of regular solutions of equations related to viscoelastic fluids*, SIAM J. Math. Anal., 33 (2001), pp. 84-112.

[8] R. Danchin, *Estimates in Besov spaces for transport and transport-diffusion equations with almost Lipschitz coefficients*, Rev. Mat. Iberoamericana, 21 (2005), pp. 863–888.

[9] W. M. Elsasser, *The hydromagnetic equations*, Phys. Rev., 79 (1950), p. 183.

[10] H. Kozono and Y. Taniuchi, *Limiting case of the Sobolev inequalities in BMO, with application to the Euler equations*, Comm. Math. Phys., 214 (2000), pp. 191–200.

[11] H. Kozono, T. Ogawa, and Y. Taniuchi, *The critical Sobolev inequalities in Besov spaces and regularity criterion to some semi-linear evolution equations*, Math. Z., 242 (2002), pp. 251–278.

[12] A. J. Majda, *Compressible fluid flow and systems of conservation laws in several space variables*, Appl. Math. Sci., 53, Springer-Verlag, New York, 1984.

[13] Y. Meyer, *Wavelets and operators*, Cambridge Stud. Adv. Math., Cambridge University Press, Cambridge, UK, 1992.

[14] F. Planchon, *An extension of the Beale-Kato-Majda criterion for the Euler equations*, Comm. Math. Phys., 232 (2003), pp. 319–326.

[15] H. Triebel, *Theory of Function Spaces*, Monograph in Mathematics 78, Birkhauser Verlag, Basel, 1983.

[16] M. Vishik, *Hydrodynamics in Besov spaces*, Arch. Ration. Mech. Anal., 145 (1998), pp. 197–214.

[17] Z-F. Zhang and X-F. Liu, *On the blow-up criterion of smooth solutions to the* 3*D ideal MHD equations*, Acta Math. Appl. Sin. Engl. Ser., E, 20 (2004), pp. 695–700.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1860–1885

THE EFFECT OF INHIBITOR ON THE PLASMID-BEARING AND PLASMID-FREE MODEL IN THE UNSTIRRED CHEMOSTAT*

JIANHUA WU†, HUA NIE†, AND GAIL S. K. WOLKOWICZ‡

Abstract. This paper deals with a chemostat model with an inhibitor in the context of competition between plasmid-bearing and plasmid-free organisms. First, sufficient conditions for coexistence of the steady-state are determined. Second, the effects of the inhibitor are considered. It turns out that the parameter μ, which represents the effect of the inhibitor, plays a very important role in deciding the number of the coexistence solutions. The results show that if μ is sufficiently large this model has at least two coexistence solutions provided that the maximal growth rate a of u lies in a certain range and has only one unique asymptotically stable coexistence solution when a belongs to another range. Finally, extensive simulations are done to complement the analytic results. The main tools used here include degree theory in cones, bifurcation theory, and perturbation technique.

Key words. chemostat, coexistence solution, perturbation theory, stability

AMS subject classifications. 35K55, 35K57,35J65, 92A17

DOI. 10.1137/050627514

1. Introduction.

The chemostat is a common model in microbial ecology. It is used as an ecological model of a simple lake, as a model of waste treatment, and as a model for commercial production of fermentation processes. It is important in ecology because the parameters are readily measurable and, thus, the mathematical results are readily testable. For a general discussion of competitive systems see [29], while a detailed mathematical description of competition in the chemostat can be found in [30].

Our study focuses on a chemostat model in the context of competition between plasmid-bearing and plasmid-free organisms. This issue has recently received considerable attention. The theoretical literature on this model includes Ryder and DiBiaso [25], Stephanopoulos and Lapidus [28], Hsu, Waltman, and Wolkowicz [17], Lu and Hadeler [22], Levin [20], Luo and Hsu [18], and Macken, Levin, and Waldstätter [23]. In industry, genetically altered organisms are frequently used to manufacture a desired product, for instance, a pharmaceutical. The alteration is accomplished by introducing a piece of DNA into the cell in the form of a plasmid. The burden imposed on the cell by the task of production can result in the genetically altered (the plasmid-bearing) organism being a less able competitor than the plasmid-free organism. Unfortunately, the plasmid can be lost in the reproductive process. Thus, it is possible for the plasmid-free organism to take over the culture. To avoid "capture" of the process by the plasmid-free organism, the obvious choice is to alter the medium in such a way as to favor the plasmid-bearing organism. An example of this would be to introduce an antibiotic into the feed bottle. See [10, 15, 16] for a detailed biological and chemical background. Models in this direction have been studied in Lenski and Hattingh [21], Hsu and Waltman [13, 15, 16], Hsu, Luo, and Waltman [12], Nie and Wu [24], and the references therein.

*Received by the editors March 24, 2005; accepted for publication (in revised form) October 20, 2006; published electronically March 15, 2007. The work is supported by the Natural Science Foundation of China (10571115 and 10071048), the Excellent Young Teachers Program of the Ministry of Education of China, and the Innovation Foundation of Shaanxi Normal University.

http://www.siam.org/journals/sima/38-6/62751.html

†College of Mathematics and Information Science, Shaanxi Normal University, Xi'an, Shaanxi 710062, People's Republic of China (wjhua@snnu.edu.cn).

‡Department of Mathematics and Statistics, McMaster University, Hamilton L8S 4K1, ON, Canada (wolkowic@mcmaster.ca).

This paper is concerned with the competition model between plasmid-bearing and plasmid-free organisms in the unstirred chemostat in the presence of an inhibitor. Here the plasmid-bearing organism devotes a partition of its resource to produce an inhibitor, which diminishes the growth rate of the plasmid-free organism but does not reduce that of the plasmid-bearing organism. The pioneering work on this model is that of Hsu and Waltman in [15]. They proposed an ODE model (see [15]) based on the study of Chao and Levin [1] and Levin [20]. Moreover, they obtained some results on the global asymptotic behavior. In our current paper, we allow a heterogeneous environment and so we remove the well-stirred hypothesis and consider the corresponding PDE system. Let $s(x,t)$ be the nutrient concentration at time t; let $u(x,t)$ and $v(x,t)$ be the concentrations of the plasmid-bearing and plasmid-free organisms in the culture vessel, respectively, and let $p(x,t)$ be the concentration of the inhibitor. Then using similar arguments as in [6, 14, 34, 32, 24] the model in the unstirred case takes the form

$$\begin{aligned}
&s_t = ds_{xx} - \tfrac{1}{r}auf_1(s) - \tfrac{1}{r}bvf_2(s)e^{-\mu p}, && x\in(0,1), t>0,\\
&u_t = du_{xx} + a(1-q-k)uf_1(s), && x\in(0,1), t>0,\\
&v_t = dv_{xx} + bvf_2(s)e^{-\mu p} + aquf_1(s), && x\in(0,1), t>0,\\
&p_t = dp_{xx} + akuf_1(s), && x\in(0,1), t>0
\end{aligned}$$

with boundary conditions and initial conditions

$$\begin{aligned}
&s_x(0,t) = -s^0, s_x(1,t) + \gamma s(1,t) = 0, && t>0,\\
&u_x(0,t) = u_x(1,t) + \gamma u(1,t) = 0, && t>0,\\
&v_x(0,t) = v_x(1,t) + \gamma v(1,t) = 0, && t>0,\\
&p_x(0,t) = p_x(1,t) + \gamma p(1,t) = 0, && t>0\\
&s(x,0) = s_0(x) \geq 0, && p(x,0) = p_0(x) \geq 0, \not\equiv 0,\\
&u(x,0) = u_0(x) \geq 0, \not\equiv 0, && v(x,0) = v_0(x) \geq 0, \not\equiv 0,
\end{aligned}$$

where $s^0 > 0$ is the input concentration of the nutrient, which is assumed to be constant; d is the diffusion rate of the chemostat; r is the growth yield constant and a, b are the maximal growth rates of the plasmid-bearing and plasmid-free organisms (without an inhibitor), respectively. The response functions are denoted by $f_i(s) = s/(k_i + s), i = 1, 2$, where k_i are the Michaelis–Menten constants. The term $e^{-\mu p}$ used by Lenski and Hattingh in [21] represents the degree of inhibition of p on the growth rate of v, where $\mu > 0$ is a constant and represents the effect of the inhibitor on v. The constant q is the fraction of plasmid lost, and k is the fraction of consumption devoted to the production of the inhibitor. Hence, $0 < q, k < 1$, and $1 - q - k > 0$. γ is a positive constant. In this model, the corresponding yield constants are assumed to be equal, just as in [17, 15, 20].

For the sake of convenience, by nondimensionalizing the parameters, which are indicated below with bars, $\bar{s} = s/s^0, \bar{u} = u/rs^0, \bar{v} = v/rs^0, \bar{p} = p/rs^0, \bar{k}_i = k_i/s^0, \bar{\mu} = rs^0\mu, f_i(\bar{s}) = f_i(s^0\bar{s})$, we can rewrite this model in the form

$$\begin{aligned}
&s_t = ds_{xx} - auf_1(s) - bvf_2(s)e^{-\mu p}, && x\in(0,1), t>0,\\
&u_t = du_{xx} + a(1-q-k)uf_1(s), && x\in(0,1), t>0,\\
&v_t = dv_{xx} + bvf_2(s)e^{-\mu p} + aquf_1(s), && x\in(0,1), t>0,\\
&p_t = dp_{xx} + akuf_1(s), && x\in(0,1), t>0,\\
&s_x(0,t) = -1, s_x(1,t) + \gamma s(1,t) = 0, && t>0,\\
&u_x(0,t) = u_x(1,t) + \gamma u(1,t) = 0, && t>0,\\
&v_x(0,t) = v_x(1,t) + \gamma v(1,t) = 0, && t>0,\\
&p_x(0,t) = p_x(1,t) + \gamma p(1,t) = 0, && t>0,\\
&s(x,0) = s_0(x) \geq 0, && p(x,0) = p_0(x) \geq 0, \not\equiv 0,\\
&u(x,0) = u_0(x) \geq 0, \not\equiv 0, && v(x,0) = v_0(x) \geq 0, \not\equiv 0.
\end{aligned} \tag{1.1}$$

For simplicity, we drop the bars over the nondimensional quantities.

Introduce the new variables $\Phi(x,t) = s + u + v + p$ and $\Psi(x,t) = p - cu$ into (1.1), where $c = k/(1-q-k)$. Then one can argue in exactly the same way as in [24, 33, 34, 36] to conclude that the limiting system of (1.1) may be written as

$$\text{(PP)}\quad \begin{aligned} &u_t = du_{xx} + a(1-q-k)uf_1(z-(1+c)u-v), && x\in(0,1), t>0,\\ &v_t = dv_{xx} + bvf_2(z-(1+c)u-v)e^{-\mu cu}\\ &\qquad +aquf_1(z-(1+c)u-v), && x\in(0,1), t>0,\\ &u_x(0,t) = u_x(1,t) + \gamma u(1,t) = 0, && t>0,\\ &v_x(0,t) = v_x(1,t) + \gamma v(1,t) = 0, && t>0,\\ &u(x,0) = u_0(x) \geq 0, \not\equiv 0,\ v(x,0) = v_0(x) \geq 0, \not\equiv 0, && x\in[0,1], \end{aligned}$$

where $z(x) = (1+\gamma)/\gamma - x, (1+c)u_0(x) + v_0(x) \leq z(x), \not\equiv z(x)$.

The purpose of the present paper is to investigate nonnegative steady-state solutions of system (1.1) and the effect of the inhibitor on coexistence states of this system. Thus we will concentrate on the simplified elliptic system:

$$\text{(EP)}\quad \begin{aligned} &du'' + a(1-q-k)uf_1(z-(1+c)u-v) = 0, && x\in(0,1),\\ &dv'' + bvf_2(z-(1+c)u-v)e^{-\mu cu} + aquf_1(z-(1+c)u-v) = 0,\\ &u'(0) = u'(1) + \gamma u(1) = 0, \qquad v'(0) = v'(1) + \gamma v(1) = 0, \end{aligned}$$

which is obtained from the steady-state system of (1.1) by introducing the variables $\Phi(x) = s + u + v + p$ and $\Psi(x) = p - cu$. Since the proof is standard, we omit it here. Interested readers can refer to [14, 24, 32, 33, 34] for details.

We are mainly interested in coexistence states of (EP), that is, the positive solutions of (EP). Hence, we redefine the response functions as follows:

$$\bar{f}_i(s) = \begin{cases} f_i(s), & s \geq 0,\\ \tan^{-1}(2s/k_i + 1) - \pi/4, & s < 0. \end{cases}$$

It is easily seen that $\bar{f}_i \in C^1(-\infty, +\infty)$. We will denote $\bar{f}_i(s)$ by $f_i(s)$ for the sake of simplicity.

This work is motivated by numerical simulations that seem to indicate that, when the parameters sit in a certain range, there exists a coexistence solution of (EP). More interestingly, it is possible that (EP) has exactly two coexistence solutions if v is a better competitor than u and the parameter μ is suitably large. From the biological standpoint, the numerical results mean that the inhibitor plays an important role in determining the number of coexistence solutions of (EP). As mentioned before, the main purpose of this paper is to determine when the numerical results hold and confirm them rigorously.

Turning now to a description of the main results, we start by introducing some notation and recalling some well-known facts. Let λ_1, σ_1 be, respectively, the principal eigenvalues of the problems

$$\begin{aligned} &d\varphi_1'' + \lambda_1 f_1(z)\varphi_1 = 0 \ \text{ in } \ (0,1), \quad \varphi_1'(0) = \varphi_1'(1) + \gamma\varphi_1(1) = 0;\\ &d\psi_1'' + \sigma_1 f_2(z)\psi_1 = 0 \ \text{ in } \ (0,1), \quad \psi_1'(0) = \psi_1'(1) + \gamma\psi_1(1) = 0, \end{aligned}$$

with the corresponding positive eigenfunctions uniquely determined by the normalization $\max_{[0,1]}\varphi_1 = \max_{[0,1]}\psi_1 = 1$. It is well known (see [14, 33]) that, if $a \leq \lambda_1/(1-q-k)$, the boundary value problem

$$\text{(1.2)}\quad du'' + a(1-q-k)uf_1(z-u) = 0, \qquad x\in(0,1), \qquad u'(0) = u'(1) + \gamma u(1) = 0$$

has zero as its unique nonnegative solution, and if $a > \lambda_1/(1-q-k)$, then it has a unique positive solution, which is denoted by ϑ and satisfies the following properties.

(A) $0 < \vartheta < z$.

(B) ϑ is continuously differentiable for $a \in (\lambda_1/(1-q-k), +\infty)$ and is pointwise increasing when a increases.

(C) $\lim_{a\to\lambda_1/(1-q-k)} \vartheta = 0$ uniformly for $x \in (0,1)$, and $\lim_{a\to\infty} \vartheta = z(x)$ for almost every $x \in (0,1)$.

(D) Let $L_a = d\frac{d^2}{dx^2} + a(1-q-k)(f_1(z-\vartheta) - \vartheta f_1'(z-\vartheta))$ be the linear operator of the above equation at ϑ. Then L_a is a differential operator in $C_B^2([0,1]) = \{u \in C^2([0,1]) : u'(0) = u'(1) + \gamma u(1) = 0\}$, and all eigenvalues of L_a are strictly negative.

Remark 1. For the other steady-state one-species problem

$$dv'' + bvf_2(z-v) = 0, \qquad x \in (0,1), \qquad v'(0) = v'(1) + \gamma v(1) = 0,$$

we have the same outcomes. Since we will need this later, we denote the unique positive solution by θ and the linear operator by $L_b = d\frac{d^2}{dx^2} + b(f_2(z-\theta) - \theta f_2'(z-\theta))$.

Next, we introduce $\hat{\lambda}_1$ as the principal eigenvalue of

$$d\hat{\varphi}_1'' + \hat{\lambda}_1 f_1(z-\theta)\hat{\varphi}_1 = 0 \ \text{ in } \ (0,1), \ \ \hat{\varphi}_1'(0) = \hat{\varphi}_1'(1) + \gamma\hat{\varphi}_1(1) = 0,$$

with the corresponding eigenfunction $\hat{\varphi}_1$ normalized by $\max_{[0,1]} \hat{\varphi}_1 = 1$.

Now we are ready to state the main results of this paper, which give analytic confirmation of some of the numerical results.

THEOREM 1.1. (EP) *has a coexistence solution if either* (i) $a > \lambda_1/(1-q-k)$, $b < \sigma_1$ *or* (ii) $a > \hat{\lambda}_1/(1-q-k)$, $b > \sigma_1$.

THEOREM 1.2. *Suppose* $b > \sigma_1$. *Then for any* $\epsilon > 0$ *small and any* $A \geq \frac{\hat{\lambda}_1}{1-q-k}$, *there exists* $M = M(\epsilon, A)$ *large such that for* $\mu \geq M$,

(i) *if* $a \in [\lambda_1/(1-q-k) + \epsilon, \hat{\lambda}_1/(1-q-k))$, *there exist at least two coexistence solutions of* (EP);

(ii) *if* $a \in [\hat{\lambda}_1/(1-q-k), A]$, *there exists a unique coexistence solution of* (EP), *and it is asymptotically stable.*

THEOREM 1.3. *Suppose* $b > \sigma_1$. *Then there exist* $\epsilon_0 > 0$ *small and* $M_0 > 0$ *large, both independent of* a, *such that if* $a \in [\hat{\lambda}_1/(1-q-k) - \epsilon_0, \hat{\lambda}_1/(1-q-k))$ *and* $\mu \geq M_0$, *then* (EP) *has exactly two coexistence solutions, one asymptotically stable and the other unstable.*

The main tools in proving Theorems 1.1–1.3 include degree theory and bifurcation theory. A crucial point of the proof for Theorems 1.2 and 1.3 is to make use of the limiting equations of (EP) which are obtained by letting $\mu \to \infty$ formally in (EP). It turns out that one of the limiting problems can be understood fully. For the other limiting problem, we can also attain some properties. Finally, perturbation theory leads to the main results of this paper.

The contents of the present paper are as follows: In section 2, some preliminary results are given which are needed in the later sections. In section 3, we consider the general case and prove Theorem 1.1. For the case μ large, the uniqueness and non-uniqueness of the coexistence solutions to (EP) are obtained in section 4. The stability is also obtained for some cases. Finally, in section 5, some numerical simulations are given complementing the analytical results.

2. Preliminaries. We begin by providing the following well-known lemmas as preliminaries without proofs. They are useful for obtaining the results in this paper.

LEMMA 2.1 (see [9, 19]). *Suppose* $q(x) \in C(\overline{\Omega})$ *and* $q(x) > 0$ *on* $\overline{\Omega}$ *in the eigenvalue problem*

$$\triangle\phi + \lambda q(x)\phi = 0, \qquad x \in \Omega, \qquad \frac{\partial\phi}{\partial n} + \gamma(x)\phi = 0, \qquad x \in \partial\Omega, \tag{2.1}$$

where $\gamma(x) \in C(\partial\Omega)$ *and* $\gamma(x) \geq 0$. *Then all eigenvalues of* (2.1) *can be listed in order*

$$0 < \lambda_1(q(x)) < \lambda_2(q(x)) \leq \cdots \to \infty$$

with the corresponding eigenfunctions $\phi_1, \phi_2, \ldots$, *where* $\phi_1 > 0$ *on* $\overline{\Omega}$, *and the principal eigenvalue*

$$\lambda_1(q) = \inf_{\phi} \frac{\int_\Omega |\nabla\phi|^2 dx + \int_{\partial\Omega} \gamma(x)\phi^2 ds}{\int_\Omega q(x)\phi^2 dx}$$

is simple. Moreover, the comparison principle holds: $\lambda_j(q_1) \leq \lambda_j(q_2)$ *for* $j \geq 1$ *if* $q_1 \geq q_2$ *on* $\overline{\Omega}$, *and strict inequality holds if* $q_1(x) \not\equiv q_2(x)$.

LEMMA 2.2 (see [27]). *Suppose* $q \in C(\overline{\Omega})$, $\gamma(x) \in C(\partial\Omega)$, *and* $\gamma(x) \geq 0$. *Let* $\sigma_1(q)$ *be the first eigenvalue of the problem* $-\triangle\omega + q\omega = \lambda\omega, x \in \Omega$, $\frac{\partial\omega}{\partial n} + \gamma(x)\omega = 0, x \in \partial\Omega$. *Then* $\sigma_1(q)$ *depends continuously on* q, *and* $q_1 \leq q_2, q_1 \not\equiv q_2$ *imply* $\sigma_1(q_1) < \sigma_1(q_2)$.

LEMMA 2.3 (see [31]). *Let* $q(x) \in C(\overline{\Omega})$ *and* $q(x) + p > 0$ *on* $\overline{\Omega}$ *with* $p > 0$, *and let* η_1 *be the first eigenvalue of the eigenvalue problem*

$$-\triangle\varphi - q(x)\varphi = \eta\varphi, \qquad x \in \Omega, \qquad \frac{\partial\varphi}{\partial n} + \gamma(x)\varphi = 0, \qquad x \in \partial\Omega,$$

where $\gamma(x) \in C(\partial\Omega)$ *and* $\gamma(x) \geq 0$. *If* $\eta_1 > 0$ *(or* $\eta_1 < 0$*), then the eigenvalue problem*

$$-\triangle\varphi + p\varphi = t(q(x) + p)\varphi, \qquad x \in \Omega, \qquad \frac{\partial\varphi}{\partial n} + \gamma(x)\varphi = 0, \qquad x \in \partial\Omega$$

has no eigenvalue less than or equal to 1 *(or has eigenvalues less than* 1*).*

Now, we introduce some more notation that will be used throughout this paper. Let X be a real Banach space, and let $W \subset X$ be a closed convex set. W is called a wedge provided that $\alpha W \subset W$ for all $\alpha \geq 0$. A wedge W is said to be a cone if $W \cap \{-W\} = 0$. Let $y \in W$, and define a wedge

$$W_y := \mathrm{cl}\{x \in X | y + \nu x \in W \ \text{ for some } \ \nu > 0\},$$

where "cl" means the closure of the set. Let S_y be the maximal linear subspace of X contained in W_y. Assume that T is a compact and Fréchet differentiable operator on X such that $y \in W$ is a fixed point of T and $T(W) \subseteq W$. Then the Fréchet derivative $T'(y)$ of T at y leaves W_y and S_y invariant (see [4, 26]). If there exists a closed linear subspace X_y of X such that $X = S_y \oplus X_y$ and W_y is generating, then the index of T at y can be found by analyzing certain eigenvalue problems in X_y and S_y as follows. Let $Q : X \to X_y$ be the projection operator of X_y along S_y. In view of Theorems 2.1 and 2.2 of [26], $\mathrm{index}_W(T, y)$ exists if the Fréchet derivative $T'(y)$ of T at y has no nonzero fixed point in W_y. Furthermore,

(1) $\mathrm{index}_W(T, y) = 0$ if $Q \circ T'(y)$ has an eigenvalue $\lambda > 1$;

(2) $\mathrm{index}_W(T, y) = \mathrm{index}_{S_y}(T'(y), 0)$ if $Q \circ T'(y)$ has no such eigenvalues.

Here $\mathrm{index}_{S_y}(T'(y), 0)$ is the index of the linear operator $T'(y)$ at 0 in the space S_y.

Next, we derive some a priori estimates for positive solutions of (EP). For an accurate estimate for positive solutions of (EP), we first consider the boundary value problem

$$\begin{aligned} &dv'' + bvf_2(z-v) + \tfrac{aq\vartheta}{1+c} f_1(z-v) = 0,\ x \in (0,1), \\ &v'(0) = v'(1) + \gamma v(1) = 0. \end{aligned} \tag{2.2}$$

LEMMA 2.4. *There exists a unique positive solution of* (2.2), *denoted by* $\bar{v}(x)$, *which satisfies* $0 < \bar{v}(x) < z$. *In particular,* $\theta < \bar{v}(x) < z$ *if* $b > \sigma_1$.

Proof. First, we claim that if $v(x)$ is a positive solution of (2.2), then $0 < v(x) < z$ and that, in addition, if $b > \sigma_1$, then $\theta < v(x) < z$. Indeed, let $\omega = z - v$. Then

$$d\omega'' - bvf_2(\omega) - \frac{aq\vartheta}{1+c} f_1(\omega) = 0, \qquad x \in (0,1), \qquad \omega'(0) = -1, \qquad \omega'(1) + \gamma\omega(1) = 0.$$

Suppose $\inf_{x\in[0,1]} \omega(x) = \omega(x_0) < 0$. Then $x_0 \notin (0,1)$. Otherwise, $\omega''(x_0) \geq 0$. By the previous equation for ω, we have $d\omega''(x_0) = bv(x_0)f_2(\omega(x_0)) + \frac{aq\vartheta(x_0)}{1+c} f_1(\omega(x_0)) < 0$, a contradiction. If $x_0 = 0$, then $\omega'(x_0) \geq 0$, contradicting the boundary condition $\omega'(0) = -1$. Similarly, we can see that $x_0 = 1$ is also impossible. Hence, one must have $\omega \geq 0, \not\equiv 0$ on [0, 1].

Assume $\omega(x_0) = 0$ for some point $x_0 \in [0,1]$. Then $x_0 = 0$ or 1 by the strong maximum principle. On the other hand, from the Hopf boundary lemma, it is easy to see that both $x = 0$ and 1 are impossible, which implies $\omega > 0$ on [0, 1]. That is, $v < z$ on [0, 1]. Moreover, since

$$dv'' + bvf_2(z-v) + \frac{aq\vartheta}{1+c} f_1(z-v) > dv'' + bvf_2(z-v),$$

it is easy to see that $v > \theta$ if $b > \sigma_1$. Hence, our assertion holds.

On the other hand, for sufficiently small $\delta > 0$, $\delta\varphi_1, z$ are the sub- and supersolutions of (2.2), respectively. It follows from the existence-comparison theorem for elliptic systems that the minimal and maximal solutions v_1, v_2 to (2.2) exist and satisfy the relation $\delta\varphi_1 < v_1 \leq v_2 < z$. Next, we show that $v_1 \equiv v_2$, to obtain the uniqueness. Since v_1, v_2 are the solutions of (2.2),

$$\begin{aligned} dv_1'' + bv_1 f_2(z-v_1) + \tfrac{aq\vartheta}{1+c} f_1(z-v_1) = 0, \\ dv_2'' + bv_2 f_2(z-v_2) + \tfrac{aq\vartheta}{1+c} f_1(z-v_2) = 0. \end{aligned}$$

Multiplying the first equation by v_2 and the second equation by v_1 and considering the integral $I = \int_0^1 d(v_1'' v_2 - v_2'' v_1)dx$, we have

$$\int_0^1 bv_1v_2(f_2(z-v_1) - f_2(z-v_2))dx + \frac{aq}{1+c}\int_0^1 \vartheta[v_2 f_1(z-v_1) - v_1 f_1(z-v_2)]dx = 0.$$

By the monotonity of $f_i (i = 1, 2)$ and since $v_1 \leq v_2$, we have $v_1 \equiv v_2$. □

The next lemma gives a priori estimates for positive solutions of (EP).

LEMMA 2.5. *Assume* (u, v) *is a nonnegative solution of* (EP) *with* $u \not\equiv 0$ *and* $v \not\equiv 0$. *Then*

1) $0 < u < \frac{\vartheta}{1+c} < \frac{z}{1+c}, 0 < v \leq \bar{v} < z$ *on* [0, 1], *where* $\bar{v}$ *defined by Lemma* 2.4;
2) $(1+c)u + v < z$ *on* [0, 1];
3) $a > \frac{\lambda_1}{1-q-k}$.

Proof. Clearly, $u > 0$ on $[0, 1]$ by the strong maximum principle and Hopf boundary lemma. Since $0 = du'' + a(1-q-k)uf_1(z-(1+c)u-v) \le du'' + a(1-q-k)uf_1(z-(1+c)u)$, it is easy to check that $u \le \frac{\vartheta}{1+c}$ and $a > \frac{\lambda_1}{1-q-k}$. Moreover, one can find that $u < \frac{\vartheta}{1+c}$ because $v \not\equiv 0$.

For v, we have

$$\begin{aligned} 0 &= dv'' + bvf_2(z-(1+c)u-v)e^{-\mu cu} + aquf_1(z-(1+c)u-v) \\ &\le dv'' + bvf_2(z-v) + \tfrac{aq\vartheta}{1+c}f_1(z-v). \end{aligned}$$

By Lemma 2.4 and the strong maximum principle, it follows that $0 < v \le \bar{v} < z$. It remains to show that $(1+c)u + v < z$ on $[0, 1]$. This proof is similar to the proof of Lemma 4.2 in [33] and so is omitted here. ∎

3. Existence of coexistence solutions. The goal of this section is to discuss the existence of coexistence solutions of (EP) in the general case and to establish Theorem 1.1.

In order to use the functional analytic framework of degree theory we introduce the spaces

$$\begin{aligned} &X = C([0,1]) \times C([0,1]), \\ &D = \{(u,v) \in X | u \le \tfrac{\vartheta}{1+c}, v \le \max_{[0,1]} \bar{v} + 1\}, \\ &W = \{(u,v) \in X | u \ge 0, v \ge 0 \ \text{ for } \ x \in [0,1]\}, \\ &D' = (\text{int}D) \cap W. \end{aligned}$$

Then W is a cone of X and D' is a bounded open set in W.

We consider the system

$$\begin{aligned} &du'' + \tau a(1-q-k)uf_1(z-(1+c)u-v) = 0, \\ &dv'' + \tau bvf_2(z-(1+c)u-v)e^{-\mu cu} + \tau aquf_1(z-(1+c)u-v) = 0, \\ &u'(0) = u'(1) + \gamma u(1) = 0, \quad v'(0) = v'(1) + \gamma v(1) = 0, \end{aligned} \tag{3.1}$$

where $\tau \in [0,1]$. Assume (u_τ, v_τ) is a nonnegative solution of (3.1). Then it is not hard to show that $u_\tau < \frac{\vartheta}{1+c}, v_\tau \le \bar{v}$ for all $\tau \in [0,1]$.

Since $f_1(z-(1+c)u-v) \ge f_1(z-(1+c)u) - K_1 v$ for some positive constant $K_1 > 0$, we can define $\mathcal{A}_\tau : X \to X, \tau \in [0,1]$ as

$$\begin{aligned} \mathcal{A}_\tau(u,v) = \ & (-d\tfrac{d^2}{dx^2} + M)^{-1}(\tau a(1-q-k)ug_1(u,v) + Mu, \\ & \tau bvg_2(u,v) + \tau aqug_1(u,v) + Mv) \end{aligned}$$

where $(-d\frac{d^2}{dx^2} + M)^{-1}$ is the inverse operator of $-d\frac{d^2}{dx^2} + M$ subject to the boundary conditions $u'(0) = u'(1) + \gamma u(1) = 0$, $g_1(u,v) = f_1(z-(1+c)u-v)$, $g_2(u,v) = f_2(z-(1+c)u-v)e^{-\mu cu}$, and M is large enough such that $M + \tau a(1-q-k)g_1(u,v) > 0$ and $M + \tau bg_2(u,v) - \tau aquK_1 > 0$ for all $(u,v) \in D'$ and $\tau \in [0,1]$. Clearly, $\mathcal{A}_\tau$ is compact. Let $\mathcal{A} = \mathcal{A}_1$. Then $\mathcal{A} : D' \to W$ is continuously differentiable. It follows from Lemma 2.5 that (EP) has nonnegative solutions if and only if the operator $\mathcal{A}$ has a fixed point in D'. Moreover, $\mathcal{A}_\tau$ has no fixed point on $\partial D'$. By the homotopic invariance of the degree, we obtain

$$\text{index}(\mathcal{A}, D', W) = \text{index}(\mathcal{A}_\tau, D', W) = \text{index}(\mathcal{A}_0, D', W) = \text{index}_W(\mathcal{A}_0, (0,0)).$$

By some standard calculations, we can obtain $\text{index}_W(\mathcal{A}_0, (0,0)) = 1$. Hence, we have the following.

LEMMA 3.1. $\text{index}(\mathcal{A}, D', W) = 1$.

LEMMA 3.2. (i) *Suppose* $a \neq \frac{\lambda_1}{1-q-k}$ *and* $b \neq \sigma_1$. *Then* $\text{index}_W(\mathcal{A}, (0,0)) = 1$ *if* $a < \frac{\lambda_1}{1-q-k}$ *and* $b < \sigma_1$, *and* $\text{index}_W(\mathcal{A}, (0,0)) = 0$ *if* $a > \frac{\lambda_1}{1-q-k}$ *or* $b > \sigma_1$.

(ii) $\text{index}_W(\mathcal{A}, (0,\theta)) = 1$ *if* $a < \frac{\hat{\lambda}_1}{1-q-k}$, *and* $\text{index}_W(\mathcal{A}, (0,\theta)) = 0$ *if* $a > \frac{\hat{\lambda}_1}{1-q-k}$.

Since the proof of this Lemma is very lengthy and quite standard, we include the proof in Appendix A. Now, we turn to prove Theorem 1.1.

Proof of Theorem 1.1. (i) If $a > \lambda_1/(1-q-k)$ and $b < \sigma_1$, then (EP) has no semitrivial nonnegative solution. In view of Lemmas 3.1 and 3.2, $\text{index}_W(\mathcal{A}, D') = 1$ and $\text{index}_W(\mathcal{A}, (0,0)) = 0$, which implies that $\mathcal{A}$ must have a positive fixed point in D'. That is, (EP) has a positive solution in D'.

(ii) If $a > \hat{\lambda}_1/(1-q-k), b > \sigma_1$, then (EP) has a semitrivial nonnegative solution $(0,\theta)$. Suppose $\mathcal{A}$ has no positive fixed point in D'. Then by Lemma 3.1 and the additivity of index,

$$\text{index}_W(\mathcal{A}, (0,0)) + \text{index}_W(\mathcal{A}, (0,\theta)) = \text{index}_W(\mathcal{A}, D') = 1.$$

However, by Lemma 3.2, $\text{index}_W(\mathcal{A}, (0,0)) = 0$, and $\text{index}_W(\mathcal{A}, (0,\theta)) = 0$ in this case, a contradiction. Hence there must exist a positive solution of (EP) in D'. This completes the proof. □

4. The effect of inhibitor. The purpose of this section is to examine the effect of the inhibitor on the multiple coexistence states. In view of the model, the effect of the inhibitor increases as the parameter μ increases. Motivated by the numerical simulations, we consider only the case of $b > \sigma_1$ and μ large enough. Using a perturbation technique, we show that the system has two positive solutions if μ is sufficiently large and the other parameters sit in some suitable range.

First of all, we observe that if a is bounded away from $\lambda_1/(1-q-k)$ and μ is large, positive solutions to (EP) are of two types. More precisely, let (u,v) be any positive solution of (EP); then either (u,v) is close to a positive solution of the problem

$$\begin{aligned} &du'' + a(1-q-k)uf_1(z-(1+c)u-v) = 0, \qquad x \in (0,1),\\ &dv'' + aquf_1(z-(1+c)u-v) = 0, \qquad x \in (0,1),\\ &u'(0) = u'(1) + \gamma u(1) = 0, \qquad v'(0) = v'(1) + \gamma v(1) = 0, \end{aligned} \tag{4.1}$$

or $(\mu u, v)$ is close to a positive solution of the problem

$$\begin{aligned} &d\omega'' + a(1-q-k)\omega f_1(z-v) = 0, \qquad x \in (0,1),\\ &dv'' + bvf_2(z-v)e^{-c\omega} = 0, \qquad x \in (0,1),\\ &\omega'(0) = \omega'(1) + \gamma\omega(1) = 0, \qquad v'(0) = v'(1) + \gamma v(1) = 0. \end{aligned} \tag{4.2}$$

Since the above two equations play an important role in determining the coexistence solutions of (EP), we first study positive solutions of (4.1) and (4.2).

LEMMA 4.1. *Assume* $a > \lambda_1/(1-q-k)$. *Then there exists a unique positive solution* $((1-q-k)\vartheta, q\vartheta)$ *of* (4.1), *and it is linearly asymptotically stable.*

Proof. Suppose that $(u,v) > 0$ solves (4.1). Let $\omega = qu - (1-q-k)v$. Then we have

$$d\omega'' = 0, \qquad \omega'(0) = \omega'(1) + \gamma\omega(1) = 0,$$

which implies $\omega \equiv 0$. That is, $v = \frac{q}{1-q-k}u$. Substituting $v = \frac{q}{1-q-k}u$ into the first

equation of (4.1), we obtain that

$$\begin{aligned} &du'' + a(1-q-k)uf_1\left(z - \frac{u}{1-q-k}\right) = 0, \qquad x \in (0,1),\\ &u'(0) = u'(1) + \gamma u(1) = 0. \end{aligned}$$

Then $u = (1-q-k)\vartheta$ due to $a > \lambda_1/(1-q-k)$, and $v = q\vartheta$. That is, (4.1) has a unique positive solution $((1-q-k)\vartheta, q\vartheta)$. It remains to establish the stability. For this purpose, noting that $c = k/(1-q-k)$, we consider the linearized eigenvalue problem

$$\begin{aligned} &d\phi'' + a(1-q-k)[f_1(z-\vartheta) - (1-q)\vartheta f_1'(z-\vartheta)]\phi\\ &\quad -a(1-q-k)^2\vartheta f_1'(z-\vartheta)\psi = -\eta\phi,\\ &d\psi'' + aq[f_1(z-\vartheta) - (1-q)\vartheta f_1'(z-\vartheta)]\phi - aq(1-q-k)\vartheta f_1'(z-\vartheta)\psi = -\eta\psi,\\ &\phi'(0) = \phi'(1) + \gamma\phi(1) = 0, \qquad \psi'(0) = \psi'(1) + \gamma\psi(1) = 0. \end{aligned} \tag{4.3}$$

Let $\omega = q\phi - (1-q-k)\psi$. Then

$$d\omega'' = -\eta\omega, \qquad \omega'(0) = \omega'(1) + \gamma\omega(1) = 0.$$

If $\omega \not\equiv 0$, then $\eta > 0$. If $\omega \equiv 0$, then $\psi = \frac{q\phi}{1-q-k}$, which leads to

$$\begin{aligned} &d\phi'' + a(1-q-k)(f_1(z-\vartheta) - \vartheta f_1'(z-\vartheta))\phi = -\eta\phi,\\ &\phi'(0) = \phi'(1) + \gamma\phi(1) = 0. \end{aligned}$$

From Lemma 2.2, $\sigma_1(a(1-q-k)(f_1(z-\vartheta) - \vartheta f_1'(z-\vartheta))) < \sigma_1(a(1-q-k)f_1(z-\vartheta)) = 0$. Hence, we can claim that $\eta > 0$. Therefore, (4.3) has no eigenvalue η with $Re\eta \le 0$ and so the stability follows. □

LEMMA 4.2. *Suppose $b > \sigma_1$ fixed. Then* (4.2) *has a positive solution if and only if $\frac{\lambda_1}{1-q-k} < a < \frac{\hat\lambda_1}{1-q-k}$. Moreover, all positive solutions of* (4.2) *are unstable.*

Proof. Suppose (ω, v) is a positive solution of (4.2). Then $a(1-q-k) = \lambda_1(f_1(z-v)) > \lambda_1(f(z)) = \lambda_1$. On the other hand,

$$0 = dv'' + bvf_2(z-v)e^{-c\omega} < dv'' + bvf_2(z-v),$$

which means $v < \theta$. Thus, $a(1-q-k) = \lambda_1(f_1(z-v)) < \lambda_1(f_1(z-\theta)) = \hat\lambda_1$. Hence, if (4.2) has a positive solution, then $\frac{\lambda_1}{1-q-k} < a < \frac{\hat\lambda_1}{1-q-k}$.

Next, we show that (4.2) has a positive solution if $\lambda_1/(1-q-k) < a < \hat\lambda_1/(1-q-k)$. To this end, we first prove that for any given $A > \hat\lambda_1/(1-q-k)$, there exists a constant $C > 0$ such that $\|\omega\|_\infty < C$ for any nonnegative solution of (4.2) with $a \in (\lambda_1/(1-q-k), A]$. At first, one can find that (4.2) has only two nonnegative solutions $(0,0)$ and $(0,\theta)$ if $a \ge \hat\lambda_1/(1-q-k)$. It remains to show that any positive solution (ω, v) of (4.2) with $\frac{\lambda_1}{1-q-k} < a < \frac{\hat\lambda_1}{1-q-k}$ satisfies $\|\omega\|_\infty < C$. Suppose this is not true. Then we may assume that there exists $a_i \to a \in [\lambda_1/(1-q-k), \hat\lambda_1/(1-q-k)]$, (ω_i, v_i) positive solutions of (4.2) with $a = a_i$ and $\|\omega_i\|_\infty \to \infty$. Set $\tilde v_i = v_i/\|v_i\|_\infty, \tilde\omega_i = \omega_i/\|\omega_i\|_\infty$. Then

$$\begin{aligned} &d\tilde\omega_i'' + a_i(1-q-k)\tilde\omega_i f_1(z - \|v_i\|_\infty\tilde v_i) = 0,\\ &d\tilde v_i'' + b\tilde v_i f_2(z-v_i)e^{-c\|\omega_i\|_\infty\tilde\omega_i} = 0,\\ &\tilde\omega_i'(0) = \tilde\omega_i'(1) + \gamma\tilde\omega_i(1) = 0, \qquad \tilde v_i'(0) = \tilde v_i'(1) + \gamma\tilde v_i(1) = 0. \end{aligned}$$

By L^p estimates and the Sobolev embedding theorem, we may assume $\tilde{\omega}_i \to \tilde{\omega} \geq 0, \not\equiv 0$, $\tilde{v}_i \to \tilde{v} \geq 0, \not\equiv 0$ in $C^1([0,1])$, and $\tilde{\omega}$ satisfies

$$d\tilde{\omega}'' + a(1-q-k)\tilde{\omega} f_1(z - B\tilde{v}) = 0, \quad \tilde{\omega}'(0) = \tilde{\omega}'(1) + \gamma\tilde{\omega}(1) = 0,$$

where $B = \lim_{i\to\infty} \|v_i\|_\infty < \infty$. (In view of the equation for v_i and $0 < v_i < \theta$, this limit exists by passing to a subsequence.) Thus $\tilde{\omega} > 0$ on $[0, 1]$ by the strong maximum principle and Hopf boundary lemma. Hence $e^{-c\omega_i} = e^{-c\|\omega_i\|_\infty \tilde{\omega}_i} \to 0$ as $i \to \infty$, which implies $v_i \to 0$, and $\tilde{v}$ satisfies

$$d\tilde{v}'' = 0, \quad \tilde{v}'(0) = \tilde{v}'(1) + \gamma\tilde{v}(1) = 0.$$

Thus $\tilde{v} \equiv 0$. This is a contradiction to $\tilde{v} \not\equiv 0$ and $\|\tilde{v}\|_\infty = 1$.

Let $\tilde{D} = \{(\omega, v) \in W : \|\omega\|_\infty \leq C + 1, \|v\|_\infty \leq \sup_{[0,1]} z + 1\}$,

$$B_\tau(\omega, v) = \left(-d\frac{d^2}{dx^2} + M\right)^{-1} \left(\tau(1-q-k)\omega f_1(z-v) + M\omega,\ bvf_2(z-v)e^{-c\omega} + Mv\right),$$

where W defined in section 3 is the positive cone of X and M is sufficiently large such that $M + \tau(1-q-k)f_1(z-v) > 0$ and $M + bf_2(z-v)e^{-c\omega} > 0$ for all $(\omega, v) \in \tilde{D}$ and $\tau \in (\lambda_1/(1-q-k), A]$.

By virtue of our a priori estimates and the homotopic invariance property of the fixed point index, we obtain $\text{index}_W(B_\tau, \tilde{D}) \equiv \text{const}$ for $\tau > \lambda_1/(1-q-k)$. On the other hand, if $a > \hat{\lambda}_1/(1-q-k)$, then (4.2) has only two nonnegative solutions $(0,0)$ and $(0,\theta)$. Hence for $\tau \in (\hat{\lambda}_1/(1-q-k), A]$, $\text{index}_W(B_\tau, \tilde{D}) = \text{index}_W(B_\tau, (0,0)) + \text{index}_W(B_\tau, (0,\theta))$. Next, we calculate the index of the two nonnegative solutions.

Let $B'_\tau(0,0)$ be the Fréchet derivative of B_τ at $(0,0)$. Then

$$B'_\tau(0,0)(\omega, v) = \left(-d\frac{d^2}{dx^2} + M\right)^{-1} (\tau(1-q-k)\omega f_1(z) + M\omega,\ bvf_2(z) + Mv)$$

for each $(\omega, v) \in X$. Therefore, an eigenvector (ω, v) of $B'_\tau(0,0)$ satisfies

$$\begin{aligned}
&-d\omega'' + M\omega = \tfrac{1}{\lambda}(\tau(1-q-k)f_1(z) + M)\omega,\\
&-dv'' + Mv = \tfrac{1}{\lambda}(bf_2(z) + M)v,\\
&\omega'(0) = \omega'(1) + \gamma\omega(1) = 0, \qquad v'(0) = v'(1) + \gamma v(1) = 0.
\end{aligned}$$

Since $b > \sigma_1, \tau > \lambda_1/(1-q-k)$, it is easy to check that $I - B'_\tau(0,0)$ is invertible in $W_{(0,0)} = \{(\omega, v) \in X : \omega \geq 0, v \geq 0\}$. Moreover, a similar argument as in the proof of Lemma 3.2 (see Appendix A) shows that $B'_\tau(0,0)$ has eigenvalues larger than 1. It follows from Theorem 2.2 of [26] that $\text{index}_W(B_\tau, (0,0)) = 0$ for $\tau > \lambda_1/(1-q-k)$.

Let $B'_\tau(0,\theta)$ denote the Fréchet derivative of B_τ at $(0,\theta)$. Then $B'_\tau(0,\theta)(\omega, v) = (-d\frac{d^2}{dx^2} + M)^{-1}(\tau(1-q-k)\omega f_1(z-\theta) + M\omega, b(f_2(z-\theta) - \theta f'_2(z-\theta))v - bc\theta f_2(z-\theta)\omega + Mv)$ for each $(\omega, v) \in X$. In order to apply Theorem 2.2 of [26], we introduce the notation $y = (0,\theta), W_y = \{(\omega, v) \in X : \omega \geq 0\}, S_y = \{(0, v) : v \in C_B([0,1])\}$, and $X_y = \{(\omega, 0) \in X : \omega \in C_B([0,1])\}$. Then $X = S_y \oplus X_y$ with projection Q given by $(\omega, v) \to (\omega, 0)$.

Suppose $(\omega, v) \in W_y$ is a fixed point of $B'_\tau(0,\theta)$. Then (ω, v) satisfies

$$\begin{aligned}
&d\omega'' + \tau(1-q-k)\omega f_1(z-\theta) = 0,\\
&dv'' + b(f_2(z-\theta) - \theta f'_2(z-\theta))v - bc\theta f_2(z-\theta)\omega = 0,\\
&\omega'(0) = \omega'(1) + \gamma\omega(1) = 0, \qquad v'(0) = v'(1) + \gamma v(1) = 0.
\end{aligned}$$

It is easy to check that $I - B'_\tau(0,\theta)$ is invertible in W_y as long as $\tau \neq \hat{\lambda}_1/(1-q-k)$. Hence, $\text{index}_W(B_\tau,(0,\theta))$ is well defined if $\tau \neq \hat{\lambda}_1/(1-q-k)$. Next, we determine the index of B_τ at $(0,\theta)$. In view of the definition $Q(\omega,v)=(\omega,0)$, every eigenfunction of $Q\circ B'_\tau(0,\theta)$ has the form $(\omega,0)$, where ω is a nonzero solution of the equation

$$-d\omega'' + M\omega = \frac{1}{\lambda}(\tau(1-q-k)f_1(z-\theta)+M)\omega,\ \omega'(0)=\omega'(1)+\gamma\omega(1)=0.$$

We can proceed further as in the proof of Lemma 3.2 (see Appendix A) to show that $\text{index}_W(B_\tau,(0,\theta))=0$ if $\tau>\hat{\lambda}_1/(1-q-k)$ and $\text{index}_W(B_\tau,(0,\theta))=1$ if $\tau<\hat{\lambda}_1/(1-q-k)$. Hence, for any $\tau \in (\hat{\lambda}_1/(1-q-k), A]$, $\text{index}_W(B_\tau,\tilde{D}) = \text{index}_W(B_\tau,(0,0)) + \text{index}_W(B_\tau,(0,\theta)) = 0$. Meanwhile, by the homotopic invariance property of the fixed point index, we can claim that $\text{index}_W(B_\tau,\tilde{D})\equiv 0$ for any $\tau \in (\lambda_1/(1-q-k), A]$. However, for $\lambda_1/(1-q-k) < \tau < \hat{\lambda}_1/(1-q-k)$, $\text{index}_W(B_\tau,(0,0)) + \text{index}_W(B_\tau,(0,\theta)) = 1 \neq \text{index}_W(B_\tau,\tilde{D})$, which implies B_τ has at least a positive fixed point in $\tilde{D}$ for $\lambda_1/(1-q-k)<\tau<\hat{\lambda}_1/(1-q-k)$. Namely, (4.2) has a positive solution when $a \in (\lambda_1/(1-q-k), \hat{\lambda}_1/(1-q-k))$.

It remains to prove the instability of any positive solution (ω_0, v_0) of (4.2). To this end, let us consider the eigenvalue problem

$$\begin{aligned} &d\varphi'' + a(1-q-k)f_1(z-v_0)\varphi - a(1-q-k)\omega_0 f_1'(z-v_0)\psi + \eta\varphi = 0,\\ &d\psi'' + b[f_2(z-v_0) - v_0 f_2'(z-v_0)]e^{-c\omega_0}\psi - cbv_0 f_2(z-v_0)e^{-c\omega_0}\varphi + \eta\psi = 0,\\ &\varphi'(0)=\varphi'(1)+\gamma\varphi(1)=0, \qquad \psi'(0)=\psi'(1)+\gamma\psi(1)=0. \end{aligned} \tag{4.4}$$

It is well known (see, e.g., [11]) that one can put this eigenvalue problem in the context of spectral theory of compact strongly positive operators with respect to the order cone $P=\{(\varphi,\psi)\in X : \varphi\geq 0, \psi\leq 0\}$. In particular, by the Krein–Rutman theorem [5, 11], one can show (4.4) has an eigenvalue η_1, which has the following properties: it is real, algebraically simple, and all other eigenvalues have their real part greater than η_1. Moreover, η_1 corresponds to an eigenfunction (φ,ψ) in the interior of P, and it is the only eigenvalue with an eigenfunction in P. Thus it is called the principal eigenvalue of (4.4). The linearized stability criterion for (ω_0, v_0) can be expressed in terms of the principal eigenvalue: (ω_0, v_0) is asymptotically stable if $\eta_1 > 0$; it is unstable if $\eta_1 < 0$. On the other hand, multiplying the first equation of (4.4) by ω_0 and integrating, we obtain

$$\eta_1\int_0^1 \varphi\omega_0 dx = a(1-q-k)\int_0^1 \omega_0^2 f_1'(z-v_0)\psi dx.$$

Noting that (φ,ψ) belongs to the interior of P, we must have $\eta_1<0$, which implies the instability. □

The rest of this section is devoted to the proof of Theorems 1.2 and 1.3, which are important in understanding the effect of the inhibitor on the number of the coexistence solutions. In order to establish Theorem 1.2, we need the following technical results.

LEMMA 4.3. *For any $\epsilon>0$ small, there exists $M=M(\epsilon)$ large such that if $a\geq\lambda_1/(1-q-k)+\epsilon$, $\mu\geq M$,* (EP) *has a positive solution $(\tilde{u},\tilde{v})$ which satisfies*

$$(1-\delta)(1-q-k)\vartheta \leq \tilde{u} \leq (1-q-k)\vartheta,\quad (1-\delta)q\vartheta \leq \tilde{v} \leq (q+\delta)\vartheta, \tag{4.5}$$

where ϑ are the unique positive solutions of (1.2)*, and $0<\delta\leq\delta_0$, where $\delta_0>0$ is small such that $(1+\delta_0)\vartheta < z$ on $[0,1]$.*

Proof. Suppose that (u, v) solves (EP). Let $\chi = (1-q-k)v - qu$. Then (u, χ) satisfies

$$\begin{aligned} &du'' + a(1-q-k)uf_1(z - \tfrac{u+\chi}{1-q-k}) = 0,\\ &d\chi'' + b(\chi + qu)f_2(z - \tfrac{u+\chi}{1-q-k})e^{-\mu c u} = 0, \end{aligned} \tag{4.6}$$

with the usual boundary conditions. Since $0 < \vartheta < z$ on $[0, 1]$, we can claim that there exists $\delta_0 > 0$ small such that $(1+\delta_0)\vartheta < z$ on $[0, 1]$. Set

$$\Sigma = \{(u, \chi) \in X : (1-\delta_0)(1-q-k)\vartheta \le u \le (1-q-k)\vartheta,\ 0 \le \chi \le \delta_0(1-q-k)\vartheta\}.$$

Next, we show (4.6) is quasi-monotone decreasing [35] on Σ provided that μ is large enough. Clearly, $h_1(u, \chi) = a(1-q-k)uf_1(z - \frac{u+\chi}{1-q-k})$ is quasi-monotone decreasing on Σ. On the other hand, let $h_2(u, \chi) = b(\chi + qu)f_2(z - \frac{u+\chi}{1-q-k})e^{-\mu c u}$. Then

$$\frac{\partial h_2(u,\chi)}{\partial u} = be^{-\mu cu}[qf_2(z - \tfrac{u+\chi}{1-q-k}) - \tfrac{\chi+qu}{1-q-k}f_2'(z - \tfrac{u+\chi}{1-q-k}) - \mu c(\chi + qu)f_2(z - \tfrac{u+\chi}{1-q-k})].$$

Recalling that $(1+\delta_0)\vartheta < z$ on $[0, 1]$, it is easy to see that $\frac{\partial h_2(u,\chi)}{\partial u} < 0$ on Σ provided that μ is large enough. That is, $h_2(u, \chi)$ is quasi-monotone decreasing on Σ provided that μ is large enough.

Let $(\bar{u}, \underline{\chi}) = ((1-q-k)\vartheta(a), 0)$ and $(\underline{u}, \bar{\chi}) = ((1-\delta)(1-q-k)\vartheta, \delta(1-q-k)\vartheta)$. By the super- and subsolution method, it suffices to show that $(\bar{u}, \underline{\chi})$ and $(\underline{u}, \bar{\chi})$ are pairs of super-sub solutions of (4.6) for large μ. That is, we need to show that the inequalities

$$\begin{aligned} &d\bar{u}'' + a(1-q-k)\bar{u}f_1(z - \tfrac{\bar{u}+\underline{\chi}}{1-q-k}) \le 0,\\ &d\underline{\chi}'' + b(\underline{\chi} + q\bar{u})f_2(z - \tfrac{\bar{u}+\underline{\chi}}{1-q-k})e^{-\mu c\bar{u}} \ge 0 \end{aligned}$$

and

$$\begin{aligned} &d\underline{u}'' + a(1-q-k)\underline{u}f_1(z - \tfrac{\underline{u}+\bar{\chi}}{1-q-k}) \ge 0,\\ &d\bar{\chi}'' + b(\bar{\chi} + q\underline{u})f_2(z - \tfrac{\underline{u}+\bar{\chi}}{1-q-k})e^{-\mu c\underline{u}} \le 0 \end{aligned}$$

hold. It is trivial to check the inequalities for $\bar{u}, \underline{\chi}$, and $\underline{u}$. For $\bar{\chi}$ to satisfy the above inequality, it suffices to have

$$e^{-\mu c(1-\delta)(1-q-k)\vartheta} \le \frac{\delta a(1-q-k)(k_2+z-\theta)}{b((1-q)\delta+q)(k_1+z-\theta)}.$$

It is well known that there exists $B > 1$ large enough such that $Bk_1 > k_2$. Hence, $\frac{k_2+z-\theta}{k_1+z-\theta} > \frac{k_2}{Bk_1}$. Since $a \ge \lambda_1/(1-q-k) + \epsilon$ and $\vartheta = \vartheta(a) \ge \vartheta(\frac{\lambda_1}{1-q-k} + \epsilon)$, we need to have only

$$e^{-\mu c(1-\delta)(1-q-k)\vartheta(\frac{\lambda_1}{1-q-k}+\epsilon)} \le \frac{\delta[\lambda_1 + \epsilon(1-q-k)]k_2}{b((1-q)\delta+q)Bk_1},$$

where $\vartheta(\frac{\lambda_1}{1-q-k}+\epsilon)$ is the unique positive solution of (1.2) with $a = \frac{\lambda_1}{1-q-k}+\epsilon$. Clearly, this inequality holds as long as μ is sufficiently large. Namely, as long as μ is large enough, we have

$$d\bar{\chi}'' + b(\bar{\chi} + q\underline{u})f_2\left(z - \frac{\underline{u}+\bar{\chi}}{1-q-k}\right)e^{-\mu c\underline{u}} \le 0.$$

Thus $(\bar{u}, \underline{\chi})$ and $(\underline{u}, \bar{\chi})$ are the order upper and lower solutions of (4.6). It follows the existence-comparison theorem for elliptic systems that (4.6) has a solution $(\tilde{u}, \tilde{\chi})$, which satisfies

$$(1-\delta)(1-q-k)\vartheta \le \tilde{u} \le (1-q-k)\vartheta, \quad 0 \le \tilde{\chi} \le \delta(1-q-k)\vartheta.$$

Noting that $v = \frac{\chi+qu}{1-q-k}$, we know that (EP) has a positive solution $(\tilde{u}, \tilde{v})$, which satisfies (4.5). □

LEMMA 4.4. *For any $\epsilon > 0$ small and any $A > \lambda_1/(1-q-k)$, there exists $M = M(\epsilon, A) > 0$ large such that if $a \in (\lambda_1/(1-q-k) + \epsilon, A]$ and $\mu \ge M$, then any positive solution of* (EP) *that satisfies* (4.5) *is nondegenerate and linearly stable.*

Proof. If $a \in (\lambda_1/(1-q-k) + \epsilon, A]$ and (u, v) satisfies (4.5), then it is easy to see that (EP) is a regular perturbation of (4.1) when μ is large. Since (4.1) has a unique positive solution $((1-q-k)\vartheta, q\vartheta)$ which is linearly stable, this lemma follows from a standard regular perturbation argument. □

As noted before, the next lemma shows rigorously that the positive solutions to (EP) are of two types.

LEMMA 4.5. *Suppose $a_i \to a \in (\frac{\lambda_1}{1-q-k}, +\infty), \mu_i \to \infty$, and (u_i, v_i) is a positive solution of* (EP) *with $(a, \mu) = (a_i, \mu_i)$. Then for large i, either (u_i, v_i) is close to $((1-q-k)\vartheta, q\vartheta)$ or $(\mu_i u_i, v_i)$ is close to (ω, v) in $C^1([0,1]) \times C^1([0,1])$, where (ω, v) is a positive solution of* (4.2). *Moreover, if $a_i \ge \frac{\hat{\lambda}_1}{1-q-k}$ for all large i and $a_i \to a$, then (u_i, v_i) converges to $((1-q-k)\vartheta, q\vartheta)$ in the C^1 norm.*

Proof. We argue by contradiction. Suppose we can find $a_i \to a \in (\frac{\lambda_1}{1-q-k}, +\infty)$, $\mu_i \to \infty$, and positive solution (u_i, v_i) bounded away from $((1-q-k)\vartheta, q\vartheta)$ and any positive solution of (4.2). First, by Lemma 2.5, $0 \le (1+c)u_i + v_i < z(x)$. Hence, by elliptic regularity and the Sobolev embedding theorems, we may assume the existence of a subsequence (if necessary), such that $u_i \to u$ and $v_i \to v$ in $C^1([0,1])$ for some $u, v \in C_B^1([0,1])$. Set $\omega_i = \mu_i u_i$ and $\chi_i = (1-q-k)v_i - qu_i$. Then (ω_i, χ_i) satisfies

$$\begin{aligned} &d\omega_i'' + a_i(1-q-k)\omega_i f_1(z-(1+c)u_i - v_i) = 0, \\ &d\chi_i'' + b(1-q-k)v_i f_2(z-(1+c)u_i - v_i)e^{-c\omega_i} = 0, \end{aligned} \tag{4.7}$$

with the usual boundary conditions. By passing to a subsequence, we have two possibilities.

Case a: $\mu_i \|u_i\|_\infty \to \infty$. In this case, one must have $\chi_i \to 0$. Indeed, it suffices to show $e^{-c\omega_i} \to 0$ almost everywhere in (0, 1) as $i \to \infty$. Let $\tilde{\omega}_i = \omega_i / \|\omega_i\|_\infty$. Then $\tilde{\omega}_i$ satisfies

$$-d\tilde{\omega}_i'' = a_i(1-q-k)\tilde{\omega}_i f_1(z-(1+c)u_i - v_i), \quad \tilde{\omega}_i'(0) = \tilde{\omega}_i'(1) + \gamma\tilde{\omega}_i(1) = 0.$$

By L^p estimates and the Sobolev embedding theorem, we may assume $\tilde{\omega}_i \to \tilde{\omega} \ge 0, \not\equiv 0$ in $C^1([0,1])$, and $\tilde{\omega}$ satisfies

$$-d\tilde{\omega}'' = a(1-q-k)\tilde{\omega} f_1(z-(1+c)u - v), \quad \tilde{\omega}'(0) = \tilde{\omega}'(1) + \gamma\tilde{\omega}(1) = 0.$$

Here $0 \le (1+c)u + v \le z$ because $0 < (1+c)u_i + v_i < z$. Therefore, $\tilde{\omega} > 0$ on [0, 1] by the strong maximum principle and Hopf boundary lemma. Thus $e^{-c\omega_i} = e^{-c\|\omega_i\|_\infty \tilde{\omega}_i} \to 0$ as $i \to \infty$, which implies $\chi_i \to 0$. Hence, $(1-q-k)v = qu$, and

$$du'' + a(1-q-k)uf_1\left(z - \frac{u}{1-q-k}\right) = 0, \qquad u'(0) = u'(1) + \gamma u(1) = 0.$$

This implies $u \equiv 0$ or $u = (1-q-k)\vartheta$. If $u \equiv 0$, then $v = \frac{q}{1-q-k}u \equiv 0$. That is, $(u_i, v_i) \to (0,0)$ as $i \to \infty$. Hence, $\tilde{u}_i = u_i/\|u_i\|_\infty$ satisfies

$$d\tilde{u}_i'' + a_i(1-q-k)\tilde{u}_i f_1(z-(1+c)u_i - v_i) = 0, \quad \tilde{u}_i'(0) = \tilde{u}_i'(1) + \gamma\tilde{u}_i(1) = 0.$$

Similarly, by L^p estimates and the Sobolev embedding theorem, we may assume that $\tilde{u}_i \to \tilde{u} \geq 0, \not\equiv 0$ in C^1, and in view of the strong maximum principle, $\tilde{u} > 0$ and satisfies

$$d\tilde{u}'' + a(1-q-k)\tilde{u}f_1(z) = 0, \quad \tilde{u}'(0) = \tilde{u}'(1) + \gamma\tilde{u}(1) = 0,$$

which means $a = \lambda_1/(1-q-k)$, a contradiction. Hence $u = (1-q-k)\vartheta$ and $v = q\vartheta$, which contradicts our assumption.

Case b: $\mu_i\|u_i\|_\infty$ is uniformly bounded, which implies $u_i \to 0$ as $i \to \infty$. Hence, $\chi_i = (1-q-k)v_i - qu_i \to (1-q-k)v$. Since ω_i is uniformly bounded, by the equation for ω_i, we may assume that $\omega_i \to \omega$ in $C^1([0,1])$. It follows from (4.7) that (ω, v) satisfies (4.2). If $\omega \geq 0, \not\equiv 0$, then the strong maximum principle tells us that $\omega > 0$. On the other hand, we claim that $v > 0$ on [0, 1]. Otherwise,

$$d\omega'' + a(1-q-k)\omega f_1(z) = 0, \qquad x \in (0,1), \qquad \omega'(0) = \omega'(1) + \gamma\omega(1) = 0,$$

which implies $a = \lambda_1/(1-q-k)$, a contradiction. Hence, (a, ω, v) is a positive solution of (4.2), which contradicts our assumption that (a_i, ω_i, v_i) is bounded away from any positive solution of (4.2). Therefore, we must have $\omega \equiv 0$. It follows that $v \equiv 0$ or $v = \theta$. Suppose $v \equiv 0$. Then $v_i \to 0$ and $\tilde{\omega}_i = \omega_i/\|\omega_i\|_\infty$ satisfies

$$d\tilde{\omega}_i'' + a_i(1-q-k)\tilde{\omega}_i f_1(z-(1+c)u_i - v_i) = 0, \quad \tilde{\omega}_i'(0) = \tilde{\omega}_i'(1) + \gamma\tilde{\omega}_i(1) = 0.$$

By L^p estimates and the Sobolev embedding theorem, we may assume $\tilde{\omega}_i \to \tilde{\omega} \geq 0, \not\equiv 0$ in $C^1([0,1])$, and by virtue of the strong maximum principle, $\tilde{\omega} > 0$ satisfies

$$d\tilde{\omega}'' + a(1-q-k)\tilde{\omega}f_1(z) = 0, \quad \tilde{\omega}'(0) = \tilde{\omega}'(1) + \gamma\tilde{\omega}(1) = 0,$$

which means $a = \lambda_1/(1-q-k)$, a contradiction. Thus $(\omega_i, v_i) \to (0, \theta)$, and hence $a_i(1-q-k) = \lambda_1(f_1(z-(1+c)u_i - v_i)) \to \lambda_1(f_1(z-\theta)) = \hat{\lambda}_1$. That is, $a_i \to \hat{\lambda}_1/(1-q-k)$. On the other hand, we can show that (4.2) has a positive solution branch bifurcating from $(a, \omega, v) = (\hat{\lambda}_1/(1-q-k), 0, \theta)$ (see Lemma 4.7). Hence, we can find $a = \tilde{a}_i \to \hat{\lambda}_1/(1-q-k)$ such that (4.2) with $a = \tilde{a}_i$ has a positive solution $(\tilde{\omega}_i, \tilde{v}_i)$ converging in L^∞ to $(0, \theta)$. Thus $(a_i, \mu_i u_i, v_i)$ is close to $(\tilde{a}_i, \tilde{\omega}_i, \tilde{v}_i)$ for i large. This again contradicts our assumption. This finishes the proof of the first part of this lemma.

Now, we prove that if $a_i \geq \hat{\lambda}_1/(1-q-k)$ for all large i and $a_i \to a$ as $i \to \infty$, then $(u_i, v_i) \to ((1-q-k)\vartheta, q\vartheta)$. Again we use an indirect argument. We suppose that this is not true. Then by the first part of this lemma and by choosing a subsequence if necessary, we may assume that $(\mu_i u_i, v_i)$ is close to a positive solution of (4.2). This implies $u_i \to 0$ as $i \to \infty$. We divide the arguments into two cases: (i) $a > \hat{\lambda}_1/(1-q-k)$ and (ii) $a = \hat{\lambda}_1/(1-q-k)$.

In case (i), suppose for any $\epsilon > 0$, there exists $a_i \to a \geq \hat{\lambda}_1/(1-q-k) + \epsilon$ such that $u_i \to 0$ as $\mu_i \to \infty$. Then $\omega_i = \mu_i u_i, \chi_i = (1-q-k)v_i - qu_i$ satisfy (4.7). Passing to a subsequence, we have two possibilities.

Case a: $\|\omega_i\|_\infty = \mu_i \|u_i\|_\infty \to \infty$. Noting Lemma 2.5, we claim that $\chi_i \to 0$ as before, which means $v_i = \frac{\chi_i + qu_i}{1-q-k} \to 0$. Let $\tilde{u}_i = u_i/\|u_i\|_\infty$. Then by L^p estimates and the Sobolev embedding theorem, we may assume $\tilde{u}_i \to \tilde{u}$ in $C^1([0,1])$, and by the strong maximum principle, $\tilde{u} > 0$. Moreover, $\tilde{u}$ satisfies

$$d\tilde{u}'' + a(1-q-k)\tilde{u}f_1(z) = 0, \qquad \tilde{u}'(0) = \tilde{u}'(1) + \gamma\tilde{u}(1) = 0,$$

which implies $a = \lambda_1/(1-q-k)$, a contradiction.

Case b: $\|\omega_i\|_\infty = \mu_i\|u_i\|_\infty \leq C$. Then by using a priori estimates for v_i (see Lemma 2.5), we may assume that $(\omega_i, v_i) \to (\omega, v)$ in $C^1([0,1])$, where $\omega, v \geq 0$ on [0, 1]. Noting that $\chi_i = (1-q-k)v_i - qu_i$ and $u_i \to 0$, one has $\chi_i \to (1-q-k)v$. It follows from the equations in (4.7) that (ω, v) satisfies (4.2). Namely, (ω, v) is exactly the nonnegative solution of (4.2). If $\omega \geq 0, \not\equiv 0$, then by the strong maximum principle, $\omega > 0$. Hence, $(\hat{\lambda}_1 <) a(1-q-k) = \lambda_1(f_1(z-v))$, which implies $v \not\equiv 0$. It follows from the strong maximum principle that $v > 0$. This contradicts Lemma 4.2; that is, (4.2) has no positive solution provided that $a > \hat{\lambda}_1/(1-q-k)$. Therefore, $\omega \equiv 0$ and $v = \theta$ (the possibility $v \equiv 0$ can be ruled out by similar arguments as in the proof of the first part of this lemma). Set $\tilde{u}_i = u_i/\|u_i\|_\infty$. A similar argument shows that $a = \hat{\lambda}_1/(1-q-k)$, a contradiction. Therefore, our assertion holds.

In case (ii), since $a_i \to \hat{\lambda}_1/(1-q-k)$ and $u_i \to 0$, one can assert that $v_i \to \theta$ and $\mu_i u_i \to 0$ in C^1 norm. Indeed, let $\tilde{u}_i = u_i/\|u_i\|_\infty$. Then $\tilde{u}_i$ satisfies

$$d\tilde{u}_i'' + a_i(1-q-k)\tilde{u}_i f_1(z-(1+c)u_i - v_i) = 0, \quad \tilde{u}_i'(0) = \tilde{u}_i'(1) + \gamma\tilde{u}_i(1) = 0.$$

Similarly, we may suppose $\tilde{u}_i \to \tilde{u}$ in $C^1([0,1])$ and $\tilde{u} > 0$ satisfies

$$d\tilde{u}'' + \hat{\lambda}_1\tilde{u}f_1(z-v) = 0, \quad \tilde{u}'(0) = \tilde{u}'(1) + \gamma\tilde{u}(1) = 0, \tag{4.8}$$

which implies $v \not\equiv 0$; otherwise, $\hat{\lambda}_1 = \lambda_1$, a contradiction. Noting that $a_i \to \hat{\lambda}_1/(1-q-k), u_i \to 0, v_i \to v \not\equiv 0$, and

$$dv_i'' + bv_i f_2(z-(1+c)u_i - v_i)e^{-c\mu_i\|u_i\|_\infty\tilde{u}_i} + a_i q u_i f_1(z-(1+c)u_i - v_i) = 0,$$

we can show that $\mu_i\|u_i\|_\infty$ is uniformly bounded. Hence we may assume that $\mu_i u_i \to \omega$ in $C^1([0,1])$ for some $\omega \geq 0$. Letting $i \to \infty$, we must have $dv'' + bvf_2(z-v) \geq 0$, which means $v \leq \theta$. Multiplying (4.8) by $\hat{\varphi}_1$, integrating over [0, 1], and applying Green's formula, we obtain

$$\hat{\lambda}_1 \int_0^1 \tilde{u}\hat{\varphi}_1(f_1(z-v) - f_1(z-\theta))dx = 0,$$

which implies $v = \theta$ since $v \leq \theta$. Moreover, $\tilde{u} = \hat{\varphi}_1$. That is, $v_i \to \theta$. Next, we show $\omega \equiv 0$. If $\omega \geq 0, \not\equiv 0$, then $\omega > 0$ by the strong maximum principle. Noting that $a_i \to \hat{\lambda}_1/(1-q-k), u_i \to 0, v_i \to \theta, \mu_i u_i \to \omega$, we have

$$d\theta'' + b\theta f_2(z-\theta)e^{-c\omega} = 0, \quad \theta'(0) = \theta'(1) + \gamma\theta(1) = 0.$$

This means $b = \lambda_1(f_2(z-\theta)e^{-c\omega}) > \lambda_1(f_2(z-\theta)) = b$, a contradiction. Hence our assertion holds. Next, we show $(1+c)u_i + v_i < \theta$ for large i. Let $Q_i = (1+c)u_i + v_i$.

Clearly, $Q_i \to \theta$, and

$$dQ_i'' + a_i u_i f_1(z - Q_i) + b v_i f_2(z - Q_i) e^{-c\mu_i u_i} = 0, \quad Q_i'(0) = Q_i'(1) + \gamma Q_i(1) = 0.$$

Hence

$$\begin{aligned}
&dQ_i'' + bQ_i f_2(z - Q_i) \\
&= u_i[b(1+c)f_2(z - Q_i) - a_i f_1(z - Q_i)] + b v_i f_2(z - Q_i)[1 - e^{-c\mu_i u_i}] \\
&= u_i[b(1+c)f_2(z - Q_i) - a_i f_1(z - Q_i) + b v_i f_2(z - Q_i) c\mu_i + O(\mu_i^2 u_i)] \\
&= u_i[b(1+c)f_2(z - Q_i) - a_i f_1(z - Q_i) + (bcv_i f_2(z - Q_i) + O(\mu_i u_i))\mu_i].
\end{aligned}$$

Since $\mu_i u_i \to 0$ and $\mu_i \to \infty$, we have $dQ_i'' + bQ_i f_2(z - Q_i) > 0$ for large i, which implies $Q_i < \theta$ for large i.

Now multiplying the equation for u_i by $\hat{\varphi}_1$ and integrating over [0, 1], we obtain

$$\int_0^1 [a_i(1 - q - k)f_1(z - Q_i) - \hat{\lambda}_1 f_1(z - \theta)]\hat{\varphi}_1 u_i dx = 0.$$

Since $a_i(1 - q - k) \geq \hat{\lambda}_1$ and $f_1(z - Q_i) > f_1(z - \theta)$ for large i, $\int_0^1 [a_i(1 - q - k)f_1(z - Q_i) - \hat{\lambda}_1 f_1(z - \theta)]\hat{\varphi}_1 u_i dx > 0$ for all large i, a contradiction. Hence $(u_i, v_i) \to ((1 - q - k)\vartheta, q\vartheta)$ in the C^1 norm when $a_i \geq \frac{\hat{\lambda}_1}{1-q-k}$ for all large i, $a_i \to a$, and $\mu_i \to \infty$. □

LEMMA 4.6. (i) *For any* $A \geq \frac{\hat{\lambda}_1}{1-q-k}$, *there exists* $M > 0$ *large such that if* $\mu > M$ *and* $a \in [\frac{\hat{\lambda}_1}{1-q-k}, A]$, *then any positive solution* (u, v) *of* (EP) *is nondegenerate and linearly asymptotically stable, and* $\mathrm{index}_W(\mathcal{A}, (u, v)) = 1$.

(ii) *For any* $\epsilon, \delta > 0$ *small, there exists* $M_{\epsilon,\delta} > 0$ *large such that if* $a \in [\frac{\lambda_1}{1-q-k} + \epsilon, \frac{\hat{\lambda}_1}{1-q-k})$ *and* $\mu \geq M_{\epsilon,\delta}$ *and if* (u, v) *is a positive solution of* (EP), *then either* (a) $\|u - (1 - q - k)\vartheta\|_{C^1} + \|v - q\vartheta\|_{C^1} < \delta$ *or* (b) $\|\mu u - \tilde{\omega}\|_{C^1} + \|v - \tilde{v}\|_{C^1} + \|a - \tilde{a}\|_{C^1} < \delta$, *where* $(\tilde{\omega}, \tilde{v})$ *is a positive solution of* (4.2) *with* $a = \tilde{a}$. *Moreover, if* (a) *occurs, then* (u, v) *is nondegenerate linearly asymptotically stable and* $\mathrm{index}_W(\mathcal{A}, (u, v)) = 1$.

Proof. (i) We prove the nondegeneracy and linear stability first. For this purpose, we consider the linearized eigenvalue problem

$$\begin{aligned}
&d\phi'' + a(1 - q - k)[f_1(z - (1+c)u - v) - (1+c)uf_1'(z - (1+c)u - v)]\phi \\
&\quad - a(1 - q - k)uf_1'(z - (1+c)u - v)\psi = -\eta\phi, \\
&d\psi'' + [b(f_2(z - (1+c)u - v) - vf_2'(z - (1+c)u - v))e^{-\mu cu} \\
&\quad - aquf_1'(z - (1+c)u - v)]\psi \\
&\quad - bv[f_2'(z - (1+c)u - v)(1+c) + \mu c f_2(z - (1+c)u - v)]e^{-\mu cu}\phi \\
&\quad + aq[f_1(z - (1+c)u - v) - (1+c)uf_1'(z - (1+c)u - v)]\phi = -\eta\psi, \\
&\phi'(0) = \phi'(1) + \gamma\phi(1) = 0, \qquad \psi'(0) = \psi'(1) + \gamma\psi(1) = 0.
\end{aligned}$$

By Lemma 4.5, (EP) has no positive solution with a small u component when $a \in [\frac{\hat{\lambda}_1}{1-q-k}, A]$ and μ is large. Therefore, we can establish this assertion by a simple variant of the proof of Lemma 4.4.

Next, we prove the statement concerning the fixed point index. Since any positive solution (u, v) to (EP) is nondegenerate, we have

$$\mathrm{index}_W(\mathcal{A}, (u, v)) = \mathrm{index}_X(\mathcal{A}, (u, v)) = \mathrm{index}_X(\mathcal{A}'(u, v), (0, 0)).$$

Let $Q_t(\phi,\psi) = (-d\frac{d^2}{dx^2} + M)^{-1}(f,g)$, where $0 \le t \le 1$ and

$$\begin{aligned} f = \;& a(1-q-k)[f_1(z-(1+c)u-v) - (1+c)uf_1'(z-(1+c)u-v)]\phi + M\phi \\ & -ta(1-q-k)uf_1'(z-(1+c)u-v)\psi, \\ g = \;& [b(f_2(z-(1+c)u-v) - vf_2'(z-(1+c)u-v))e^{-\mu cu} \\ & -taquf_1'(z-(1+c)u-v)]\psi + M\psi \\ & -bv[(1+c)f_2'(z-(1+c)u-v) + \mu cf_2(z-(1+c)u-v)]e^{-\mu cu}\phi \\ & +aq[f_1(z-(1+c)u-v) - (1+c)uf_1'(z-(1+c)u-v)]\phi. \end{aligned}$$

Then there exists a neighborhood $U_\delta \subset X$ of $(0,0)$ such that Q_t has no fixed point on ∂U_δ provided μ is large enough. Moreover, we can choose U_δ such that $\mathcal{A}'(u,v)(\phi,\psi) = (\phi,\psi)$ has only the solution $(\phi,\psi) = (0,0)$ in U_δ. By similar arguments as in Lemma 2.5 in [7] and Theorem 3.1 in [8], we can show $\operatorname{index}_X(\mathcal{A}'(u,v),(0,0)) = \operatorname{index}_X(\mathcal{A}'(u,v),U_\delta) = \operatorname{index}_X(Q_1,U_\delta) = \operatorname{index}_X(Q_0,U_\delta) = \operatorname{index}_X(Q_0,(0,0)) = 1$. Hence $\operatorname{index}_W(\mathcal{A},(u,v)) = 1$.

(ii) The statement on the location of the positive solutions follows directly from Lemma 4.5. The other statements are proved in the same way as in (i) above. □

Proof of Theorem 1.2. (i) For any $\epsilon > 0$ small, let $M = \max\{M(\epsilon), M(\epsilon, \hat{\lambda}_1/(1-q-k))\}$, where $M(\epsilon), M(\epsilon,\hat{\lambda}_1/(1-q-k))$ are given by Lemmas 4.3 and 4.4, respectively. Assume that for $\mu \ge M$ and $a \in [\lambda_1/(1-q-k)+\epsilon, \hat{\lambda}_1/(1-q-k))$, (EP) has only a unique positive solution $(\tilde{u},\tilde{v})$ as shown in Lemma 4.3. In view of Lemma 4.4, $I - \mathcal{A}'(\tilde{u},\tilde{v})$ is invertible in X and $\mathcal{A}'(\tilde{u},\tilde{v})$ has no real eigenvalue greater than one, where $\mathcal{A}'(\tilde{u},\tilde{v})$ is the Fréchet derivative of $\mathcal{A}$ at $(\tilde{u},\tilde{v})$. We can argue in the same way as in the proof of Theorem 3.1 in [8] to draw a conclusion that $\operatorname{index}_W(\mathcal{A},(\tilde{u},\tilde{v})) = 1$. By virtue of Lemmas 3.1 and 3.2, it follows that

$$1 = \operatorname{index}_W(\mathcal{A},D') = \operatorname{index}_W(\mathcal{A},(0,0)) + \operatorname{index}_W(\mathcal{A},(0,\theta)) + \operatorname{index}_W(\mathcal{A},(\tilde{u},\tilde{v})) = 2.$$

This contradiction completes the proof.

(ii) It follows from Lemma 4.6 that for any $A \ge \frac{\hat{\lambda}_1}{1-q-k}$, there exists $M > 0$ large such that any positive solution (u,v) of (EP) is nondegenerate and linearly asymptotically stable for $a \in [\frac{\hat{\lambda}_1}{1-q-k}, A]$ and $\mu \ge M$. Hence, it suffices to show the uniqueness. Set $D_1 = \{(u,v) \in X : \frac{(1-q-k)}{2}\vartheta < u < \frac{\vartheta}{1+c}, \frac{q}{2}\vartheta < v < \max_{[0,1]}\bar{v} + 1\}$, and define $F_\tau : D_1 \to W$ by

$$F_\tau(u,v) = \left(-d\frac{d^2}{dx^2} + K\right)^{-1}(a(1-q-k)ug_1(u,v)+Ku,\ \tau bvg_2(u,v)+aqug_1(u,v)+Kv),$$

where $\tau \in [0,1]$, $g_1(u,v) = f_1(z-(1+c)u-v)$, $g_2(u,v) = f_2(z-(1+c)u-v)e^{-\mu cu}$, and K is large enough such that $K+a(1-q-k)g_1(u,v) > 0$ and $K+\tau bg_2(u,v)-aquK_1 > 0$ (K_1 is given in section 3) for all $(u,v) \in D_1$ and $\tau \in [0,1]$. Clearly, F_τ is a compact and continuously differentiable operator. Moreover, it follows from Lemma 4.5 that there exists $M > 0$ large such that if $\mu \ge M$ and $a \in [\hat{\lambda}_1/(1-q-k), A]$, then any positive solution (u,v) of (EP) is close to $((1-q-k)\vartheta, q\vartheta)$. Hence, $(u,v) \in D_1$ for $a \in [\hat{\lambda}_1/(1-q-k), A]$ and $\mu \ge M$. Namely, if $a \in [\hat{\lambda}_1/(1-q-k), A]$ and $\mu \ge M$, then (u,v) is a positive solution of (EP) if and only if it is a fixed point of F_1 in D_1. Again by Lemma 4.5, F_τ has no fixed point on ∂D_1 for $a \in [\hat{\lambda}_1/(1-q-k), A]$ and $\mu \ge M$. Therefore, $\operatorname{index}_W(F_\tau, D_1) \equiv$ const. In particular, $\operatorname{index}_W(F_1,D_1) = \operatorname{index}_W(F_0,D_1)$. It is easy to show that F_0 has a unique fixed point $((1-q-k)\vartheta, q\vartheta)$ in D_1 and $\operatorname{index}_W(F_0,D_1) = \operatorname{index}_W(F_0,((1-q-k)\vartheta, q\vartheta)) = 1$. Hence, $\operatorname{index}_W(F_1,D_1) = 1$.

As mentioned before, from Lemma 4.6, we know that, for $\mu > M$ and $a \in [\hat{\lambda}_1/(1-q-k), A]$, all fixed points of F_1 in D_1 are nondegenerate and linearly stable. Hence by a compactness argument it is easy to show that there are at most finitely many fixed points of F_1, which are denoted by $\{(u_i, v_i)\}_{i=1}^n$. By Lemma 4.6 again, $\text{index}_W(\mathcal{A}, (u_i, v_i)) = 1$. In view of the additivity property of the fixed point index, we have for $a \in [\hat{\lambda}_1/(1-q-k), A]$

$$n = \sum_{i=1}^{n} \text{index}_W(F_1, (u_i, v_i)) = \text{index}_W(F_1, D_1) = 1.$$

Hence for $\mu \geq M$ and $a \in [\hat{\lambda}_1/(1-q-k), A]$, (EP) has only a unique positive solution and it is stable. The proof of Theorem 1.2 is completed. □

Next we wish to establish Theorem 1.3, but first we give the following lemma, which is crucial in proving Theorem 1.3.

LEMMA 4.7. *There exists $\epsilon > 0$ small such that if $\hat{\lambda}_1/(1-q-k) - \epsilon \leq a < \hat{\lambda}_1/(1-q-k)$, then* (4.2) *has a unique positive solution.*

Proof. Here, we prove this lemma by the local bifurcation theorem of Crandall and Rabinowitz [3]. We regard a as the bifurcation parameter and try to construct a positive solution branch from the semitrivial nonnegative solution branch $\{(a, 0, \theta) : a \in R^+\}$.

After some standard calculations, we obtain that $(\hat{\lambda}_1/(1-q-k), 0, \theta)$ is a bifurcation point. Close to this bifurcation point, (4.2) has a positive solution $(a(s), s(\hat{\varphi}_1 + \Phi(s)), \theta + s(\chi_1 + \Psi(s)))$ $(0 < s \ll 1)$, where $a(0) = \hat{\lambda}_1/(1-q-k), \chi_1 = bcL_b^{-1}(\theta f_2(z-\theta)\hat{\varphi}_1) < 0, \Phi(0) = \Psi(0) = 0$. Putting this positive solution into the first equation of (4.2), dividing by s, and differentiating with respect to s, it follows that the derivative of $a(s)$ with respect to s at $s = 0$ is less than 0. That is, $a'(0) < 0$, which implies the positive solution bifurcation branch is to the left. Namely, there exists $\epsilon > 0$ sufficiently small such that if $\hat{\lambda}_1/(1-q-k) - \epsilon \leq a < \hat{\lambda}_1/(1-q-k)$, then (4.2) has a positive solution with the form of $(a(s), s(\hat{\varphi}_1 + \Phi(s)), \theta + s(\chi_1 + \Psi(s)))$ $(0 < s \ll 1)$. Furthermore, it is unique as long as ϵ is sufficiently small. In fact, it is also unstable. We leave the proof of this assertion to the reader. □

Proof of Theorem 1.3. First we show that for large μ (EP) has a unique asymptotically stable positive solution which is close to $((1-q-k)\vartheta, q\vartheta)$. In fact, if we choose $\delta > 0$ small enough in Lemma 4.6, then by Lemma 4.6 any positive solution of (EP) close to $((1-q-k)\vartheta, q\vartheta)$ is nondegenerate and linearly stable. Next, by a simple variant of the proof of part (ii) of Theorem 1.2, we can find that (EP) has only one positive solution of type (a), and it is asymptotically stable.

On the other hand, we can show that (EP) has a unique unstable positive solution of type (b). If this assertion holds, then by Lemma 4.6 our proof is completed. Hence, our main task is to establish this assertion.

Suppose (u, v) is a positive solution of type (b) of (EP). It follows from Lemmas 4.6 and 4.7; $(\mu u, v)$ is close to (ω, v), where (ω, v) is the unique positive solution of (4.2). Hence to prove the uniqueness, it suffices to show that, for $a \in [\hat{\lambda}_1/(1-q-k) - \epsilon_0, \hat{\lambda}_1/(1-q-k))$ and $\mu \geq M_0$, there is a unique pair $(\mu u, v)$ close to (ω, v) for certain ϵ_0 and M_0.

Set $\hat{u} = \mu u, \epsilon = \frac{1}{\mu}$, and consider the following problem with the usual boundary conditions

$$\begin{aligned} &d\hat{u}'' + a(1-q-k)\hat{u}f_1(z-(1+c)\epsilon\hat{u}-v) = 0, \qquad x \in (0,1), \\ &dv'' + bvf_2(z-(1+c)\epsilon\hat{u}-v)e^{-c\hat{u}} + aq\epsilon\hat{u}f_1(z-(1+c)\epsilon\hat{u}-v) = 0. \end{aligned} \tag{4.9}$$

Clearly, (u, v) is a solution of (EP) if and only if $(\mu u, v)$ is a solution of (4.9) with $\epsilon = 1/\mu$. Thus it suffices to prove the uniqueness of (4.9). For fixed $\epsilon \geq 0$, regarding a as a bifurcation parameter, we see that $(\hat{\lambda}_1/(1-q-k), 0, \theta)$ is a simple bifurcation point of (4.9). By virtue of a variant of Theorem 1 in Crandall and Rabinowitz [2], there exists $\delta_1 > 0$ and C^1 curves

$$\Gamma_\epsilon = \{(a(\epsilon, s), \hat{u}(\epsilon, s), v(\epsilon, s)) : 0 < s < \delta_1\}, \qquad 0 \leq \epsilon \leq \delta_1,$$

such that if $0 \leq \epsilon \leq \delta_1$, then all positive solutions of (4.9) close to $(\hat{\lambda}_1/(1-q-k), 0, \theta) = (a(0,0), \hat{u}(0,0), v(0,0))$ lie on the curve Γ_ϵ. Hence, we need show only that for fixed ϵ, Γ_ϵ uniformly cover an a-range: $a \in [\hat{\lambda}_1/(1-q-k) - \epsilon_0, \hat{\lambda}_1/(1-q-k))$ only once for suitably chosen ϵ_0. It is easy to obtain

$$\frac{\partial a}{\partial s}(0,0) = \frac{\hat{\lambda}_1 \int_0^1 \hat{\varphi}_1 f_1'(z-\theta)\chi_1}{(1-q-k)\int_0^1 \hat{\varphi}_1^2 f_1(z-\theta)} < 0$$

based on $\chi_1 = L_b^{-1}(bc\theta f_2(z-\theta)\hat{\varphi}_1) < 0$. By taking δ_1 small, we may assume that $\frac{\partial a}{\partial s}(\epsilon, s) < 0$ for $0 \leq \epsilon, s \leq \delta_1$. Hence $\hat{\lambda}_1/(1-q-k) - a(0, \delta_1) = a(0,0) - a(0,\delta_1) > 0$. Since $a(\epsilon, s)$ is continuous, there exists $\delta \in (0, \delta_1]$ such that $\epsilon_0 = \min_{0 \leq \epsilon \leq \delta}(\hat{\lambda}_1/(1-q-k) - a(\epsilon, \delta_1)) > 0$. Therefore, if $a \geq \hat{\lambda}_1/(1-q-k) - \epsilon_0$, then $a(\epsilon, \delta_1) \leq a$ for any $\epsilon \in [0, \delta]$. This shows that for each $\epsilon \in [0, \delta]$, Γ_ϵ covers the a-range $[\hat{\lambda}_1/(1-q-k) - \epsilon_0, \hat{\lambda}_1/(1-q-k))$. Moreover, since $\frac{\partial a}{\partial s}(\epsilon, s) < 0$ for $0 \leq \epsilon, s \leq \delta_1$, each curve covers the range only once. By taking $M_0 = 1/\delta$, we see that, for $\mu \geq M_0$ and $\hat{\lambda}_1/(1-q-k) - \epsilon_0 \leq a < \hat{\lambda}_1/(1-q-k)$, (EP) has exactly one positive solution of type (b).

It remains to show the instability. A simple computation shows that η is an eigenvalue of the linearization of (EP) at (u, v) with eigenfunction (ϕ, ψ) if and only if it is an eigenvalue of that of (4.9) with $\epsilon = 1/\mu$ at $(\mu u, v)$ with eigenfunction $(\mu\phi, \psi)$. Hence it suffices to show that the linearization of (4.9) has a negative eigenvalue at any point on the bifurcation curves Γ_ϵ. This follows from a simple application of a variant of Theorem 1.16 in Crandall and Rabinowitz [3]. More precisely, by Lemma 1.3 in [3], we can obtain a variant of Corollary 1.13 there. That is, there exist $\tau > 0$ and C^1 functions $\gamma : (\hat{\lambda}_1/(1-q-k) - \tau, \hat{\lambda}_1/(1-q-k) + \tau) \times (-\tau, \tau) \to R^1$ and $\beta : (-\tau, \tau) \times (-\tau, \tau) \to R^1$ such that $\gamma(a, \epsilon)$ is a simple eigenvalue of the linearization of (4.9) at $(a, 0, \theta)$ and $\beta(s, \epsilon)$ is a simple eigenvalue of the linearization of (4.9) at $(a, u, v) = (a(\epsilon, s), \hat{u}(\epsilon, s), v(\epsilon, s))$ with $0 \leq \epsilon, s \leq \tau$. Moreover, $\gamma(\hat{\lambda}_1/(1-q-k), \epsilon) = \beta(0, \epsilon) = 0$. It is easy to check that, in fact, $\gamma(a, \epsilon)$ is a simple eigenvalue of

$$d\phi'' + a(1-q-k)\phi f_1(z-\theta) = -\gamma(a, \epsilon)\phi$$

with the usual boundary conditions. Hence, $\frac{\partial \gamma}{\partial a}(\hat{\lambda}_1/(1-q-k), \epsilon) < 0$ because of the monotone property. Then it follows from Theorem 1.16 in [3] that $\beta(s, 0) \sim -s\frac{\partial a}{\partial s}(0, s)\frac{\partial \gamma}{\partial a}(\hat{\lambda}_1/(1-q-k), 0)$ for $0 < s \ll 1$, which implies $\beta(s, 0) < 0$ and the positive solution of type (b) of (EP) is unstable. This completes the proof of Theorem 1.3. □

5. Numerical simulation. In this section, we present some results of our numerical simulations that complement the analytic results of the previous sections. All computations in this section are performed with Matlab.

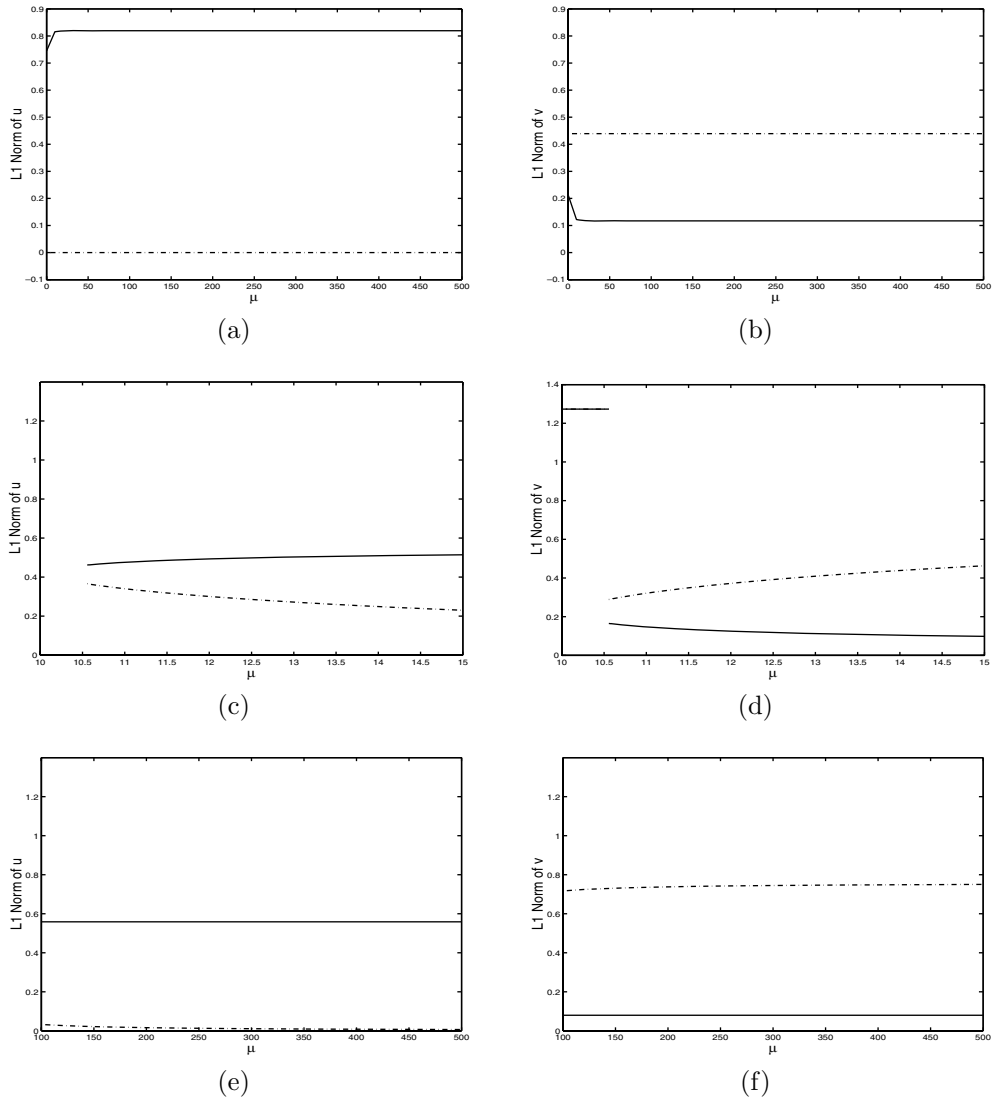


FIG. 1. *Effect of μ:* (a) *and* (b) *are the bifurcation diagrams of u and v, respectively, with respect to μ with the parameters $a = 4, b = 1.5$. Here the two solid lines in* (a) *and* (b) *represent the L^1 norm of components u and v of the stable coexistence solution (u, v), respectively. The two dashed lines in* (a) *and* (b) *represent the L^1 norm of components 0 and θ of the unstable semitrivial nonnegative solution $(0, \theta)$, respectively. Similarly, the pair of* (c) *and* (d) *and the pair of* (e) *and* (f) *are the bifurcation diagrams of u and v, respectively, with respect to μ all with $a = 2.5, b = 5$. Here solid lines denote the stable solutions and dashed lines represent the unstable solutions. Note that $\mu \in [10, 15]$ in* (c) *and* (d) *and that $\mu \in [100, 500]$ in* (e) *and* (f). *The aim of plotting in the above domain is to explicitly show the change tendency of u and v.*

Several parameters are common for all simulations: the diffusion rate $d = 1.0$ and parameters $k_1 = 1, k_2 = 1.1, \gamma = 1, q = 0.1$, and $k = 0.2$. The other parameters are varied in order to illustrate different outcomes. In Figures 1 and 2, the vertical axis is the L_1 norm of u or v. In Figures 3 and 4, the coexistence solutions to (PP) are plotted.

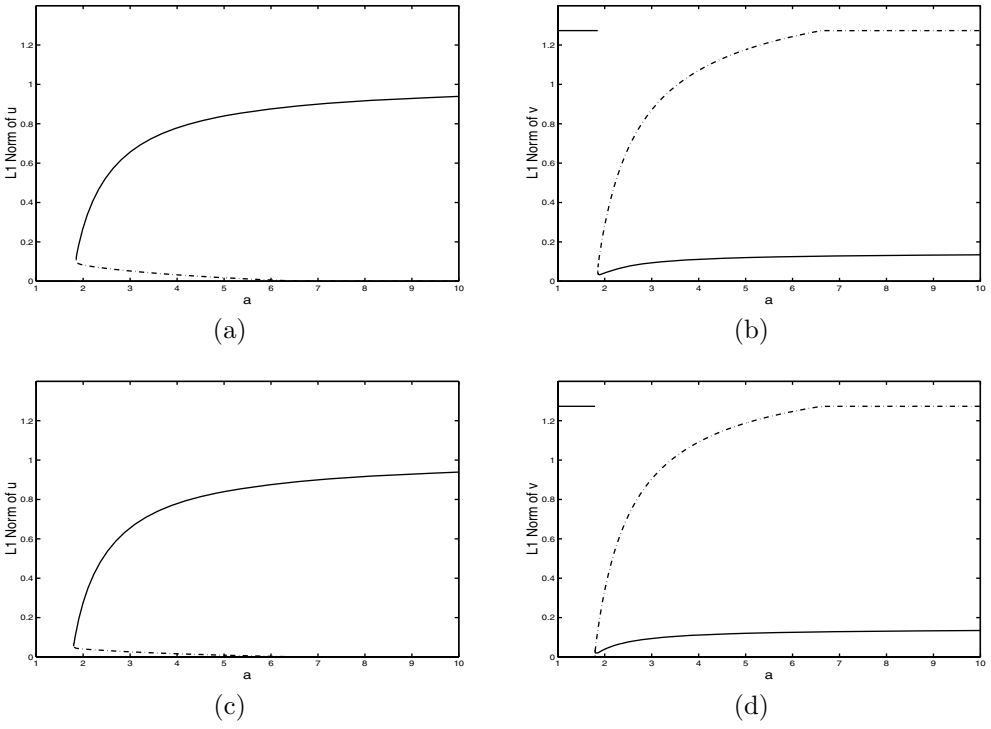


FIG. 2. *Bifurcation diagrams with respect to a:* (a) *and* (b) *with* $\mu = 50, b = 5$ *and* (c) *and* (d) *with* $\mu = 100, b = 5$ *also represent the bifurcation graphs of* u *and* v *with respect to* a*, respectively. Here solid lines denote the stable solutions and dashed lines represent the unstable solutions.*

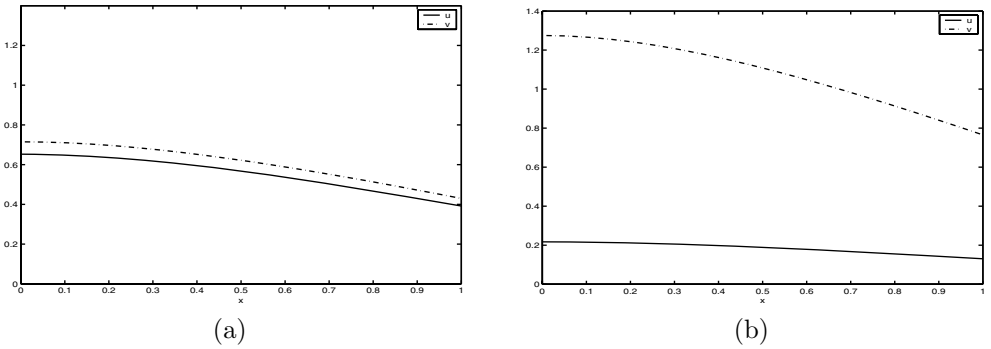


FIG. 3. *Two coexistence solutions of* (PP) *with* $\mu = 1, a = 6.4, b = 5$*. This indicates* (PP) *also has two coexistence solutions when* μ *is not large.*

The simulations presented below illustrate the following major outcomes of the plasmid-bearing and plasmid-free competition in the unstirred chemostat with an internal inhibitor.

(1) If u is a better competitor than v, there exists only a unique globally stable coexistence state of (PP) for any $\mu > 0$ (see Figures 1(a) and 1(b)). That is, if u is a better competitor, then it cannot eliminate its competitor but forces the existence of a coexistence state. This reflects the difference between the plasmid model and the standard competition model in the chemostat.

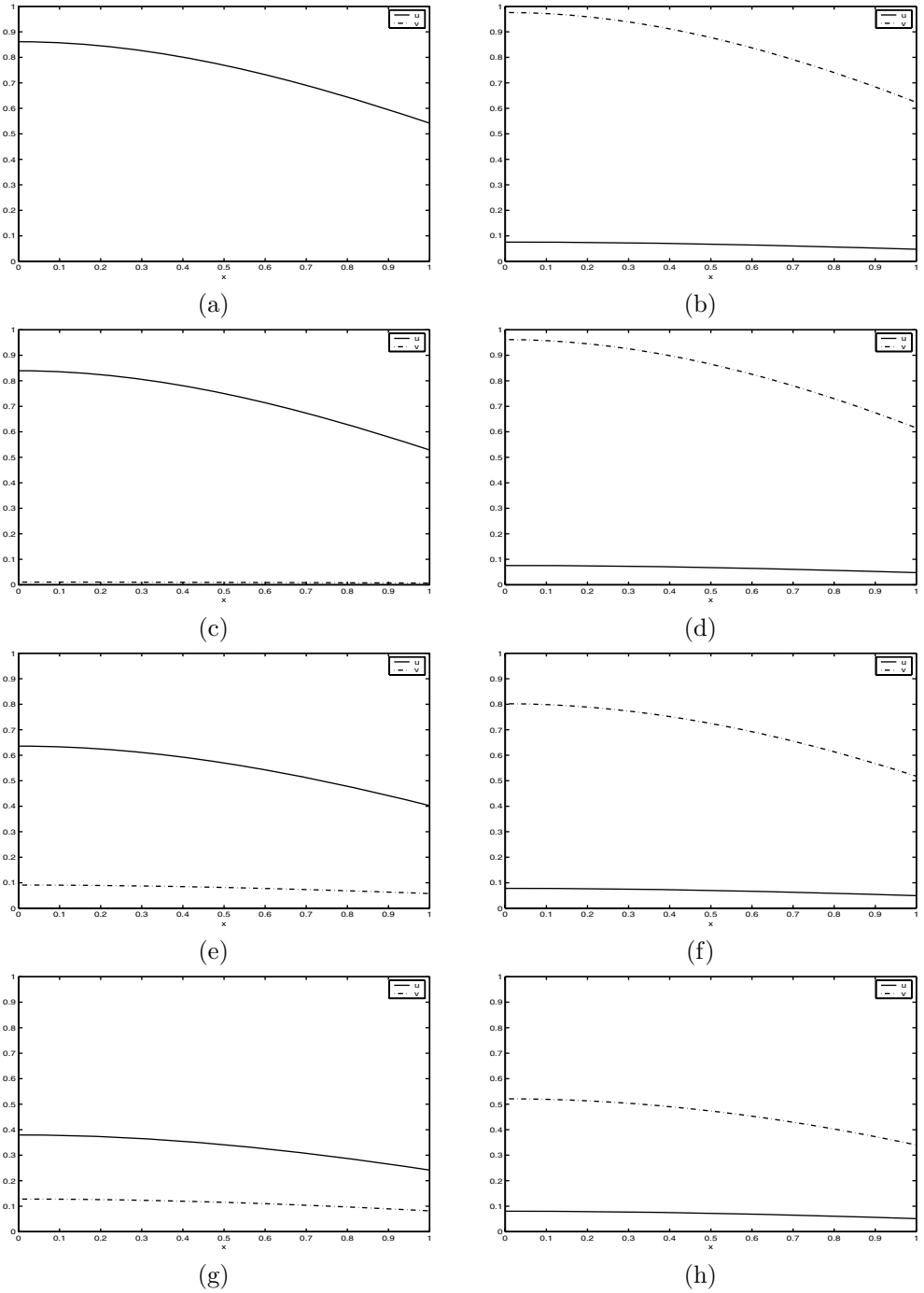


FIG. 4. *The difference between the plasmid model and the standard chemostat competition model in the presence of inhibitor:* (a) *and* (b) *with* $q = 0$, (c) *and* (d) *with* $q = 0.01$, (e) *and* (f) *with* $q = 0.1$, *and* (g) *and* (h) *with* $q = 0.2$. *Here* $a = 2.5, b = 5, \mu = 50$. *The simulations suggest that for large* μ *the plasmid model* $(q > 0)$ *has two coexistence solutions; one asymptotically stable,* (c), (e), *and* (g) *with* $q = 0.01, 0.1, 0.2$, *respectively, and the others unstable,* (d), (f) *and* (h) *with* $q = 0.01, 0.1, 0.2$, *respectively. However, the basic chemostat model* $(q = 0)$ *seems to have only one unstable coexistence solution,* (b). *Moreover, when* $q \to 0+$, *the stable coexistence solution of* (PP) *goes to the semitrivial nonnegative solution* $(\vartheta, 0)$, (a).

(2) If u is a weaker competitor than v, then there exists a unique number $\mu^* > 0$ such that if $\mu < \mu^*$ there is no coexistence state of (PP) and the semitrivial nonnegative solution $(0, \theta)$ is globally stable; if $\mu > \mu^*$ there are exactly two coexistence states of (PP) (see Figures 1(c)–(f)). One is asymptotically stable, and the theoretical results and plenty of numerical analysis strongly suggest the other coexistence state is the most possibly unstable. Namely, if v is the better competitor, then it will eliminate u unless the effect of the inhibitor is sufficiently large, reflected by the condition $\mu > \mu^*$. This result exactly indicates that the inhibitor can help the genetically altered (plasmid-bearing) organism to avoid capture of the process by the plasmid-free organism.

(3) If μ is sufficiently large and b suitably large, then there exists a unique constant $a_\mu > \frac{\lambda_1}{1-q-k}$ such that (PP) has exactly two coexistence states for $a_\mu < a < \frac{\hat{\lambda}_1}{(1-q-k)}$: one asymptotically stable and the other (most possibly) unstable. Meanwhile, the semitrivial nonnegative solution $(0, \theta)$ is stable as well. But for $a \geq \frac{\hat{\lambda}_1}{(1-q-k)}$, (PP) has only a unique coexistence state, and it is asymptotically stable (see Figure 2). The simulations indicate that it is also globally stable, but we cannot give a rigorous proof. Furthermore, a_μ goes to $\frac{\lambda_1}{1-q-k}$ when $\mu \to \infty$, which is just consistent with our analytic outcomes.

(4) In fact, (PP) may also have two coexistence states in the case that μ is not large enough. For example, taking the parameters $\mu = 1, a = 6.4$, and $b = 5$ and the same parameters as above, (PP) has two positive solutions; see Figure 3. Moreover, the simulations also suggest that the coexistence solution in Figure 3(a) is asymptotically stable and the coexistence solution in Figure 3(b) is (most possibly) unstable.

(5) We discuss the difference between the plasmid model and the standard chemostat competition model in the presence of inhibitor. In (1), we mention the difference between the above two kinds of chemostat models when the plasmid-bearing organism is a better competitor. Here, we mainly concentrate on the case that the plasmid-bearing organism is a weaker competitor than the plasmid-free organism. It is easy to see that the introduction of the plasmid-free organism destroys the competitive property of the system. However, it is this property of the plasmid model that leads to the complex dynamical behavior. Now, numerical simulations help us understand this; see Figure 4. Take the parameters $a = 2.5, b = 5$, and $\mu = 50$ and the same parameters as before except that $q = 0, 0.01, 0.1, 0.2$ for Figures 4(a)–(h). Simulations convince us that when the effect of the inhibitor is very large, represented by large μ, if $q = 0$, that is, for the standard chemostat model with inhibitor, there is only one positive coexistence solution (see Figure 4(b)). Moreover, both the analytic results and many numerical simulations convince us that it is unstable. But once $q > 0$, the plasmid model has one asymptotically stable coexistence solution (see Figures 4(c), 4(e), 4(g)) and one (most likely) unstable coexistence solution (see Figures 4(d), 4(f), 4(h)).

Appendix A. In this section, we give the proof of Lemma 3.2.

Proof. (i) Let $y = (0, 0)$. By computation $W_y = \{(u, v) \in X : u \geq 0, v \geq 0\}$, $S_y = (0, 0)$. Hence $X_y = X$, and $Q = I$ (I is the identity operator in X). We first examine the eigenvalues of $\mathcal{A}'(0, 0)$, where $\mathcal{A}'(0, 0)$ is the Fréchet derivative of $\mathcal{A}$ with respect to (u, v) at $(0, 0)$. By direct computation,

$$\mathcal{A}'(0,0)(u,v) = \left(-d\frac{d^2}{dx^2} + M\right)^{-1} \times (a(1-q-k)uf_1(z) + Mu,\ bvf_2(z) + Mv + aquf_1(z))$$

for each $(u,v)\in X$. Hence an eigenvector (u,v) of $\mathcal{A}'(0,0)$ satisfies

$$\begin{aligned}&-du'' + Mu = \tfrac{1}{\lambda}(a(1-q-k)f_1(z)+M)u,\\ &-dv'' + Mv = \tfrac{1}{\lambda}((bf_2(z)+M)v + aquf_1(z)),\\ &u'(0)=u'(1)+\gamma u(1)=0, \qquad v'(0)=v'(1)+\gamma v(1)=0.\end{aligned}$$

It is easy to see that $I-\mathcal{A}'(0,0)$ is invertible in W_y since $a\neq \lambda_1/(1-q-k)$ and $b\neq\sigma_1$.

If $u\equiv 0$, then λ is an eigenvalue of

$$-dv'' + Mv = \frac{1}{\lambda}(bf_2(z)+M)v,\ v'(0)=v'(1)+\gamma v(1)=0. \tag{A.1}$$

Let η_1 be the principal eigenvalue of

$$-d\omega'' - bf_2(z)\omega = \eta_1\omega,\quad \omega'(0)=\omega'(1)+\gamma\omega(1)=0.$$

Then $\eta_1>0$ if $b<\sigma_1$, and $\eta_1<0$ if $b>\sigma_1$. It follows from Lemma 2.3 that if $b<\sigma_1$, then (A.1) has no eigenvalue larger than or equal to 1; if $b>\sigma_1$, then (A.1) has eigenvalues larger than 1. Namely, $\mathcal{A}'(0,0)$ has no eigenvalue larger than or equal to 1 with the corresponding eigenvector of the form $(0,v)$ if $b<\sigma_1$; $\mathcal{A}'(0,0)$ has eigenvalues larger than 1 with the corresponding eigenvector of the form $(0,v)$ if $b>\sigma_1$.

If $u\not\equiv 0$, then λ is an eigenvalue of

$$-du'' + Mu = \frac{1}{\lambda}(a(1-q-k)f_1(z)+M)u, \qquad u'(0)=u'(1)+\gamma u(1)=0.$$

By Lemma 2.3, we know that if $a<\lambda_1/(1-q-k)$, then $\mathcal{A}'(0,0)$ has no eigenvalue larger than or equal to 1 with the associated eigenfunction (u,v), where $u\not\equiv 0$; if $a>\lambda_1/(1-q-k)$, then $\mathcal{A}'(0,0)$ has eigenvalues larger than 1 with the associated eigenfunction $(u,v)(u\not\equiv 0)$. Hence, by Theorem 2.2 in [26], $\operatorname{index}_W(\mathcal{A},(0,0))=1$ if $a<\lambda_1/(1-q-k)$ and $b<\sigma_1$, and $\operatorname{index}_W(\mathcal{A},(0,0))=0$ if $a>\lambda_1/(1-q-k)$ or $b>\sigma_1$.

(ii) Let $y=(0,\theta)$. By computation,

$$W_y=\{(u,v)\in X: u\geq 0\},\ S_y=\{(0,v): v\in C_B([0,1])\}.$$

Define $X_y=\{(u,0): u\in C_B([0,1])\}$. Then $X=S_y\oplus X_y$ with projection Q given by $(u,v)\to(u,0)$. We first determine the existence of $\operatorname{index}_W(\mathcal{A},(0,\theta))$. Let $\mathcal{A}'(0,\theta)$ denote the Fréchet derivative of $\mathcal{A}$ with respect to (u,v) at $(0,\theta)$. Then

$$\mathcal{A}'(0,\theta)(u,v)=\left(\left(-d\frac{d^2}{dx^2}+M\right)^{-1}g(u,v),\ \left(-d\frac{d^2}{dx^2}+M\right)^{-1}(h_1(u,v)+h_2(u,v))\right)$$

for $(u,v)\in X$, where

$$\begin{aligned}&g(u,v)=(a(1-q-k)f_1(z-\theta)+M)u,\\ &h_1(u,v)=(-b(1+c)\theta f_2'(z-\theta)-b\mu c\theta f_2(z-\theta)+aqf_1(z-\theta))u,\\ &h_2(u,v)=(b(f_2(z-\theta)-\theta f_2'(z-\theta))+M)v.\end{aligned}$$

Let $(u,v)\in W_y$ be a fixed point of $\mathcal{A}'(0,\theta)$. Then (u,v) satisfies

$$\begin{aligned}&du'' + a(1-q-k)uf_1(z-\theta)=0,\\ &dv'' + b(f_2(z-\theta)-\theta f_2'(z-\theta))v\\ &\quad=(b(1+c)\theta f_2'(z-\theta)+b\mu c\theta f_2(z-\theta)-aqf_1(z-\theta))u.\end{aligned}$$

Clearly, if $a \neq \hat{\lambda}_1/(1-q-k)$, then $u \equiv v \equiv 0$. That is, $I - \mathcal{A}'(0,\theta)$ is invertible in W_y, and $\text{index}_W(\mathcal{A},(0,\theta))$ is well defined. Next, we consider the eigenvalues of $Q \circ \mathcal{A}'(0,\theta)$. By virtue of definition $Q(u,v) = (u,0)$, every eigenvector of $Q \circ \mathcal{A}'(0,\theta)$ has the form $(u,0)$, where u is a nonzero solution of the equation

$$-du'' + Mu = \frac{1}{\lambda}(a(1-q-k)f_1(z-\theta) + M)u, \qquad u'(0) = u'(1) + \gamma u(1) = 0.$$

Let η_1 be the first eigenvalue of

$$-d\omega'' - a(1-q-k)\omega f_1(z-\theta) = \eta_1\omega, \qquad \omega'(0) = \omega'(1) + \gamma\omega(1) = 0.$$

Then $\eta_1 > 0$ if $a < \hat{\lambda}_1/(1-q-k)$; $\eta_1 < 0$ if $a > \hat{\lambda}_1/(1-q-k)$. It follows from Lemma 2.3 that $Q \circ \mathcal{A}'(0,\theta)$ has no eigenvalue larger than or equal to 1 if $a < \hat{\lambda}_1/(1-q-k)$; $Q \circ \mathcal{A}'(0,\theta)$ has an eigenvalue larger than 1 if $a > \hat{\lambda}_1/(1-q-k)$. In view of Theorem 2.2 in [26], $\text{index}_W(\mathcal{A},(0,\theta)) = 0$ if $a > \hat{\lambda}_1/(1-q-k)$; $\text{index}_W(\mathcal{A},(0,\theta)) = \text{index}_{S_y}(\mathcal{A}'(0,\theta),(0,0)) = (-1)^\sigma$ if $a < \hat{\lambda}_1/(1-q-k)$. Here σ is the sum of multiplicities of the eigenvalues λ of $\mathcal{A}'(0,\theta)$ restricted in S_y such that $\lambda > 1$.

It remains to prove that $\text{index}_W(\mathcal{A},(0,\theta)) = 1$ for $a < \hat{\lambda}_1/(1-q-k)$. It suffices to show $\sigma = 0$. Suppose λ is an eigenvalue of $\mathcal{A}'(0,\theta)$ in S_y with the corresponding eigenvector (u,v). Then $u = 0$ and v is a nonzero solution of the equation

$$-dv'' + Mv = \frac{1}{\lambda}(b(f_2(z-\theta) - \theta f_2'(z-\theta)) + M)v, \qquad v'(0) = v'(1) + \gamma v(1) = 0. \tag{A.2}$$

It follows from Lemma 2.3 that (A.2) has no eigenvalue larger than or equal to 1, which implies $\sigma = 0$ and $\text{index}_W(\mathcal{A},(0,\theta)) = \text{index}_{S_y}(\mathcal{A}'(0,\theta),(0,0)) = 1$. The proof of this lemma is completed. □

Acknowledgments. The authors would like to give their thanks to Professor Sze-Bi Hsu who gave us some important references on this model. And the authors also would like to give their sincere thanks to the anonymous referees for their valuable suggestions leading to an improvement of the paper.

REFERENCES

[1] L. Chao and B. R. Levin, *Structured habitats and the evolution of anti-competitor toxins in bacteria*, Proc. Nat. Acad. Sci. USA, 75 (1981), pp. 6324–6328.
[2] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.
[3] M. G. Crandall and P. H. Rabinowitz, *Bifurcation, perturbation of simple eigenvalues and linearized stability*, Arch. Ration. Mech. Anal., 52 (1973), pp. 161–180.
[4] E. N. Dancer, *On the indices of fixed points of mappings in cones and applications*, J. Math. Anal. Appl., 91 (1983), pp. 131–151.
[5] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.
[6] L. Dung and H. L. Smith, *A parabolic system modelling microbial competition in an unmixed bio-reactor*, J. Differential Equations, 130 (1996), pp. 59–91.
[7] Y. Du, *Uniqueness, multiplicity and stability for positive solutions of a pair of reaction-diffusion equations*, Proc. Roy. Soc. Edinburgh Sect. A, 126A (1996), pp. 777–809.
[8] Y. Du and Y. Lou, *Some uniqueness and exact mulitiplicity results for a predator-prey model*, Trans. Amer. Math. Soc., 349 (1997), pp. 2443–2475.
[9] D. G. Figueiredo and J. P. Gossez, *Strict monotonicity of eigenvalues and unique continuation*, Comm. Partial Differential Equations, 17 (1992), pp. 339–346.
[10] A. G. Fredrickson and G. Stephanopoulos, *Microbial competition*, Science, 213 (1981), pp. 972–979.

[11] P. Hess, *Periodic Parabolic Boundary Value Problems and Positivity*, Longman Scientific and Technical, Harlow, UK, 1991.
[12] S. B. Hsu, T. K. Luo, and P. Waltman, *Competition between plasmid-bearing and plasmid-free organisms in a chemostat with an inhibitor*, J. Math. Biol., 34 (1995), pp. 225–238.
[13] S. B. Hsu and P. Waltman, *Analysis of a model of two competitors in a chemostat with an external inhibitor*, SIAM J. Appl. Math., 52 (1992), pp. 528–540.
[14] S. B. Hsu and P. Waltman, *On a system of reaction-diffusion equations arising from competition in an un-stirred chemostat*, SIAM J. Appl. Math., 53 (1993), pp. 1026–1044.
[15] S. B. Hsu and P. Waltman, *Competition between plasmid-bearing and plasmid-free organisms in selective media*, Chem. Eng. Sci., 52 (1997), pp. 23–35.
[16] S. B. Hsu and P. Waltman, *A survey of mathematical models of competition with an inhibitor*, Math. Biosci., 187 (2004), pp. 53–97.
[17] S. B. Hsu, P. Waltman, and G. S. K. Wolkowicz, *Global analysis of a model of plasmid-bearing, plasmid-free competition in the chemostat*, J. Math. Biol., 32 (1994), pp. 731–742.
[18] T. K. Luo and S. B. Hsu, *Global analysis of a model of plasmid-bearing and plasmid-free competition in a chemostat with inhibition*, J. Math. Biol., 34 (1995), pp. 41–76.
[19] J. Keener, *Principles of Applied Mathematics*, Addison–Wesley, Reading, MA, 1987.
[20] B. R. Levin, *Frequency-dependent selection in bacterial population*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 319 (1988), pp. 459–472.
[21] R. E. Lenski and S. Hattingh, *Coexistence of two competitors on one resource and one inhibitor: A chemostat model based on bacteria and antibiotics*, J. Theoret. Biol., 122 (1986), pp. 83–93.
[22] Z. Lu and K. P. Hadeler, *Model of plasmid-bearing, plasmid-free competition in the chemostat with nutrient recycling and an inhibitor*, Math. Biosci., 148 (1998), pp. 147–159.
[23] C. A. Macken, S. A. Levin, and R. Waldstätter, *The dynamics of bacteria-plasmid system*, J. Math. Biol., 32 (1994), pp. 123–145.
[24] H. Nie and J. Wu, *A system of reaction-diffusion equations in the unstirred chemostat with an inhibitor*, Internat. J. Bifur. Chaos, 16 (2006), pp. 989–1009.
[25] D. F. Ryder and D. DiBiasio, *An operational strategy for unstable recombinant DNA cultures*, Biotechnology and Bioengineering, 26 (1984), pp. 947–952.
[26] W. Ruan and W. Feng, *On the fixed point index and multiple steady-state solutions of reaction-diffusion systems*, Differential Integral Equations, 8 (1995), pp. 371–392.
[27] H. H. Schaefer, *Topological Vector Spaces*, Macmillan, New York, 1966.
[28] G. Stephanopoulos and G. Lapidus, *Chemostat dynamics of plasmid-bearing plasmid-free mixed recombinant cultures*, Chem. Eng. Sci., 43 (1988), pp. 49–57.
[29] H. L. Smith, *Monotone Dynamical Systems: An Introduction to The Theory of Competitive and Cooperative Systems*, Math. Surveys Monogr. 41, American Mathematical Society, Providence, RI, 1995.
[30] H. L. Smith and P. Waltman, *The Theory of the Chemostat*, Cambridge University Press, Cambridge, UK, 1995.
[31] M. X. Wang, *Nonlinear Parabolic Equation*, Science Press, Beijing, 1993 (in Chinese).
[32] J. Wu, H. Nie, and G. S. K. Wolkowicz, *A mathematical model of competition for two essential resources in the unstirred chemostat*, SIAM J. Appl. Math., 65 (2004), pp. 209–229.
[33] J. Wu, *Global bifurcation of coexistence state for the competition model in the chemostat*, Nonlinear Anal., 39 (2000), pp. 817–835.
[34] J. Wu and G. S. K. Wolkowicz, *A system of resource-based growth models with two resources in the un-stirred chemostat*, J. Differential Equations, 172 (2001), pp. 300–332.
[35] Q. Ye and Z. Li, *Introduction to Reaction-Diffusion Equations*, Science Press, Beijing, 1990.
[36] S. Zheng and J. Liu, *Coexistence solutions for a reaction-diffusion system of un-stirred chemostat model*, Appl. Math. Comput., 145 (2003), pp. 579–590.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1886–1911

SINGULAR LIMIT PROBLEM FOR SOME ELLIPTIC SYSTEMS*

YOSHIHITO OSHITA†

Abstract. For the sharp interface problem arising in the singular limit of some elliptic systems, we prove the existence and the nondegeneracy of solutions whose interface is a distorted circle in a two-dimensional bounded domain without any assumption on the symmetry of the domain.

Key words. sharp interface, singular perturbation, elliptic system

AMS subject classifications. 35B25, 35R35, 35P15

DOI. 10.1137/060656632

1. Introduction.

We consider

$$\begin{cases} -\Delta v = \mathbf{1}_{\Omega^+} - m & \text{in } \Omega, \\ \dfrac{\partial v}{\partial n} = 0 & \text{on } \partial\Omega, \\ \beta v + \kappa = 0 & \text{on } \Gamma, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with the smooth boundary $\partial\Omega$, $\partial/\partial n$ is the normal derivative on $\partial\Omega$, Ω^+ is an open set in Ω, $\Gamma = \partial\Omega^+ \subset \Omega$ is a C^2-curve embedded in Ω, κ is the curvature of Γ, $\beta > 0$ is a parameter, $m \in (0,1)$ is a constant, and $\mathbf{1}_{\Omega^+}$ denotes the characteristic function of Ω^+.

This problem comes from the following reaction-diffusion systems:

$$u_\tau = \epsilon^2 \Delta u + f(u) - \mu v, \qquad v_\tau = \Delta v + g(u,v), \tag{1.2}$$

where $u = u(y,\tau)$ and $v = v(y,\tau)$ are real-valued functions on $(y,\tau) \in \mathbb{R}^2 \times \mathbb{R}^+$; $\epsilon, \mu > 0$ are positive constants; $f \in C^1(\mathbb{R})$ is a function satisfying $f(i) = 0$, $f'(i) < 0$ $(i = 0, 1)$, $f(a) = 0$, $f'(a) > 0$ with $a \in (0,1)$, and $f(s) = -W'(s)$ with $W \in C^2(\mathbb{R})$ being a double-equal-well potential satisfying

$$W(0) = W(1) = 0 < W(s) \quad \forall s \in \mathbb{R}\backslash\{0,1\},$$
$$W''(0)W''(1) > 0;$$

and $g \in C^1(\mathbb{R}^2)$ is a function such that $g(1,0) = 1 - m > 0$, $g(-1,0) = -m < 0$. It follows $\int_0^1 f(s)\,ds = 0$.

A typical example of (f,g) is the FitzHugh–Nagumo type: $f(s) = s(s-a)(1-s)$, $g(u,v) = u - \gamma v - m$ ($\gamma \geq 0$ is a nonnegative constant). The general case is referred to as the activator-inhibitor system.

The system (1.2) describes the reaction and the diffusion phenomena of substances. When the ratio of the diffusion constants, ϵ^2, is extremely small, very interesting stationary patterns, such as stripes or spots, often appear. As a mathematical approach to understanding this pattern formation, we consider the limit $\epsilon \to 0$. Then

*Received by the editors April 7, 2006; accepted for publication (in revised form) November 27, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/sima/38-6/65663.html
†Department of Mathematics, Okayama University, Tsushima-Naka 3-1-1, Okayama 700-8530, Japan (oshita@math.okayama-u.ac.jp).

usually the domain is divided into two regions, and the remaining part becomes a thin layer. In some cases, the width of the internal transition layer approaches 0 in the limit, and the discontinuity surface inside the domain, which is called the sharp interface, appears. Recently very fine layered patterns of (1.2) have attracted a great deal of attention. See [4, 13, 14]. We consider this fine pattern which has the space scale of $\epsilon^{1/3}$ order. This is the unique scale that the order of the two driving forces of the sharp interface, the inhibitor v and the curvature of the sharp interface, balances. See [11]. This scale also appeared in [4]. After rescaling

$$x = \frac{y}{\epsilon^{1/3}}, \qquad t = \epsilon^{4/3}\tau, \qquad \varepsilon = \epsilon^{2/3},$$

we obtain

$$\begin{cases} u_t = \Delta u + \dfrac{1}{\varepsilon^2}(f(u) - \mu v), \\ \varepsilon^3 v_t = \Delta v + \varepsilon g(u, v). \end{cases} \tag{1.3}$$

We consider the stationary solutions of (1.3) subject to the homogeneous Neumann boundary condition:

$$\begin{cases} -\varepsilon^2 \Delta u = f(u) - \mu v & \text{in } \Omega, \\ -\Delta v = \varepsilon g(u, v) & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} = \dfrac{\partial v}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.4}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with the smooth boundary $\partial\Omega$.

The reduced equation in the singular limit $\varepsilon \to 0$ becomes

$$\begin{cases} -\Delta v_1 = \mathbf{1}_{\Omega^+} - m & \text{in } \Omega, \\ \dfrac{\partial v_1}{\partial n} = 0 & \text{on } \partial\Omega, \\ \frac{\mu}{\sigma} v_1 + \kappa = 0 & \text{on } \Gamma, \end{cases} \tag{RP}$$

where Ω^+ is an open set in Ω such that $\Gamma = \partial\Omega^+$ is a curve embedded in Ω and $\mathbf{1}_{\Omega^+}$ denotes the characteristic function of Ω^+. Here $u \to \mathbf{1}_{\Omega^+}$, $v/\varepsilon \to v_1$ as $\varepsilon \to 0$. See Appendix A of this paper or [12].

The associated functional becomes

$$J[\Gamma] = \sigma|\Gamma| + \frac{\mu}{2}\int_\Omega |\nabla v_1|^2 \, dx,$$

where $|\Gamma|$ is the length of Γ and v_1 is a solution to

$$\begin{cases} -\Delta v_1 = \mathbf{1}_{\Omega^+} - m & \text{in } \Omega, \\ \dfrac{\partial v_1}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

See Lemma 3.4 in section 3.

Essentially the same equation as (RP) was obtained in [12] by using the matched expansion method. Once you have a "nondegenerate" solution of (RP) in some sense, you can find a layered solution for the singular perturbation problem (1.4) with $g(u, v) = u - m$ and $\mu = 1$. See Appendix B.

For the reduction from the parabolic system to the sharp interface model, see [17]. In the case of the FitzHugh–Nagumo type, the functional $J[\Gamma]$ can also be obtained mathematically by using the notion of the Gamma convergence. See [14].

The direct method of calculus of variations implies the existence of global minimizers of $J[\Gamma]$. This gives a solution of (RP). However, it is usually difficult to know the profile and the nondegeneracy of the global minimizers. Here we consider the problem of finding a solution of (RP) that does not necessarily correspond to the global minimizers. The radially symmetric case for related problems is studied in [5, 6, 12, 15, 16, 18]. We do not assume any symmetry of the domain.

In order to state the result, we define the Green's function and its harmonic part.

DEFINITION 1.1. *For each $y \in \Omega$, let $G(x,y)$ be the solution to*

$$\begin{cases} -\Delta_x G(x,y) = \delta(x-y) - \dfrac{1}{|\Omega|}, & x \in \Omega, \\ \dfrac{\partial G}{\partial n_x}(x,y) = 0, & x \in \partial\Omega, \\ \displaystyle\int_\Omega G(x,y)\,dx = 0. \end{cases}$$

Set

$$G(x,y) = -\frac{1}{2\pi}\log|x-y| + \frac{|x-y|^2}{4|\Omega|} + H(x,y), \quad x,y \in \Omega.$$

Then it is known that $H(x,y)$ is symmetric and harmonic in both x and y. Let $\mathcal{H}(x) = H(x,x)$.

We define the following two conditions.

(A1) $0 \in \Omega$ is a strict local minimum point of $\mathcal{H}$. More precisely, there exists a neighborhood U of 0 in Ω such that $\mathcal{H}(0) < \mathcal{H}(x)$ for all $x \in U\backslash\{0\}$.

(A2) $0 \in \Omega$ is a nondegenerate critical point of $\mathcal{H}$.

Remark. When $\Omega = \{x \in \mathbb{R}^2\,;\, |x| < 1\}$, $x = 0$ is a unique minimum point of $\mathcal{H}$, and both (A1) and (A2) are satisfied. Indeed, we have $\mathcal{H}(x) = -\frac{1}{2\pi}\log(1-|x|^2) + \frac{|x|^2}{2\pi} + \mathcal{H}(0)$, and hence $\frac{\partial^2\mathcal{H}}{\partial x_i \partial x_j}(0) = \frac{2}{\pi}\delta_{ij}$.

The regular part of Green's function subject to the homogeneous Dirichlet boundary condition has a unique nondegenerate minimum point when $\Omega \subset \mathbb{R}^2$ is convex (see [2]). The regular part of Green's function subject to the homogeneous Neumann boundary condition is considered in [7].

We denote by d_{H} the Hausdorff metric

$$d_{\mathrm{H}}(K_1,K_2) = \max[\sup\{\mathrm{dist}(x,K_2)\,;\, x \in K_1\}, \sup\{\mathrm{dist}(y,K_1)\,;\, y \in K_2\}],$$

$S_r(0) = \{x \in \mathbb{R}\,;\, |x| = r\}$, and $B_r(0) = \{x \in \mathbb{R}\,;\, |x| < r\}$.

THEOREM 1.1. *Assume that* (A1) *or* (A2) *holds. If*

$$r_0 := \sqrt{\frac{m|\Omega|}{\pi}} < \mathrm{dist}(0,\partial\Omega),$$

then there exists a constant $\beta_0 > 0$ such that (1.1) *has a solution*

$$\begin{cases} \Gamma = \Gamma_\beta, \\ v = v_\beta, \\ \Omega^+ = \Omega^+_\beta \end{cases}$$

for all $\beta < \beta_0$ satisfying $d_{\mathrm{H}}(\Gamma_\beta, S_{r_0}(0)) \to 0$ as $\beta \to 0$.

This paper is organized as follows. In section 2, we prepare some preliminaries and notations. In section 3, we prove Theorem 1.1, and in section 4, we consider the linearized nondegeneracy of the problem.

2. Notation. We identify 2π-periodic functions on $\mathbb{R}$ with the functions on $S^1 = \{x \in \mathbb{R}^2\,;\, |x| = 1\} \cong \mathbb{R}/2\pi\mathbb{Z}$. For $q \in C^2(S^1)$, we use the following notation:

$$\dot{q}(\omega) = \frac{dq}{d\omega}(\omega) = \frac{d}{d\theta} q(\cos\theta, \sin\theta), \quad \omega = (\cos\theta, \sin\theta) \in S^1$$

and

$$\ddot{q}(\omega) = \frac{d^2q}{d\omega^2}(\omega) = \frac{d^2}{d\theta^2} q(\cos\theta, \sin\theta), \quad \omega = (\cos\theta, \sin\theta) \in S^1.$$

We set $X = C^2(S^1)$,

$$\|q\|_X = \max_{\omega\in S^1} |q(\omega)| + \max_{\omega\in S^1} |\dot{q}(\omega)| + \max_{\omega\in S^1} |\ddot{q}(\omega)|,$$

$Y = C(S^1)$, and

$$\|q\|_Y = \max_{\omega\in S^1} |q(\omega)|.$$

For $q_1, q_2 \in L^2(S^1)$, denote

$$\langle q_1, q_2\rangle = \int_{S^1} q_1(\omega) q_2(\omega)\, d\omega = \int_0^{2\pi} q_1(\cos\theta, \sin\theta) q_2(\cos\theta, \sin\theta)\, d\theta$$

and $\|q_1\|^2 = \langle q_1, q_1\rangle$.

Let $\Pi_{n^2} : L^2(S^1) \to L^2(S^1)$ denote the projections with respect to $\langle\cdot,\cdot\rangle$ onto $\operatorname{span}\{\cos i\theta, \sin i\theta\,;\, i = 0, 1, \dots, n\}$ for $n = 0, 1, \dots$. Let $\Pi_{n^2}^{\perp} = \mathbf{Id} - \Pi_{n^2}$.

Define $\Phi_0(\omega) = 1/\sqrt{2\pi}$, $\Phi_1(\omega) = \omega_1/\sqrt{\pi}$, and $\Phi_2(\omega) = \omega_2/\sqrt{\pi}$ for $\omega = (\omega_1, \omega_2) \in S^1$. Then $\Pi_0^{\perp}, \Pi_1^{\perp}$ are the projections onto the orthogonal complements of $\operatorname{span}\{\Phi_0\}$ and $\operatorname{span}\{\Phi_i\,;\, i = 0, 1, 2\}$, respectively.

3. Proof of Theorem 1.1. For brevity's sake, we assume that $r_0 = 1 < \operatorname{dist}(0, \partial\Omega)$.

For $\ell > 0$, define

$$X_\ell = \{q \in X\,;\, \|q\|_X \le \ell\}.$$

We can choose a constant $\delta \in (0, 1/2)$ such that $B_{1+\delta}(0) \subset \Omega$ by the assumption. For $q \in X_{\delta/2}$, define

$$\Gamma(q) = \{\sqrt{1 + q(\omega)}\,\omega\,;\, \omega \in S^1\},$$

$$\Omega^+(q) = \{r\omega\,;\, 0 \le r \le \sqrt{1 + q(\omega)}, \omega \in S^1\}.$$

Let $q \in X_{\delta/2} \cap \Pi_0^{\perp} X$. Then $\Gamma(q) \subset \Omega$, and $|\Omega^+(q)| = \pi$. Indeed since $\sqrt{1+q} \le 1 + \frac{1}{2}q \le 1 + \frac{\delta}{4}$, we have $\Gamma(q) \subset B_{1+\delta/2}(0) \subset \Omega$. In addition, since $\langle q, 1\rangle = 0$, we have

$$|\Omega^+(q)| = \int_{S^1} \int_0^{\sqrt{1+q(\omega)}} r\, dr d\omega = \int_{S^1} \frac{1 + q(\omega)}{2}\, d\omega = \pi.$$

Let M_β be the map from $X_{\delta/2}$ to Y defined by

$$M_\beta(q)(\omega) = K(q)(\omega) + \beta \int_{\Omega^+(q)} G(\sqrt{1+q(\omega)}\omega, y)\, dy, \qquad \omega \in S^1$$

for $q \in X_{\delta/2}$, where

$$K(q) = \frac{1 + q + \frac{3\dot{q}^2}{4(1+q)} - \frac{1}{2}\ddot{q}}{\left[1 + q + \frac{\dot{q}^2}{4(1+q)}\right]^{3/2}}$$

is the curvature of $\Gamma(q)$. Indeed, set $x_1(\theta) = r(\theta)\cos\theta$, $x_2(\theta) = r(\theta)\sin\theta$ with

$$r(\theta) = \sqrt{1 + q(\cos\theta, \sin\theta)}.$$

Then the curvature of $\Gamma(q)$ can be computed as follows.

$$\frac{\dot{x}_1\ddot{x}_2 - \ddot{x}_2\dot{x}_1}{(\dot{x}_1^2 + \dot{x}_2^2)^{3/2}} = \frac{r^2 + 2\dot{r}^2 - r\ddot{r}}{(r^2 + \dot{r}^2)^{3/2}} = \frac{1 + q + \frac{3\dot{q}^2}{4(1+q)} - \frac{1}{2}\ddot{q}}{\left[1 + q + \frac{\dot{q}^2}{4(1+q)}\right]^{3/2}}.$$

In order to solve (1.1), we need prove only the following.

PROPOSITION 3.1. *Suppose either* (A1) *or* (A2) *holds. If* $1 = \sqrt{m|\Omega|/\pi} < \operatorname{dist}(0, \partial\Omega)$, *then there exists a constant* $\beta_0 > 0$ *such that* $\Pi_0^\perp M_\beta(q) = 0$ *has a solution* $q = q_\beta \in X_{\delta/2} \cap \Pi_0^\perp X$ *for all* $\beta \in (0, \beta_0)$ *satisfying* $q_\beta \to 0$ *in* X *as* $\beta \to 0$.

Indeed, if $q \in X_{\delta/2} \cap \Pi_0^\perp X$ is a solution of $\Pi_0^\perp M_\beta(q) = 0$, then there exists a constant C_1 such that

$$M_\beta(q) \equiv C_1.$$

Now set

$$v(x) = \int_{\Omega^+(q)} G(x, y)\, dy - \frac{1}{\beta} C_1, \quad x \in \Omega.$$

Then v satisfies

$$\begin{cases} -\Delta v = \mathbf{1}_{\Omega^+(q)} - m & \text{in } \Omega, \\ \dfrac{\partial v}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

Hence we see that

$$\Gamma = \Gamma(q), \quad v(x) = \int_{\Omega^+(q)} G(x, y)\, dy - \frac{1}{\beta} C_1, \quad \Omega^+ = \Omega^+(q)$$

solves our (1.1) and completes the proof of Theorem 1.1.

3.1. Proof of Proposition 3.1. First we construct approximate solutions and prepare some lemmas. Next we find a solution by decomposing the equation $\Pi_0^\perp M_\beta(q) = 0$ into the system of equations:

1. $\Pi_1^\perp M_\beta(q) = 0$,
2. $(\Pi_1 - \Pi_0) M_\beta(q) = 0$.

Approximate solutions. We shall construct approximate solutions. Fix $\eta = (\eta_1, \eta_2) \in \mathbb{R}^2$ ($|\eta| < 1$). For $\omega \in S^1$, $r = \eta \cdot \omega + \sqrt{1 + (\eta \cdot \omega)^2 - |\eta|^2}$ is the solution of $|r\omega - \eta| = 1, r > 0$. Now set

$$q(\omega; \eta) := 2(\eta \cdot \omega)^2 - |\eta|^2 + 2(\eta \cdot \omega)\sqrt{1 + (\eta \cdot \omega)^2 - |\eta|^2} \quad \text{for } \omega \in S^1.$$

Then $\Gamma(q(\cdot; \eta))$ is a circle of radius 1 with center at η. Also it is easy to see that $\langle q(\cdot; \eta), \Phi_0 \rangle = 0$. Set

$$\xi_i = \langle q(\cdot; \eta), \Phi_i \rangle, \quad i = 1, 2. \tag{3.1}$$

Noting that $q(\omega; \eta) + |\eta|^2 - 2(\eta \cdot \omega)\sqrt{1 + q(\omega; \eta)} = 0$, it follows that

$$\xi_i = 2\sqrt{\pi} \sum_{j=1}^{2} \eta_j \langle \sqrt{1 + q(\omega; \eta)} \Phi_j, \Phi_i \rangle, \quad i = 1, 2.$$

Then we have $\langle q(\cdot; 0), \Phi_i \rangle = 0$, $i = 1, 2$, and $\partial \xi_i / \partial \eta_j(0) = 2\sqrt{\pi} \langle \Phi_j, \Phi_i \rangle = 2\sqrt{\pi} \delta_{ij}$. From the inverse mapping theorem, we conclude that there exists a neighborhood $U_1 \subset \mathbb{R}^2$ of 0 and a smooth map $P = (P_1, P_2)$ on U_1 such that $\xi_i = \langle q(\cdot; P(\xi)), \Phi_i \rangle$ for $\xi = (\xi_1, \xi_2) \in U_1$. Set $\psi_\xi = q(\cdot; P(\xi))$ for $\xi \in U_1$. Taking a smaller U_1 if necessary, we may assume the following.

LEMMA 3.1. *For any $\xi = (\xi_1, \xi_2) \in U_1$, there holds $\psi_\xi \in X_{\delta/4} \cap \Pi_0^\perp X$,*

$$\Pi_1 \psi_\xi = \xi_1 \Phi_1 + \xi_2 \Phi_2,$$

and $\Gamma(\psi_\xi) = \{\omega \sqrt{1 + \psi_\xi(\omega)}\,;\, \omega \in S^1\}$ is a circle of radius 1 with center at $P(\xi) \in \mathbb{R}^2$. Moreover P is smooth, $P(0) = 0$, and $\partial P_i / \partial \xi_j(0) = (2\sqrt{\pi})^{-1} \delta_{ij}$.

Since $K(\psi_\xi) \equiv 1$, we have $M_\beta(\psi_\xi) = 1 + O(\beta)$, and hence $\Pi_0 M_\beta(\psi_\xi) = \langle 1 + O(\beta), \Phi_0 \rangle \Phi_0 = 1 + O(\beta)$. Thus we have the following.

PROPOSITION 3.2. $\Pi_0^\perp M_\beta(\psi_\xi) = O(\beta)$ *as $\beta \to 0$ uniformly in $\xi \in U_1$.*

Linearization. For $t > -1$, $p \in \mathbb{R}$, and $s \in \mathbb{R}$, set

$$L(t, p, s) = \frac{1 + t + \frac{3p^2}{4(1+t)} - \frac{1}{2}s}{\left[1 + t + \frac{p^2}{4(1+t)}\right]^{3/2}}. \tag{3.2}$$

Then K is C^1 on $X_{\delta/2}$, and there hold

$$K'(q)\zeta = L_s(q, \dot{q}, \ddot{q})\ddot{\zeta} + L_p(q, \dot{q}, \ddot{q})\dot{\zeta} + L_t(q, \dot{q}, \ddot{q})\zeta \quad \text{for } \zeta \in X. \tag{3.3}$$

Moreover since

$$M_\beta(q)(\omega) = K(q)(\omega) + \beta \int_{S^1} \int_{-1}^{q(\hat{\omega})} G(\sqrt{1 + q(\omega)}\omega, \sqrt{1 + \hat{q}}\hat{\omega}) \frac{d\hat{q}}{2} d\hat{\omega}, \qquad \omega \in S^1$$

for $q \in X_{\delta/2}$, we see that M_β is also C^1 and

$$\begin{aligned}
[M'_\beta(q)\zeta](\omega) &= [K'(q)\zeta](\omega) + \frac{\beta}{2}\int_{S^1} G(\sqrt{1+q(\omega)}\omega, \sqrt{1+q(\hat\omega)}\hat\omega)\zeta(\hat\omega)\,d\hat\omega \\
&\quad + \beta\zeta(\omega)\int_{S^1}\int_{-1}^{q(\hat\omega)} \frac{\nabla_x G(\sqrt{1+q(\omega)}\omega, \sqrt{1+\hat q}\hat\omega)\cdot\omega}{2\sqrt{1+q(\omega)}}\,\frac{d\hat q}{2}\,d\hat\omega \\
&= \frac{d}{d\omega}[L_s(q,\dot q,\ddot q)\dot\zeta] + L_t(q,\dot q,\ddot q)\zeta \\
&\quad + \frac{\beta}{2}\int_{S^1} G(\sqrt{1+q(\omega)}\omega, \sqrt{1+q(\hat\omega)}\hat\omega)\zeta(\hat\omega)\,d\hat\omega \\
&\quad + \frac{\beta\zeta(\omega)}{2\sqrt{1+q(\omega)}}\int_{\Omega^+(q)} \omega\cdot\nabla_x G(\sqrt{1+q(\omega)}\omega, y)\,dy, \quad \omega \in S^1
\end{aligned} \tag{3.4}$$

for $q \in X_{\delta/2}$ and $\zeta \in X$.

LEMMA 3.2. *$\langle K'(q)\zeta_1, \zeta_2\rangle = \langle \zeta_1, K'(q)\zeta_2\rangle$, and $\langle M'_\beta(q)\zeta_1, \zeta_2\rangle = \langle \zeta_1, M'_\beta(q)\zeta_2\rangle$ for all $\zeta_1, \zeta_2 \in X$.*

Proof. Since

$$\begin{aligned}
L_s(q,\dot q,\ddot q) &= -\frac{1}{2}\left[1+q+\frac{\dot q^2}{4(1+q)}\right]^{-3/2}, \\
L_p(q,\dot q,\ddot q) &= \frac{3\dot q}{16}\left\{4 - \frac{\dot q^2}{(1+q)^2} + \frac{2\ddot q}{1+q}\right\}\left[1+q+\frac{\dot q^2}{4(1+q)}\right]^{-5/2},
\end{aligned}$$

we have $\frac{d}{d\omega}L_s(q,\dot q,\ddot q) = L_p(q,\dot q,\ddot q)$. Hence it follows that for $\zeta \in X$

$$\begin{aligned}
K'(q)\zeta &= L_s(q,\dot q,\ddot q)\ddot\zeta + L_p(q,\dot q,\ddot q)\dot\zeta + L_t(q,\dot q,\ddot q)\zeta \\
&= \frac{d}{d\omega}[L_s(q,\dot q,\ddot q)\dot\zeta] + L_t(q,\dot q,\ddot q)\zeta.
\end{aligned} \tag{3.5}$$

The assertion of the Lemma follows from (3.5) and (3.4). ∎

PROPOSITION 3.3. *There exists a constant $C > 0$, independent of small β, such that*

$$\|M'_\beta(q) - M'_\beta(\psi_\xi)\|_{\mathcal{L}(X,Y)} \le Ca^{1/3}$$

for any q in $\{q \in X_{\delta/2}\,;\, \|q-\psi_\xi\|_X \le a\}$, $a \le \delta/4$, and $\xi \in U_1$.

Proof. We shall show that there exists a constant $C > 0$, independent of small β, such that

$$\|M'_\beta(q)\zeta - M'_\beta(\psi_\xi)\zeta\|_Y \le Ca^{1/3}\|\zeta\|_X$$

for any q in $\{q \in X_{\delta/2}\,;\, \|q-\psi_\xi\|_X \le a\}$, $a \le \delta/4$, $\xi \in U_1$, and $\zeta \in X$.

Let $q \in X_{\delta/2}, \|q-\psi_\xi\|_X \le a$, $a \le \delta/4$, $\xi \in U_1$, and $\zeta \in X$. In this proof, for the sake of notational simplicity, we will use the same letters C to denote some positive constants, independent of q, ξ, a, ζ, whose values may vary from line to line.

From the expression (3.3) and the fact that L is smooth in $(-1,\infty)\times\mathbb{R}\times\mathbb{R}$, we easily see that

$$\|K'(q)\zeta - K'(\psi_\xi)\zeta\|_X \le Ca\|\zeta\|_Y. \tag{3.6}$$

Next let

$$v(x) = \int_{\Omega^+(q)} G(x,y)\,dy,$$
$$v_\xi(x) = \int_{\Omega^+(\psi_\xi)} G(x,y)\,dy.$$

Then since $\|\mathbf{1}_{\Omega^+(q)} - \mathbf{1}_{\Omega^+(\psi_\xi)}\|_{L^3(\Omega)} \le C\|q-\psi_\xi\|^{1/3}_{L^\infty(S^1)} \le Ca^{1/3}$, it follows from elliptic estimates and the Sobolev embedding theorem that $v, v_\xi \in W^{2,3}(\Omega) \subset C^{1,1/3}(\bar\Omega)$, $\|v - v_\xi\|_{C^1(\bar\Omega)} \le Ca^{1/3}$, and $\|v\|_{C^{1,1/3}(\bar\Omega)} < \infty$. Then since

$$\begin{aligned}
&\frac{\omega\cdot\nabla v(\sqrt{1+q(\omega)}\omega)}{\sqrt{1+q(\omega)}} - \frac{\omega\cdot\nabla v_\xi(\sqrt{1+\psi_\xi(\omega)}\omega)}{\sqrt{1+\psi_\xi(\omega)}}\\
&= \frac{\omega\cdot\nabla v(\sqrt{1+\psi_\xi(\omega)}\omega) - \omega\cdot\nabla v_\xi(\sqrt{1+\psi_\xi(\omega)}\omega)}{\sqrt{1+\psi_\xi(\omega)}}\\
&\quad + \frac{\omega\cdot\nabla v(\sqrt{1+\psi_\xi(\omega)}\omega)}{\sqrt{1+q(\omega)}} - \frac{\omega\cdot\nabla v(\sqrt{1+\psi_\xi(\omega)}\omega)}{\sqrt{1+\psi_\xi}}\\
&\quad + \frac{\omega\cdot\nabla v(\sqrt{1+q(\omega)}\omega) - \omega\cdot\nabla v(\sqrt{1+\psi_\xi(\omega)}\omega)}{|\sqrt{|1+q(\omega)} - \sqrt{1+\psi_\xi(\omega)}|^{1/3}} \frac{|\sqrt{|1+q(\omega)} - \sqrt{1+\psi_\xi(\omega)}|^{1/3}}{\sqrt{1+q(\omega)}},
\end{aligned}$$

we have

$$\sup_{\omega\in S^1}\left|\frac{\omega\cdot\nabla v(\sqrt{1+q(\omega)}\omega)}{\sqrt{1+q(\omega)}} - \frac{\omega\cdot\nabla v_\xi(\sqrt{1+\psi_\xi(\omega)}\omega)}{\sqrt{1+\psi_\xi(\omega)}}\right| \le Ca^{1/3}. \tag{3.7}$$

On the other hand let

$$w_\xi(x) = \int_{S^1} G\left(x, \sqrt{1+\psi_\xi(\hat\omega)}\hat\omega\right)\zeta(\hat\omega)\,d\hat\omega.$$

Then w_ξ is the solution to

$$\int_\Omega \nabla w_\xi\cdot\nabla\varphi\,dx = A_{\psi_\xi}\varphi := \int_{S^1}\zeta(\hat\omega)\varphi\left(\sqrt{1+\psi_\xi(\hat\omega)}\hat\omega\right)\,d\hat\omega - \frac{1}{|\Omega|}\int_{S^1}\zeta\int_\Omega\varphi$$

for all $\varphi\in W^{1,3/2}(\Omega)$. Here

$$A_{\psi_\xi} : \varphi \mapsto \int_{S^1}\zeta(\hat\omega)\varphi\left(\sqrt{1+\psi_\xi(\hat\omega)}\hat\omega\right)\,d\hat\omega - \frac{1}{|\Omega|}\int_{S^1}\zeta\int_\Omega\varphi$$

is a bounded linear functional on $W^{1,3/2}(\Omega)$. Indeed for $\varphi\in W^{1,3/2}(\Omega)$, we have

$$\begin{aligned}
&\left|\int_{S^1}\zeta(\hat\omega)\varphi\left(\sqrt{1+\psi_\xi(\hat\omega)}\hat\omega\right)\,d\hat\omega\right|\\
&\qquad\le \left(\int_{S^1}\left|\varphi\left(\sqrt{1+\psi_\xi(\hat\omega)}\hat\omega\right)\right|^{3/2}d\hat\omega\right)^{2/3}\left(\int_{S^1}|\zeta(\hat\omega)|^3\,d\hat\omega\right)^{1/3}\\
&\qquad\le (2\pi)^{1/3}\|\zeta\|_Y\|\tilde\varphi\|_{L^{3/2}(S^1)}\\
&\qquad\le C\|\zeta\|_Y\|\tilde\varphi\|_{W^{1,3/2}(B_1(0))} \le C\|\zeta\|_Y\|\varphi\|_{W^{1,3/2}(\Omega)},
\end{aligned}$$

where $\tilde{\varphi}(x) = \varphi\big(\sqrt{1+\psi_\xi(x/|x|)}\,x\big)$ and hence $A_{\psi_\xi} \in W^{-1,3}(\Omega)$ and $\|A_{\psi_\xi}\|_{W^{-1,3}(\Omega)} \le C\|\zeta\|_Y$. Then by elliptic estimates, we see that $w_\xi \in W^{1,3}(\Omega)$, and there exists a positive constant C, independent of $\xi \in U_1$, such that $\|w_\xi\|_{W^{1,3}(\Omega)} \le C\|\zeta\|_Y$.

Let

$$w(x) = \int_{S^1} G\left(x, \sqrt{1+q(\hat{\omega})}\hat{\omega}\right)\zeta(\hat{\omega})\,d\hat{\omega},$$
$$A_q : \varphi \mapsto \int_{S^1} \zeta(\hat{\omega})\varphi\left(\sqrt{1+q(\hat{\omega})}\hat{\omega}\right)\,d\hat{\omega}.$$

Then similarly we have $w \in W^{1,3}(\Omega)$ and $A_q \in W^{-1,3}(\Omega)$.

In addition, we estimate for $\varphi \in W^{1,3/2}(\Omega)$,

$$\begin{aligned}
|A_q\varphi - A_{\psi_\xi}\varphi| &= \left|\int_{S^1} \zeta(\hat{\omega})\varphi\left(\sqrt{1+q(\hat{\omega})}\hat{\omega}\right) - \zeta(\hat{\omega})\varphi\left(\sqrt{1+\psi_\xi(\hat{\omega})}\hat{\omega}\right)\,d\hat{\omega}\right| \\
&= \left|\int_{S^1}\int_{\sqrt{1+\psi_\xi(\hat{\omega})}}^{\sqrt{1+q(\hat{\omega})}} \varphi_r(r\hat{\omega})\zeta(\hat{\omega})\,dr\,d\hat{\omega}\right| \\
&\le \frac{\|\zeta\|_Y}{\sqrt{1-\delta/2}}\int_D |\varphi_r|\,dx \\
&\le \frac{\|\zeta\|_Y}{\sqrt{1-\delta/2}}\left(\int_D |\varphi_r|^{3/2}\,dx\right)^{2/3}\left(\int_D dx\right)^{1/3} \\
&\le \frac{(\pi a)^{1/3}}{\sqrt{1-\delta/2}}\|\zeta\|_Y\|\varphi\|_{W^{1,3/2}(\Omega)},
\end{aligned}$$

where

$$\begin{aligned}
D = \Bigg\{ & r\omega \in \Omega\,; \\
& \min\left(\sqrt{1+\psi_\xi(\omega)}, \sqrt{1+q(\omega)}\right) \le r \le \max\left(\sqrt{1+\psi_\xi(\omega)}, \sqrt{1+q(\omega)}\right), \omega \in S^1 \Bigg\}
\end{aligned}$$

and $\varphi_r(r\omega) = \omega\cdot\nabla\varphi(r\omega)$. We then have $\|A_q - A_{\psi_\xi}\|_{W^{-1,3}(\Omega)} \le Ca^{1/3}\|\zeta\|_Y$, and thus by elliptic estimates, $\|w - w_\xi\|_{L^\infty(\Omega)} \le Ca^{1/3}\|\zeta\|_Y$.

Therefore we have

$$\begin{aligned}
&\left|w\left(\sqrt{1+q(\omega)}\omega\right) - w_\xi\left(\sqrt{1+\psi_\xi(\omega)\omega}\right)\right| \\
&\le \left|w_\xi\left(\sqrt{1+\psi_\xi(\omega)\omega}\right) - w_\xi\left(\sqrt{1+q(\omega)}\omega\right)\right| \\
&\qquad + \left|w_\xi\left(\sqrt{1+q(\omega)}\omega\right) - w\left(\sqrt{1+q(\omega)}\omega\right)\right| \\
&\le Ca^{1/3}\|\zeta\|_Y
\end{aligned}$$

and hence

$$\sup_{\omega\in S^1}\left|w\left(\sqrt{1+q(\omega)}\omega\right) - w_\xi\left(\sqrt{1+\psi_\xi(\omega)\omega}\right)\right| \le Ca^{1/3}\|\zeta\|_Y. \tag{3.8}$$

Our assertion follows from (3.4), (3.6), (3.7), and (3.8). ∎

LEMMA 3.3. *For $\zeta \in X$,*

$$K'(0)\zeta = -\frac{1}{2}\left\{\ddot{\zeta} + \zeta\right\}.$$

Proof. From a simple computation, we have $L_t(0,0,0) = L_s(0,0,0) = -\frac{1}{2}$, $L_p(0,0,0) = 0$, and thus

$$K'(0)\zeta = L_s(0,0,0)\ddot{\zeta} + L_p(0,0,0)\dot{\zeta} + L_t(0,0,0)\zeta = -\frac{1}{2}\left\{\ddot{\zeta} + \zeta\right\}$$

for $\zeta \in X$. □

Note that (i) $\{\frac{1}{2}(n^2-1)\}_{n=0,1,2,\ldots}$ are the set of all the eigenvalues of $K'(0)$, (ii) the eigenspace associated with the eigenvalue $\frac{1}{2}(n^2-1)$ is $\operatorname{span}\{\cos n\theta, \sin n\theta\}$, and (iii) $K'(0)$ maps $\Pi_1^{\perp}X$ onto $\Pi_1^{\perp}Y$.

PROPOSITION 3.4. *There exists a constant C and a neighborhood $U_2 \subset U_1$ of 0, independent of small β, such that for all $\zeta \in \Pi_1^{\perp}X$ and $\xi \in U_2$,*

$$\|\Pi_1^{\perp} M'_{\beta}(\psi_{\xi})\zeta\|_Y \geq C\|\zeta\|_X.$$

Proof. Assume by contrary that there exist $\beta_n > 0$, $\zeta_n \in \Pi_1^{\perp}X$, and $\xi_n \in U_1$ such that $\|\zeta_n\|_X = 1$, $\lim_{n\to\infty}\beta_n = 0$, $\lim_{n\to\infty}\xi_n = 0$, and

$$\lim_{n\to\infty} \|\Pi_1^{\perp} M'_{\beta_n}(\psi_{\xi_n})\zeta_n\|_Y = 0.$$

We write ψ_n instead of ψ_{ξ_n} for simplicity. Taking a subsequence if necessary, we may assume that $\zeta_n \to \zeta$ in C^1 and $\psi_n \to 0$ in X. By Proposition 3.3,

$$\|M'_{\beta_n}(\psi_n) - M'_{\beta_n}(0)\| \leq C\|\psi_n\|_X^{1/3}$$

for n large. Hence

$$\|M'_{\beta_n}(\psi_n)\zeta_n - M'_{\beta_n}(0)\zeta_n\|_Y \leq C\|\psi_n\|_X^{1/3}.$$

Since

$$\Pi_1^{\perp} M'_{\beta_n}(0)\zeta_n = \Pi_1^{\perp} M'_{\beta_n}(\psi_n)\zeta_n - \Pi_1^{\perp}(M'_{\beta_n}(\psi_n)\zeta_n - M'_{\beta_n}(0)\zeta_n)$$

and

$$M'_{\beta}(0)\zeta = K'(0)\zeta + O(\beta)\zeta,$$

we have

$$\lim_{n\to\infty} \|\Pi_1^{\perp} K'(0)\zeta_n\|_Y = \lim_{n\to\infty} \|\Pi_1^{\perp} M'_{\beta_n}(\psi_{\xi_n})\zeta_n\|_Y = 0.$$

Then we have, for any $\varphi \in \Pi_1^{\perp}X$,

$$0 = \lim_{n\to\infty} \int_{S^1} \varphi \Pi_1^{\perp} K'(0)\zeta_n = \lim_{n\to\infty} \int_{S^1} \zeta_n \Pi_1^{\perp} K'(0)\varphi = \int_{S^1} \zeta \Pi_1^{\perp} K'(0)\varphi$$

since $\zeta_n \to \zeta$ in C^1. Since $K'(0)\varphi \in \Pi_1^{\perp}Y$ and $K'(0)\Pi_1^{\perp}X = \Pi_1^{\perp}Y$, we then have $\zeta = 0$. Thus we conclude that $\zeta_n \to 0$ in C^1, and it follows from $\|\zeta_n\|_X \leq \|\ddot{\zeta}_n + \zeta_n\|_Y + \|\dot{\zeta}_n\|_Y + 2\|\zeta_n\|_Y$ that

$$\frac{1}{2} = \frac{1}{2}\lim_{n\to\infty} \|\zeta_n\|_X \leq \lim_{n\to\infty} \|K'(0)\zeta_n\|_Y = \lim_{n\to\infty} \|\Pi_1^{\perp} K'(0)\zeta_n\|_Y = 0.$$

This is a contradiction and completes the proof. □

Energy functional.

DEFINITION 3.1. *For $q \in X_{\delta/2}$, define*

$$E_\beta[q] := \frac{1}{\beta}|\Gamma(q)| + \frac{1}{2}\int_\Omega |\nabla v|^2\,dx,$$

where

$$v(x) = \int_{\Omega^+(q)} G(x,y)\,dy, \quad x \in \Omega.$$

Note that

$$\begin{aligned}\int_\Omega |\nabla v|^2\,dx &= -\int_\Omega v\Delta v\,dx = \int_\Omega v(\mathbf{1}_{\Omega^+(q)} - m)\,dx \\ &= \int_{\Omega^+(q)}\int_{\Omega^+(q)} G(x,y)\,dxdy. \end{aligned} \tag{3.9}$$

LEMMA 3.4. *Let $T : \mathbb{I} \to X_{\delta/2}$ be a C^1-map from an open interval $\mathbb{I} \subset \mathbb{R}$ to $X_{\delta/2}$. Then*

$$\frac{d}{dt}E_\beta[T(t)] = \frac{1}{2\beta}\langle M_\beta(q), T'(t)\rangle, \tag{3.10}$$

$$\frac{d}{dt}|\Gamma(T(t))| = \frac{1}{2}\langle K(T(t)), T'(t)\rangle. \tag{3.11}$$

Proof. We write $q = T(t)$, $\zeta = T'(t)$, $R = \sqrt{1 + T(t)}$, and $\partial R/\partial t = \zeta/(2R)$. Then

$$|\Gamma(q)| = \int_{S^1} \sqrt{R^2 + \dot{R}^2}\,d\omega.$$

Since

$$\frac{R}{\sqrt{R^2+\dot{R}^2}} - \frac{d}{d\omega}\frac{\dot{R}}{\sqrt{R^2+\dot{R}^2}} = R\frac{R^2 + 2\dot{R}^2 - R\ddot{R}}{\left[R^2+\dot{R}^2\right]^{3/2}} = RK(q),$$

we have

$$\begin{aligned}\frac{d}{dt}|\Gamma(q)| &= \int_{S^1} RK(q)\frac{\partial R}{\partial t}\,d\omega \\ &= \frac{1}{2}\langle K(q), \zeta\rangle. \end{aligned} \tag{3.12}$$

Next let

$$v(x) = \int_{\Omega^+(q)} G(x,y)\,dy, \quad x \in \Omega.$$

Then from (3.9), we have

$$\int_\Omega |\nabla v|^2\,dx = \int_{S^1}\int_{S^1}\int_0^{R(\varphi)}\int_0^{R(\omega)} G(r\omega, s\varphi)rs\,drdsd\omega d\varphi,$$

and thus

$$\frac{d}{dt}\int_\Omega |\nabla v|^2\,dx = \int_{S^1}\int_{S^1}\left\{R(\varphi)\frac{\partial R}{\partial t}(\varphi)\int_0^{R(\omega)} G(r\omega,\varphi R(\varphi))r\,dr \right.$$
$$\left. + R(\omega)\frac{\partial R}{\partial t}(\omega)\int_0^{R(\varphi)} G(\omega R(\omega),s\varphi)s\,ds\right\}d\omega d\varphi$$
$$= \int_{S^1} v(R(\omega)\omega)\zeta\,d\omega. \tag{3.13}$$

Here we used the identities $G(x,y)=G(y,x)$ and

$$\frac{d}{dt}\int_0^{a(t)}\int_0^{b(t)} f(r,s)\,drds = a'(t)\int_0^{b(t)} f(r,a(t))\,dr + b'(t)\int_0^{a(t)} f(b(t),s)\,ds.$$

Hence (3.10) follows from (3.12) and (3.13). □

We apply the following Lemma used in [1, 8, 9, 10].

LEMMA 3.5. *Let X,Y be Banach spaces, $a>0$, $B_a = B_a(z) = \{z\in X\,;\,\|z-z_0\|\le a\}$. Suppose that F is a C^1-map of B_a into Y with $F'(z_0)$ invertible and satisfying, for some $0<\vartheta<1$,*

$$\|F'(z_0)^{-1}F(z_0)\| \le (1-\vartheta)a,$$
$$\|F'(z_0)^{-1}\|\,\|F'(z)-F'(z_0)\| \le \vartheta \quad \forall z\in B_a.$$

Then there is a unique solution in B_a of $F(z)=0$.

Completion of the proof. Define a map $F_\beta(\cdot) = \Pi_1^\perp M_\beta(\psi_\xi+\cdot)$ from $X_{\delta/4}\cap\Pi_1^\perp X$ to $\Pi_1^\perp Y$. Then we can take $a(\beta)>0$ such that $a(\beta)=O(\beta)$ as $\beta\to 0$ and $\|F'_\beta(\psi_\xi)^{-1}F_\beta(\psi_\xi)\|\le\frac{a(\beta)}{2}$. Then we see that for sufficiently small β,

$$\|F'_\beta(\psi_\xi)^{-1}\|\|F'_\beta(\psi_\xi+v)-F'_\beta(\psi_\xi)\| \le \frac{1}{2}$$

for all $v\in X_{a(\beta)}\cap\Pi_1^\perp X$. Then applying Lemma 3.5, we deduce that there exists a function $q^\perp_{\beta,\xi}\in X_{a(\beta)}\cap\Pi_1^\perp X$ solving

$$0 = F(q^\perp_{\beta,\xi}) = \Pi_1^\perp M_\beta(\psi_\xi + q^\perp_{\beta,\xi}). \tag{3.14}$$

Now $\Pi_0^\perp M_\beta(\psi_\xi+q^\perp_{\beta,\xi})=0$ if and only if $\langle M_\beta(\psi_\xi+q^\perp_{\beta,\xi}),\Phi_i\rangle=0$ for $i=1,2$. It follows from Lemma 3.4 and (3.14) that

$$\frac{\partial}{\partial\xi_i}E_\beta[\psi_\xi+q^\perp_{\beta,\xi}] = \frac{1}{2\beta}\langle M_\beta(\psi_\xi+q^\perp_{\beta,\xi}),\Phi_i\rangle,\quad i=1,2. \tag{3.15}$$

We have the following.

LEMMA 3.6. *There hold*

$$E_\beta[\psi_\xi+q^\perp_{\beta,\xi}] - \frac{2\pi}{\beta} = \frac{\pi}{16} + \frac{\pi^2}{8|\Omega|} + \frac{\pi^2}{2}\mathcal{H}(P(\xi)) + o(1), \tag{3.16}$$

and

$$\frac{\partial}{\partial\xi_i}E_\beta[\psi_\xi+q^\perp_{\beta,\xi}] = \frac{\pi^2}{2}\sum_{j=1}^2\frac{\partial\mathcal{H}}{\partial x_j}(P(\xi))\frac{\partial P_j}{\partial\xi_i}(\xi) + o(1),\quad i=1,2, \tag{3.17}$$

as $\beta\to 0$ uniformly in $\xi\in U_2$.

Proof. First we prove (3.16). It follows from (3.12) that

$$|\Gamma(\psi_\xi + q^\perp_{\beta,\xi})| = |\Gamma(\psi_\xi)| + \frac{1}{2}\langle 1, q^\perp_{\beta,\xi}\rangle + o(\|q^\perp_{\beta,\xi}\|_X) = 2\pi + o(\beta) \tag{3.18}$$

as $\beta \to 0$. In addition, we see that

$$\int_{\Omega^+(\psi_\xi + q^\perp_{\beta,\xi})}\int_{\Omega^+(\psi_\xi + q^\perp_{\beta,\xi})} G(x,y)\,dxdy = \int_{\Omega^+(\psi_\xi)}\int_{\Omega^+(\psi_\xi)} G(x,y)\,dxdy + o(1)$$

as $\beta \to 0$. Hence we need prove only

$$E_\beta[\psi_\xi] = \frac{2\pi}{\beta} + \frac{\pi}{16} + \frac{\pi^2}{8|\Omega|} + \frac{\pi^2}{2}\mathcal{H}(P(\xi)), \quad \xi \in U_2. \tag{3.19}$$

By the translation invariance, we have

$$\int_{\Omega^+(\psi_\xi)}\int_{\Omega^+(\psi_\xi)} \log|x-y|\,dxdy = \int_{\Omega^+(0)}\int_{\Omega^+(0)} \log|x-y|\,dxdy = -\frac{1}{4}\pi^2.$$

Similarly using the translation invariance and the fact that there holds

$$\int_{\Omega^+(0)} x \cdot y\,dx = 0$$

for any $y \in \mathbb{R}^2$, we have

$$\begin{aligned}\int_{\Omega^+(\psi_\xi)}\int_{\Omega^+(\psi_\xi)} |x-y|^2\,dxdy &= \int_{\Omega^+(0)}\int_{\Omega^+(0)} |x-y|^2\,dxdy \\ &= \int_{\Omega^+(0)}\int_{\Omega^+(0)} (|x|^2 + |y|^2)\,dxdy = \pi^2.\end{aligned}$$

In addition, by the mean value theorem for harmonic functions, we have

$$\int_{\Omega^+(\psi_\xi)}\int_{\Omega^+(\psi_\xi)} H(x,y)\,dxdy = \pi^2\mathcal{H}(P(\xi)).$$

Hence (3.19) follows and so does (3.16).

Next we prove (3.17). By Lemma 3.4 and (3.14), we see that

$$\begin{aligned}\frac{1}{\beta}\langle M_\beta(\psi_\xi + q^\perp_{\beta,\xi}), \Phi_i\rangle &= \frac{1}{\beta}\left\langle M_\beta(\psi_\xi + q^\perp_{\beta,\xi}), \frac{\partial\psi_\xi}{\partial\xi_i}\right\rangle \\ &= \frac{1}{\beta}\left\langle M_\beta(\psi_\xi), \frac{\partial\psi_\xi}{\partial\xi_i}\right\rangle + \frac{1}{\beta}\left\langle M'_\beta(\psi_\xi)q^\perp_{\beta,\xi}, \frac{\partial\psi_\xi}{\partial\xi_i}\right\rangle + o(1) \qquad (3.20)\end{aligned}$$

as $\beta \to 0$. By Lemma 3.4, we have

$$\frac{1}{\beta}\left\langle M_\beta(\psi_\xi), \frac{\partial\psi_\xi}{\partial\xi_i}\right\rangle = 2\frac{\partial}{\partial\xi_i}E_\beta[\psi_\xi]. \tag{3.21}$$

Noting that $K'(\psi_\xi)\frac{\partial\psi_\xi}{\partial\xi_i} \equiv 0$ since $K(\psi_\xi) \equiv 1$, it follows from Lemma 3.2 that

$$\begin{aligned}\frac{1}{\beta}\left\langle M'_\beta(\psi_\xi)q^\perp_{\beta,\xi}, \frac{\partial\psi_\xi}{\partial\xi_i}\right\rangle &= \frac{1}{\beta}\left\langle K'(\psi_\xi)q^\perp_{\beta,\xi}, \frac{\partial\psi_\xi}{\partial\xi_i}\right\rangle + O(\beta) \\ &= \frac{1}{\beta}\left\langle q^\perp_{\beta,\xi}, K'(\psi_\xi)\frac{\partial\psi_\xi}{\partial\xi_i}\right\rangle + O(\beta) \\ &= O(\beta) \qquad (3.22)\end{aligned}$$

as $\beta \to 0$. Thus (3.17) follows from (3.15), (3.19), (3.20), (3.21), and (3.22). ∎

Since

$$\begin{aligned}\frac{\partial}{\partial\xi_k}\sum_{j=1}^{2}\frac{\partial\mathcal{H}}{\partial x_j}(P(\xi))\frac{\partial P_j}{\partial\xi_i}(\xi)\bigg|_{\xi=0}&\\ &=\sum_{j,n=1}^{2}\frac{\partial^2\mathcal{H}}{\partial x_j\partial x_n}(0)\frac{\partial P_n}{\partial\xi_k}(0)\frac{\partial P_j}{\partial\xi_i}(0)+\sum_{j=1}^{2}\frac{\partial\mathcal{H}}{\partial x_j}(0)\frac{\partial P_j}{\partial\xi_i\partial\xi_k}(0)\\ &=\frac{1}{4\pi}\frac{\partial^2\mathcal{H}}{\partial x_i\partial x_k}(0),\end{aligned}$$

it follows from (3.15) and (3.17) that if $\frac{\partial^2\mathcal{H}}{\partial x_i\partial x_k}(0)$ is nondegenerate, then there exists a $\xi_\beta\in U_2$ such that $\langle M_\beta(\psi_{\xi_\beta}+q^{\perp}_{\beta,\xi_\beta}),\Phi_i\rangle=0$ $(i=1,2)$ and $\lim_{\beta\to0}\xi_\beta=0$. Thus $\psi_{\xi_\beta}+q^{\perp}_{\beta,\xi_\beta}$ is a solution of $\Pi_0^{\perp}M_\beta(\cdot)=0$.

If 0 is a strict local minimum point of $\mathcal{H}$, it follows from (3.15) and (3.16) that there exists a $\xi_\beta\in U_2$ such that $\langle M_\beta(\psi_{\xi_\beta}+q^{\perp}_{\beta,\xi_\beta}),\Phi_i\rangle=0$ $(i=1,2)$ and $\lim_{\beta\to0}\xi_\beta=0$. The proof is complete. □

4. Linearized nondegeneracy. In this section, we continue to use the notation defined in section 3. Throughout this section, we assume that there exists a compact subset $\mathcal{N}\subset\Omega$ satisfying $\operatorname{dist}(\mathcal{N},\partial\Omega)>1$. We linearize the equation around $\boldsymbol{P}+\Gamma(q)=\{\boldsymbol{P}+\sqrt{1+q(\omega)}\omega\,;\,\omega\in S^1\}$ for $\boldsymbol{P}\in\mathcal{N}$. Set

$$M_\beta(q\,;\boldsymbol{P})(\omega):=K(q)(\omega)+\beta\int_{\boldsymbol{P}+\Omega^+(q)}G(\boldsymbol{P}+\sqrt{1+q(\omega)}\omega,y)\,dy,\qquad\omega\in S^1$$

for $q\in X_{\delta/2}$, where $\boldsymbol{P}+\Omega^+(q)$ is the region surrounded by $\boldsymbol{P}+\Gamma(q)$.

THEOREM 4.1. *Suppose that*

(B1) *for every small $\beta>0$, there exist $\tilde q_\beta\in X$ and $\boldsymbol{P}\in\mathcal{N}$ such that*

$$(\Pi_4-\Pi_1)M_\beta(\tilde q_\beta\,;\boldsymbol{P})=0,$$

(B2) $\|\tilde q_\beta\|_X=O(\beta)$ *as* $\beta\to0$, *and*

(B3) *the Hessian matrix $\big(\frac{\partial^2\mathcal{H}}{\partial x_i\partial x_j}(\boldsymbol{P})\big)_{1\le i,j\le2}$ of $\mathcal{H}$ is nondegenerate for any $\boldsymbol{P}\in\mathcal{N}$.*

Then for sufficiently small β, $\mathcal{L}=\Pi_0^{\perp}M'_\beta(\tilde q_\beta\,;\boldsymbol{P})$ is nondegenerate in the sense that $\mathcal{L}\zeta=0$, $\int_{S^1}\zeta\,d\omega=0$ implies that $\zeta=0$.

Let $q_\beta=\psi_{\xi_\beta}+q^{\perp}_{\beta,\xi_\beta}$ be a solution obtained in Proposition 3.1. If you define $\tilde q_\beta$ such as $\Gamma(\psi_{\xi_\beta}+q^{\perp}_{\beta,\xi_\beta})=P(\xi_\beta)+\Gamma(\tilde q_\beta)$, then (B1) with $\boldsymbol{P}=P(\xi_\beta)$ and (B2) hold. Thus we have the following.

COROLLARY 4.1. *Suppose* (A2). *Then the solution obtained in Theorem* 1.1 *is nondegenerate in the sense of Theorem* 4.1.

Proof of Theorem 4.1. For brevity's sake, we write $q=\tilde q_\beta$. Set

$$\begin{aligned}B(\zeta,\zeta)=&\int_{S^1}[-L_s(q,\dot q,\ddot q)\dot\zeta^2+L_t(q,\dot q,\ddot q)\zeta^2]\,d\omega\\ &+\frac{\beta}{2}\int_{S^1}\int_{S^1}\zeta(\omega)G(\boldsymbol{P}+\sqrt{1+q(\omega)}\omega,\boldsymbol{P}+\sqrt{1+q(\hat\omega)}\hat\omega)\zeta(\hat\omega)\,d\omega d\hat\omega\\ &+\frac{\beta}{2}\int_{S^1}d\omega\frac{\zeta(\omega)^2}{\sqrt{1+q(\omega)}}\int_{\boldsymbol{P}+\Omega^+(q)}\omega\cdot\nabla_xG(\boldsymbol{P}+\sqrt{1+q(\omega)}\omega,y)\,dy\end{aligned}\tag{4.1}$$

for $\zeta \in H^1(S^1)$. We regard $\mathcal{L}$ as the operator on $\Pi_0^\perp H^2(S^1)$ satisfying $B(\zeta,\zeta) = \langle \mathcal{L}\zeta, \zeta\rangle$ for all $\zeta \in \Pi_0^\perp H^2(S^1)$.

LEMMA 4.1. *Suppose* (B2). *Let* $\lambda_1 \le \lambda_2 \le \lambda_3 \le \cdots$ *be the eigenvalues of* $\mathcal{L} : \Pi_0^\perp H^2(S^1) \to \Pi_0^\perp L^2(S^1)$ *and* $\zeta_i \in \Pi_0^\perp H^2(S^1)$ *be the normalized eigenfunctions associated with* λ_i. *Then*

$$\lambda_1 = \inf_{\zeta \in \Pi_0^\perp H^1(S^1), \|\zeta\|=1} B(\zeta,\zeta) = B(\zeta_1,\zeta_1) = O(\beta), \tag{4.2}$$

$$\lambda_2 = \inf_{\substack{\zeta \in \Pi_0^\perp H^1(S^1), \|\zeta\|=1 \\ \zeta \perp \zeta_1}} B(\zeta,\zeta) = B(\zeta_2,\zeta_2) = O(\beta), \tag{4.3}$$

$$\lambda_3 = \inf_{\substack{\zeta \in \Pi_0^\perp H^1(S^1), \|\zeta\|=1 \\ \zeta \perp \mathrm{span}\{\zeta_1,\zeta_2\}}} B(\zeta,\zeta) = B(\zeta_3,\zeta_3) = \frac{3}{2} + O(\beta). \tag{4.4}$$

Proof. The characterization of eigenvalues of $\mathcal{L}$ comes from the min-max principle of eigenvalues:

$$\lambda_i = \inf_{V_i \in \mathcal{S}_i} \max_{\substack{\zeta \in V_i \\ \|\zeta\|=1}} B(\zeta,\zeta), \quad i = 1,2,\ldots, \tag{4.5}$$

where $\mathcal{S}_i$ is the set of all i-dimensional subspace in $\Pi_0^\perp H^1(S^1)$. We need show only that $\lambda_i = \frac{1}{2}\left(\left[\frac{i+1}{2}\right]^2 - 1\right) + O(\beta)$ for $i = 1,2,\ldots$, where $[\frac{i+1}{2}]$ denotes the largest integer which is less than or equal to $\frac{i+1}{2}$.

We see that there exists a constant C such that $B(\zeta,\zeta) \ge \langle K'(0)\zeta,\zeta\rangle - C\beta\|\zeta\|^2 - C\beta\|\dot\zeta\|^2 \ge \frac{1}{2}(\|\dot\zeta\|^2 - \|\zeta\|^2) - C\beta\|\zeta\|^2 - C\beta\|\dot\zeta\|^2$ for all $\zeta \in H^1(S^1)$ and all small β.

For any $V_i \in \mathcal{S}_i$, let $V_{i+1} = V_i \oplus \mathrm{span}\{1\} \in \mathcal{T}_{i+1}$. Note that there holds

$$\left[\frac{i+1}{2}\right]^2 = \inf_{W_{i+1} \in \mathcal{T}_{i+1}} \max_{\substack{\zeta \in W_{i+1} \\ \|\zeta\|=1}} \left\|\dot\zeta\right\|^2,$$

where $\mathcal{T}_{i+1}$ is the set of all $(i+1)$-dimensional subspace in $H^1(S^1)$. It follows that $\max_{\zeta \in V_{i+1}} \|\dot\zeta\|^2/\|\zeta\|^2 \ge [\frac{i+1}{2}]^2$. This means that there exists a $\zeta^* \in V_{i+1}$ such that $\|\dot{\zeta^*}\|^2 \ge [\frac{i+1}{2}]^2$ and $\|\zeta^*\| = 1$. Setting $\zeta^* = \Pi_0\zeta^* + \zeta^\perp$, we have $0 < \|\zeta^\perp\| \le 1$ since $\|\dot{\zeta^\perp}\|^2 = \|\dot{\zeta^*}\|^2$ and $\|\zeta^*\|^2 = (\Pi_0\zeta^*)^2 + \|\zeta^\perp\|^2$. Then $\zeta^{**} = \frac{\zeta^\perp}{\|\zeta^\perp\|}$ satisfies $\zeta^{**} \in V_i$, $\|\zeta^{**}\| = 1$, and $\|\dot{\zeta^{**}}\|^2 \ge [\frac{i+1}{2}]^2$. Hence $\max_{\substack{\zeta \in V_i \\ \|\zeta\|=1}} B(\zeta,\zeta) \ge B(\zeta^{**},\zeta^{**}) \ge (\frac{1}{2} - C\beta)\|\dot{\zeta^{**}}\|^2 - (\frac{1}{2} + C\beta)\|\zeta^{**}\|^2 \ge ([\frac{i+1}{2}]^2 - 1)/2 - C([\frac{i+1}{2}]^2 + 1)\beta$. Hence we conclude that $\lambda_i \ge \frac{1}{2}\left([\frac{i+1}{2}]^2 - 1\right) + O(\beta)$.

On the other hand, let $V_1 = \mathrm{span}\{\cos\theta\}$ and

$$V_i = \begin{cases} \mathrm{span}\{\cos j\theta, \sin j\theta\,;\, j = 1,\ldots,n\} & \text{if } i = 2n \\ \mathrm{span}\{\cos j\theta, \sin j\theta, \cos(n+1)\theta\,;\, j = 1,\ldots,n\} & \text{if } i = 2n+1. \end{cases}$$

Then we see that $V_i \in \mathcal{S}_i$ and

$$\max_{\substack{\zeta \in V_i \\ \|\zeta\|=1}} \left\|\dot\zeta\right\|^2 = \left[\frac{i+1}{2}\right]^2.$$

Therefore we have

$$\begin{aligned} \lambda_i \le \max_{\substack{\zeta \in V_i \\ \|\zeta\|=1}} B(\zeta,\zeta) &\le \max_{\substack{\zeta \in V_i \\ \|\zeta\|=1}} \left\{ \frac{1}{2}\left(\left\|\dot\zeta\right\|^2 - \|\zeta\|^2\right) + C\beta\|\zeta\|^2 \right\} \\ &\le \frac{1}{2}\left(\left[\frac{i+1}{2}\right]^2 - 1\right) + O(\beta). \end{aligned}$$

The proof is complete. □

LEMMA 4.2.

1. *There hold* $L_{ts}(0,0,0) = L_{tt}(0,0,0) = L_{pp}(0,0,0) = \frac{3}{4}$ *and* $L_{ss}(0,0,0) = L_{ps}(0,0,0) = L_{tp}(0,0,0) = 0$ *(the definition of* $L(t,p,s)$ *is* (3.2)*).*
2. *There hold*

$$\int_{S^1} d\omega \Phi_j(\omega)\Phi_k(\omega)\omega \cdot \nabla_x H(\boldsymbol{P}+\omega,\boldsymbol{P}) = \frac{1}{2}\frac{\partial^2 H}{\partial x_j \partial x_k}(x,y)\Big|_{x=y=\boldsymbol{P}} \tag{4.6}$$

and

$$\int_{S^1}\int_{S^1} \Phi_j(\omega)H(\boldsymbol{P}+\omega,\boldsymbol{P}+\hat{\omega})\Phi_k(\hat{\omega})\, d\omega d\hat{\omega} = \pi\frac{\partial^2 H}{\partial x_j \partial y_k}(x,y)\Big|_{x=y=\boldsymbol{P}} \tag{4.7}$$

for each $j,k = 1,2$.

3. *Suppose* (B1) *and* (B2) *hold. Then*

$$\lim_{\beta\to 0}\frac{1}{\beta}\langle \dot{q}\Phi_k, \dot{\Phi}_j\rangle = -\frac{\pi}{3}\frac{\partial^2 H}{\partial x_j \partial x_k}(x,y)\Big|_{x=y=\boldsymbol{P}} \tag{4.8}$$

for each $j,k = 1,2$.

Proof. Part 1 follows from a simple computation.

By Green's theorem and the mean value theorem for harmonic functions,

$$\begin{aligned}\int_{S_r} ds_x \frac{\partial H}{\partial x_j}(\boldsymbol{P}+x,\boldsymbol{P})x_k &= r\int_{S_r} ds_x \frac{\partial H}{\partial x_j}(\boldsymbol{P}+x,\boldsymbol{P})\frac{x_k}{r}\\ &= r\int_{B_r}\frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P}+x,\boldsymbol{P})\, dx\\ &= \pi r^3 \frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P},\boldsymbol{P}),\end{aligned} \tag{4.9}$$

and hence

$$\begin{aligned}\int_{B_1}\frac{\partial H}{\partial x_j}(\boldsymbol{P}+x,\boldsymbol{P})x_k\, dx &= \int_0^1 dr\int_{S_r} ds_x \frac{\partial H}{\partial x_j}(\boldsymbol{P}+x,\boldsymbol{P})x_k\\ &= \int_0^1 \pi r^3\, dr \frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P},\boldsymbol{P})\\ &= \frac{\pi}{4}\frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P},\boldsymbol{P}).\end{aligned} \tag{4.10}$$

It then follows from Green's theorem that

$$\begin{aligned}&\pi\int_{S^1} d\omega \Phi_j(\omega)\Phi_k(\omega)\omega\cdot\nabla_x H(\boldsymbol{P}+\omega,\boldsymbol{P})\\ &\qquad = \sum_{i=1}^2 \int_{S^1}\omega_i\omega_j\omega_k \frac{\partial H}{\partial x_i}(\boldsymbol{P}+\omega,\boldsymbol{P})\, d\omega\\ &\qquad = \int_{B_1}\frac{\partial H}{\partial x_j}(\boldsymbol{P}+x,\boldsymbol{P})x_k + \frac{\partial H}{\partial x_k}(\boldsymbol{P}+x,\boldsymbol{P})x_j\, dx\\ &\qquad = \frac{\pi}{2}\frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P},\boldsymbol{P}),\end{aligned} \tag{4.11}$$

and thus (4.6) follows.

By Green's theorem and the mean value theorem, we have

$$\begin{aligned}
\pi \int_{S^1}\int_{S^1} \Phi_j(\omega) H(\boldsymbol{P}+\omega, \boldsymbol{P}+\hat{\omega})\Phi_k(\hat{\omega})\, d\omega d\hat{\omega} & \\
&= \int_{S^1}\int_{S^1} H(\boldsymbol{P}+x, \boldsymbol{P}+\hat{x}) x_j \hat{x}_k \, ds_x ds_{\hat{x}} \\
&= \int_{S^1}\int_{B_1} \frac{\partial H(\boldsymbol{P}+x, \boldsymbol{P}+\hat{x})}{\partial \hat{x}_k} x_j \, ds_x d\hat{x} \\
&= \pi \int_{S^1} x_j \frac{\partial H}{\partial \hat{x}_k}(\boldsymbol{P}+x, \boldsymbol{P}) ds_x \\
&= \pi \int_{B_1} \frac{\partial^2 H}{\partial x_j \partial \hat{x}_k}(\boldsymbol{P}+x, \boldsymbol{P}) dx \\
&= \pi^2 \frac{\partial^2 H}{\partial x_j \partial \hat{x}_k}(\boldsymbol{P}, \boldsymbol{P}). \qquad (4.12)
\end{aligned}$$

Thus (4.7) follows.

On the other hand,

$$\begin{aligned}
\langle \dot{q}\Phi_k, \dot{\Phi}_j\rangle &= -\langle q, \dot{\Phi}_k \dot{\Phi}_j\rangle - \langle q, \Phi_k \ddot{\Phi}_j\rangle \\
&= -\langle q, \dot{\Phi}_j \dot{\Phi}_k\rangle + \langle q, \Phi_j \Phi_k\rangle \\
&= \left\langle q, 2\Phi_j\Phi_k - \frac{\delta_{jk}}{\pi}\right\rangle. \qquad (4.13)
\end{aligned}$$

Set $\Phi_{jk} = 2\Phi_j\Phi_k - \frac{\delta_{jk}}{\pi}$. From Lemmas 3.3 and 3.2 we have

$$\langle q, \Phi_{jk}\rangle = \frac{2}{3}\langle q, K'(0)\Phi_{jk}\rangle = \frac{2}{3}\langle K'(0)q, \Phi_{jk}\rangle. \qquad (4.14)$$

Since $\langle M_\beta(q\,;\boldsymbol{P}), \Phi_{jk}\rangle = 0$ by (B1), we have $\langle M_\beta(0\,;\boldsymbol{P}), \Phi_{jk}\rangle + \langle M'_\beta(0\,;\boldsymbol{P})q, \Phi_{jk}\rangle = o(\beta)$. Noting (B2), we see that

$$\begin{aligned}
\lim_{\beta\to 0}\left\langle \frac{K'(0)q}{\beta}, \Phi_{jk}\right\rangle &= -\int_{S^1}\int_{B_1} G(\boldsymbol{P}+\omega, \boldsymbol{P}+y)\, dy \Phi_{jk}(\omega)\, d\omega \\
&= -\pi \int_{S^1} G(\boldsymbol{P}+\omega, \boldsymbol{P})\Phi_{jk}(\omega)\, d\omega \\
&= -\pi \int_{S^1} H(\boldsymbol{P}+\omega, \boldsymbol{P})\Phi_{jk}(\omega)\, d\omega. \qquad (4.15)
\end{aligned}$$

On the other hand, it follows from (4.10) and Green's theorem that

$$\begin{aligned}
\pi \int_{S^1} d\omega \Phi_j(\omega)\Phi_k(\omega) H(\boldsymbol{P}+\omega, \boldsymbol{P}) &= \int_{S^1} d\omega \omega_j \omega_k H(\boldsymbol{P}+\omega, \boldsymbol{P}) \\
&= \int_{B_1} \frac{\partial H}{\partial x_j}(\boldsymbol{P}+x, \boldsymbol{P}) x_k + \delta_{jk} H(\boldsymbol{P}+x, \boldsymbol{P})\, dx \\
&= \frac{\pi}{4}\frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P}, \boldsymbol{P}) + \delta_{jk}\pi H(\boldsymbol{P}, \boldsymbol{P}). \qquad (4.16)
\end{aligned}$$

Thus by (4.13), (4.15), (4.14), and (4.16), we conclude that

$$\lim_{\beta\to 0}\frac{1}{\beta}\langle\dot{q}\Phi_j,\dot{\Phi}_k\rangle = -\frac{2}{3}\pi\int_{S^1} H(\boldsymbol{P}+\omega,\boldsymbol{P})\left(2\Phi_j\Phi_k-\frac{\delta_{jk}}{\pi}\right)d\omega$$

$$= -\frac{\pi}{3}\frac{\partial^2 H}{\partial x_j\partial x_k}(\boldsymbol{P},\boldsymbol{P}). \tag{4.17}$$

The proof of (4.8) is complete. □

LEMMA 4.3. *Suppose* (B1) *and* (B2) *hold. Then there exists an orthogonal matrix* $(c_{ij})_{1\le i,j\le 2}$ *such that for* $i=1,2$, $\zeta_i^R=\zeta_i-(c_{1i}\Phi_1+c_{2i}\Phi_2)$ *satisfies* $\|\zeta_i^R\|^2=O(\beta)$ *as* $\beta\to 0$. *In addition, there holds*

$$\sum_{k=1}^{2}\frac{\pi}{4}\frac{\partial^2\mathcal{H}}{\partial x_j\partial x_k}(\boldsymbol{P})c_{ki} = o(1)+\frac{\lambda_i}{\beta}c_{ji} \tag{4.18}$$

for each $i,j=1,2$.

Proof. Set

$$\Phi_1 = d_{11}\zeta_1+d_{12}\zeta_2+\zeta_1^{\perp}, \tag{4.19}$$

$$\Phi_2 = d_{21}\zeta_1+d_{22}\zeta_2+\zeta_2^{\perp}, \tag{4.20}$$

where $\zeta_i^{\perp}\in\Pi_0^{\perp}H^2(S^1)$ and $\zeta_i^{\perp}\perp\operatorname{span}\{\zeta_1,\zeta_2\}$ for $i=1,2$. Then we have, for $i=1,2$,

$$\begin{aligned} O(\beta) &= B(\Phi_i,\Phi_i)\\ &= \langle\mathcal{L}\Phi_i,\Phi_i\rangle\\ &= d_{i1}^2\lambda_1+d_{i2}^2\lambda_2+\langle\mathcal{L}\zeta_i^{\perp},\zeta_i^{\perp}\rangle\\ &\ge O(\beta)+\|\zeta_i^{\perp}\|^2 \end{aligned} \tag{4.21}$$

and thus $\|\zeta_i^{\perp}\|^2=O(\beta)$. On the other hand, we have $1=d_{i1}^2+d_{i2}^2+\|\zeta_i^{\perp}\|^2$ and $0=d_{11}d_{21}+d_{12}d_{22}+\langle\zeta_1^{\perp},\zeta_2^{\perp}\rangle$. Hence we see that there exists an orthogonal matrix (c_{ij}) such that $d_{ij}=c_{ij}+O(\sqrt{\beta})$. Then since

$$\begin{aligned} \begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix}-\begin{pmatrix}c_{11}&c_{21}\\ c_{12}&c_{22}\end{pmatrix}\begin{pmatrix}\Phi_1\\ \Phi_2\end{pmatrix} &= \begin{pmatrix}c_{11}&c_{21}\\ c_{12}&c_{22}\end{pmatrix}\begin{pmatrix}c_{11}-d_{11}&c_{12}-d_{12}\\ c_{21}-d_{21}&c_{22}-d_{22}\end{pmatrix}\begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix}\\ &\quad -\begin{pmatrix}c_{11}&c_{21}\\ c_{12}&c_{22}\end{pmatrix}\begin{pmatrix}d_{10}\zeta_0+\zeta_1^{\perp}\\ d_{20}\zeta_0+\zeta_2^{\perp}\end{pmatrix}, \end{aligned} \tag{4.22}$$

$\zeta_i^R=\zeta_i-\sum_{k=1}^{2}c_{ki}\Phi_k$ satisfies the desired estimate. This completes the proof of the first part of Lemma 4.3.

Multiplying Φ_j to $\lambda_i\zeta_i=\mathcal{L}\zeta_i$ and integrating over S^1, we have $\lambda_i\langle\zeta_i,\Phi_j\rangle=\langle\mathcal{L}\zeta_i,\Phi_j\rangle$. Substituting $\zeta_i=\sum_{k=1}^{2}c_{ki}\Phi_k+\zeta_i^R$, we then have

$$\begin{aligned} \lambda_i c_{ji}+\lambda_i\langle\zeta_i^R,\Phi_j\rangle &= \sum_{k=1}^{2}c_{ki}\langle\mathcal{L}\Phi_k,\Phi_j\rangle+\langle\mathcal{L}\zeta_i^R,\Phi_j\rangle\\ &= \sum_{k=1}^{2}c_{ki}\langle\mathcal{L}\Phi_k,\Phi_j\rangle+\langle\zeta_i^R,\mathcal{L}\Phi_j\rangle, \end{aligned} \tag{4.23}$$

and hence $\sum_{k=1}^{2} c_{ki}\langle \mathcal{L}\Phi_k, \Phi_j\rangle = \lambda_i c_{ji} + o(\beta)$ by $\lambda_i = O(\beta)$, $\mathcal{L}\Phi_j = K'(0)\Phi_j + O(\beta) = O(\beta)$, and $\zeta_i^R = O(\sqrt{\beta})$. We need prove only that

$$\lim_{\beta\to 0} \frac{1}{\beta}\langle \mathcal{L}\Phi_k, \Phi_j\rangle = \frac{\pi}{4}\frac{\partial^2 \mathcal{H}}{\partial x_j \partial x_k}(\boldsymbol{P}). \tag{4.24}$$

Since $K''(0)(q_1, q_2) = \frac{3}{4}(\ddot{q}_1 q_2 + \dot{q}_1\dot{q}_2 + q_1 q_2 + q_1\ddot{q}_2)$ by Lemma 4.2 part 1 and $K'(0)\Phi_k = -\frac{1}{2}\{\ddot{\Phi}_k + \Phi_k\} = 0$, integrating by parts yields

$$\begin{aligned}
\langle K'(q)\Phi_k, \Phi_j\rangle &= \langle K'(0)\Phi_k, \Phi_j\rangle + \langle K''(0)(q, \Phi_k), \Phi_j\rangle + o(\beta) \\
&= \frac{3}{4}\langle \ddot{q}\Phi_k + \dot{q}\dot{\Phi}_k + q\Phi_k + q\ddot{\Phi}_k, \Phi_j\rangle + o(\beta) \\
&= -\frac{3}{4}\langle \dot{q}\Phi_k, \dot{\Phi}_j\rangle + o(\beta). \qquad (4.25)
\end{aligned}$$

Hence by (4.8), we have

$$\lim_{\beta\to 0} \frac{1}{\beta}\langle K'(q)\Phi_k, \Phi_j\rangle = \frac{\pi}{4}\frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P}, \boldsymbol{P}). \tag{4.26}$$

On the other hand, we have

$$\begin{aligned}
&\lim_{\beta\to 0}\int_{S^1}\int_{S^1} \Phi_j(\omega) G(\boldsymbol{P} + \sqrt{1+q(\omega)}\omega, \boldsymbol{P} + \sqrt{1+q(\hat{\omega})}\hat{\omega})\Phi_k(\hat{\omega})\, d\omega d\hat{\omega} \\
&\qquad = \int_{S^1}\int_{S^1} \Phi_j(\omega) G(\boldsymbol{P} + \omega, \boldsymbol{P} + \hat{\omega})\Phi_k(\hat{\omega})\, d\omega d\hat{\omega} \\
&\qquad = \frac{1}{2}\delta_{jk} - \frac{\pi}{2|\Omega|}\delta_{jk} + \pi\frac{\partial^2 H}{\partial x_j \partial y_k}(\boldsymbol{P}, \boldsymbol{P}). \qquad (4.27)
\end{aligned}$$

Here we used (4.7),

$$-\frac{1}{2\pi}\int_{S^1} \log|\omega - \hat{\omega}|\Phi_j(\hat{\omega})\, d\hat{\omega} = \frac{1}{2}\Phi_j(\omega), \tag{4.28}$$

and

$$\begin{aligned}
\int_{S^1} \frac{|\omega - \hat{\omega}|^2}{4}\Phi_j(\hat{\omega})\, d\hat{\omega} &= \int_{B_1} \frac{y_j - \omega_j}{2\sqrt{\pi}}\, dy \\
&= -\frac{\pi}{2}\Phi_j(\omega). \qquad (4.29)
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\lim_{\beta\to 0}\int_{S^1} \Phi_j(\omega)\frac{\Phi_k(\omega)}{\sqrt{1+q(\omega)}}\int_{\Omega^+(q)} \omega\cdot\nabla_x G(\boldsymbol{P} + \sqrt{1+q(\omega)}\omega, \boldsymbol{P} + y)\, dy \\
&\qquad = \int_{S^1} d\omega \Phi_j(\omega)\Phi_k(\omega)\int_{B_1} \omega\cdot\nabla_x G(\boldsymbol{P} + \omega, \boldsymbol{P} + y)\, dy \\
&\qquad = -\frac{1}{2}\delta_{jk} + \frac{\pi}{2|\Omega|}\delta_{jk} + \frac{\pi}{2}\frac{\partial^2 H}{\partial x_j \partial x_k}(\boldsymbol{P}, \boldsymbol{P}). \qquad (4.30)
\end{aligned}$$

Here we used

$$\begin{aligned}\int_{S^1} d\omega\Phi_j(\omega)\Phi_k(\omega)\int_{B_1}\omega\cdot\nabla_x H(\boldsymbol{P}+\omega,\boldsymbol{P}+y)\,dy&\\ =\pi\int_{S^1} d\omega\Phi_j(\omega)\Phi_k(\omega)\omega\cdot\nabla_x H(\boldsymbol{P}+\omega,\boldsymbol{P})&\\ =\frac{\pi}{2}\frac{\partial^2 H}{\partial x_j\partial x_k}(\boldsymbol{P},\boldsymbol{P})\qquad\text{by (4.6)},&\end{aligned}$$

$$-\frac{1}{2\pi}\nabla_x\int_{B_1}\log|x-y|\,dy=-\frac{x}{2},$$

and

$$\nabla_x\int_{B_1}\frac{|x-y|^2}{4}\,dy=\frac{\pi}{2}x.$$

Consequently by (4.26), (4.27), and (4.30), we have

$$\lim_{\beta\to 0}\frac{1}{\beta}\langle\mathcal{L}\Phi_k,\Phi_j\rangle=\frac{\pi}{2}\frac{\partial^2 H}{\partial x_j\partial y_k}(\boldsymbol{P},\boldsymbol{P})+\frac{\pi}{2}\frac{\partial^2 H}{\partial x_j\partial x_k}(\boldsymbol{P},\boldsymbol{P})=\frac{\pi}{4}\frac{\partial^2\mathcal{H}}{\partial x_j\partial x_k}(\boldsymbol{P}),\tag{4.31}$$

and the proof is complete. □

Completion of the proof of Theorem 4.1. Assume by contrary that there exists a sequence ζ_β such that $\mathcal{L}\zeta_\beta=0$, $\|\zeta_\beta\|=1$, and $\int_{S^1}\zeta_\beta\,d\omega=0$. This means that ζ_β is an eigenfunction of $\mathcal{L}$ associated with the eigenvalue 0. We see that for sufficiently small β, either λ_1 or λ_2 is equal to 0. Then by Lemma 4.3, we have $\zeta_\beta=c_1\Phi_1+c_2\Phi_2+\zeta^R$ such that $(c_1,c_2)\in S^1$ and $\|\zeta^R\|^2=O(\beta)$, and

$$\sum_{k=1}^{2}\frac{\partial^2\mathcal{H}}{\partial x_j\partial x_k}(\boldsymbol{P})c_k=o(1)\quad\text{for } j=1,2$$

as $\beta\to 0$. Taking a subsequence if necessary, we may assume that $(c_1,c_2)\to(\hat{c}_1,\hat{c}_2)\in S^1$ and

$$\sum_{k=1}^{2}\frac{\partial^2\mathcal{H}}{\partial x_j\partial x_k}(\boldsymbol{P})\hat{c}_k=0\quad\text{for } j=1,2.$$

It follows from (B3) that $\hat{c}_1=\hat{c}_2=0$. This is a contradiction and completes the proof.

Appendix A. Formal derivation of reduced problem. We shall formally deduce the reduced problem. If we assume $u\to u_0$ and $v\to v_0$ in the limit $\varepsilon\to 0$, we have

$$f(u_0)=\mu v_0,\quad \Delta v_0=0\quad\text{in }\Omega,$$

$$\frac{\partial v_0}{\partial n}=0\quad\text{on }\partial\Omega.$$

Hence v_0 is a constant. Now assume that v_0 is close to 0 and

$$u_0=f_1^{-1}(\mu v_0)\mathbf{1}_{\Omega^+}+f_0^{-1}(\mu v_0)\mathbf{1}_{\Omega^-},$$

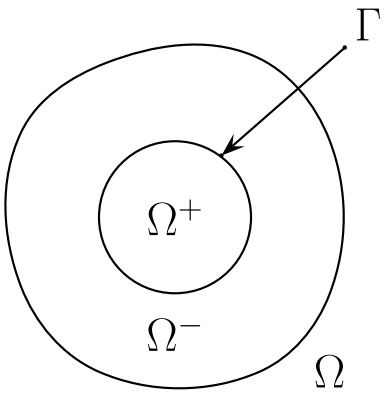


FIG. 1. *Sharp interface* Γ *and the domain* Ω.

where Ω^+, Ω^- are mutually disjoint open sets in Ω such that $\Gamma = \Omega \backslash (\Omega^+ \cup \Omega^-)$ is a curve embedded in Ω, $\mathbf{1}_{\Omega^\pm}$ denote the characteristic functions of $\Omega^\pm$, and $u = f_i^{-1}(v)$ are the inverse functions of $v = f(u)$ near $u = i$ $(i = 0, 1)$, respectively. We call Γ the sharp interface (see Figure 1). We shall identify the profile of u near Γ.

It is known that there exists a constant $\tau > 0$, depending on f, such that for any $v \in (-\tau, \tau)$, the equation for u, $u_t = u_{xx} + f(u) - v$, has a traveling wave solution $u(x,t) = Q(x - ct\,; v)$ with the speed $c = c(v)$ and the profile $Q = Q(\xi; v)$. More precisely, $c(v)$ and $Q(\xi; v)$ for $v \in (-\tau, \tau), \xi \in \mathbb{R}$ satisfy

$$\begin{cases} \ddot{Q} + c(v)\dot{Q} + f(Q) - v = 0 \quad \text{in } \mathbb{R}, \\ \lim_{\xi \to -\infty} Q(\xi; v) = f_1^{-1}(v), \\ \lim_{\xi \to +\infty} Q(\xi; v) = f_0^{-1}(v), \\ c(0) = 0. \end{cases}$$

Here $\dot{}$ means $d/d\xi$. See, for example, [3]. Near the sharp interface Γ, consider the function

$$u(x) = Q\left(\frac{d(x)}{\varepsilon}; v\right),$$

where $d = d(x)$ is the signed distance function from Γ such that $d(x) > 0$ if $x \in \Omega^-$ and $d(x) < 0$ if $x \in \Omega^+$. If the above function satisfies the first equation of (1.4) for each prescribed v, noting that $|\nabla d| = 1$, there holds

$$\ddot{Q} + \varepsilon(\Delta d)\dot{Q} + f(Q) - \mu v = 0.$$

Since Δd is equal to the curvature κ of Γ on the interface Γ (here we choose the sign such that $\kappa > 0$ when Ω^+ is a disk), it follows that

$$c(\mu v) = \varepsilon\kappa \qquad \text{on } \Gamma.$$

Since $c(0) = 0$ by the assumption, we may assume that $v_0 = 0$ and $u_0 = \mathbf{1}_{\Omega^+}$.

Next we consider the higher order term. Assume

$$v = \varepsilon v_1 + O(\varepsilon^2).$$

Then we obtain the reduced problem

$$\text{(A.1)}\qquad \begin{cases} -\Delta v_1 = g(u_0,0) = \mathbf{1}_{\Omega^+} - m & \text{in } \Omega, \\ \dfrac{\partial v_1}{\partial n} = 0 & \text{on } \partial\Omega, \\ \mu c'(0) v_1 = \kappa & \text{on } \Gamma. \end{cases}$$

Since it is easily seen that $c'(0) = -\frac{1}{\sigma} < 0$ with

$$\sigma = \int_0^1 \sqrt{2W(s)}\, ds = \int_{-\infty}^{\infty} [\dot{Q}(\xi;0)]^2\, d\xi,$$

(A.1) becomes (RP) with $\beta = \mu/\sigma$.

Appendix B. Layered solutions for elliptic systems. In this section, we will show that (1.4) with $g(u,v) = u - m$, $\mu = 1$ has a layered solution from Theorem 1.2 of [12].

The equation $v = f(u)$ has three solutions $f_0^{-1}(v) < f_a^{-1}(v) < f_1^{-1}(v)$ for small $|v|$, where $u = f_i^{-1}(v)$ is the inverse function of $v = f(u)$ near $u = i$ $(i = 0, a, 1)$. We then have $f_u(f_i^{-1}(v)) < 0$ $(i = 0, 1)$, and $J(0) = 0$, $J'(0) = -1$, where $J(v) := \int_{f_0^{-1}(v)}^{f_1^{-1}(v)} [f(s) - v]\, ds$. Thus the assumptions (A1)–(A3) of [12] are satisfied.

We use the notation in sections 2 and 3. Under the assumption (A2), we consider the solution $q \in X_{\delta/2}$ of $\Pi_0^{\perp} M_\beta(q) = 0$ obtained in Proposition 3.1. For brevity's sake, we assume that (B1) holds with $\boldsymbol{P} = 0$. Then, $\Pi_0^{\perp} M'_\beta(q)\zeta = 0$ and $\zeta \in \Pi_0^{\perp} X$ implies that $\zeta = 0$.

Let $(\Gamma(q), v, \Omega^+(q))$ be a corresponding solution of (1.1). We write $\Gamma^* = \Gamma(q)$, $\Omega^+ = \Omega^+(q)$, and $\Omega^- = \Omega \backslash (\Omega^+ \cup \Gamma^*)$.

Set $r = r(\theta) = \sqrt{1 + q(\theta)}$. Let

$$\mathbf{e}_1 = \left(\frac{\dot{r}\cos\theta - r\sin\theta}{\sqrt{r^2 + \dot{r}^2}}, \frac{\dot{r}\sin\theta + r\cos\theta}{\sqrt{r^2 + \dot{r}^2}} \right)$$

be a unit tangent vector on Γ^* and

$$\mathbf{e}_2 = \left(-\frac{\dot{r}\sin\theta + r\cos\theta}{\sqrt{r^2 + \dot{r}^2}}, \frac{\dot{r}\cos\theta - r\sin\theta}{\sqrt{r^2 + \dot{r}^2}} \right)$$

be the unit normal vector on Γ^* which points the interior of Ω^+.

Set $V^+ = v|_{\Omega^+}$ and $V^- = v|_{\Omega^-}$. Then $-\Delta V^+ = 1 - m$ in Ω^+, $\beta V^+ + \kappa = 0$ on Γ^*, $-\Delta V^- = -m$ in Ω^-, $\beta V^- + \kappa = 0$ on Γ^*, and $\partial V^-/\partial n = 0$ on $\partial\Omega$. Moreover $\partial V^+/\partial \mathbf{e}_2 = \partial V^-/\partial \mathbf{e}_2$ on Γ^* since $v \in W^{2,\infty}(\Omega) \subset C^1(\Omega)$. Thus (A4) of [12] is satisfied.

By a bootstrap argument, q is smooth and so is Γ^*. Fix $\alpha \in (0,1)$. For any $I \in C^{2,\alpha}(\Gamma^*)$, let $w^{\pm}$ be the unique solution to

$$\begin{cases} \Delta w^+ = 0 & \text{in } \Omega^+, \\ w^+ = I & \text{on } \Gamma^* \end{cases}$$

and

$$\begin{cases} \Delta w^- = 0 & \text{in } \Omega^-, \\ \frac{\partial w^-}{\partial n} = 0 & \text{on } \partial\Omega, \\ w^- = I & \text{on } \Gamma^*, \end{cases}$$

respectively.

Let $Z(x) = \frac{\partial w^-}{\partial \mathbf{e}_2}(x) - \frac{\partial w^+}{\partial \mathbf{e}_2}(x)$ for $x \in \Gamma^*$. Then it follows from Green's theorem that $w^+(x) = \int_{\Gamma^*} G(x,y)Z(y)\,ds_y + \frac{1}{|\Omega|}(\int_{\Omega^+} w^+\,dx + \int_{\Omega^-} w^-\,dx)$ for $x \in \Omega^+$, $w^-(x) = \int_{\Gamma^*} G(x,y)Z(y)\,ds_y + \frac{1}{|\Omega|}(\int_{\Omega^+} w^+\,dx + \int_{\Omega^-} w^-\,dx)$ for $x \in \Omega^-$, and $I(x) = \int_{\Gamma^*} G(x,y)Z(y)\,ds_y + \frac{1}{|\Omega|}(\int_{\Omega^+} w^+\,dx + \int_{\Omega^-} w^-\,dx)$ for $x \in \Gamma^*$. Moreover we see that $\int_{\Omega^+} \nabla w^+ \cdot \nabla\varphi\,dx + \int_{\Omega^-} \nabla w^- \cdot \nabla\varphi\,dx = \int_{\Gamma^*} Z\varphi\,ds$ for all $\varphi \in C^\infty(\bar\Omega)$. Let $\mathcal{S} : C^{2,\alpha}(\Gamma^*) \to C^{1,\alpha}(\Gamma^*)$ be the linear map defined by $\mathcal{S}(I) = Z$. Then $\mathcal{S}$ has one dimensional kernel which is spanned by $I \equiv \text{const}$, and $\mathrm{Im}(\mathcal{S}) = \{Z \in C^{1,\alpha}(\Gamma^*)\,;\, P(Z) = 0\}$, where P is a projection defined by $P(Z) = \frac{1}{|\Gamma^*|}\int_{\Gamma^*} Z\,ds$ for $Z \in C^\alpha(\Gamma^*)$.

On the other hand, for $Z \in C^{1,\alpha}(\Gamma^*)$, define $w(x) = \int_{\Gamma^*} G(x,y)Z(y)\,ds_y$ for $x \in \Omega$. Then we see that $\Delta w(x) = \frac{1}{|\Omega|}\int_{\Gamma^*} Z\,ds$ for $x \in \Omega^\pm$. Moreover there holds $\int_\Omega \nabla w \cdot \nabla\varphi\,dx = \int_{\Gamma^*} Z\varphi\,ds - \frac{1}{|\Omega|}\int_{\Gamma^*} Z\,ds\int_\Omega \varphi\,dx$ for all $\varphi \in C^\infty(\bar\Omega)$. Let $\mathcal{T} : C^{1,\alpha}(\Gamma^*) \to C^{2,\alpha}(\Gamma^*)$ be the linear operator defined by $\mathcal{T}(Z) = w|_{\Gamma^*}$. Then $\mathcal{S}(\mathbf{Id} - P)\mathcal{T} = \mathbf{Id}$ on $\{Z \in C^{1,\alpha}(\Gamma^*)\,;\, P(Z) = 0\}$, and $(\mathbf{Id} - P)\mathcal{T}\mathcal{S} = \mathbf{Id}$ on $\{I \in C^{2,\alpha}(\Gamma^*)\,;\, P(I) = 0\}$.

Let

$$\ell = \int_0^\theta \sqrt{r^2 + \dot r^2}\,d\theta$$

be the arc-length parameter of Γ^*. We change the variable θ into ℓ and regard functions of ℓ as functions on Γ^*.

PROPOSITION B.1. *For any $\zeta \in \Pi_0^\perp X$,*

$$M'_\beta(q)\zeta = \tilde{\mathcal{L}}Z := \frac{d^2Z}{d\ell^2} + \kappa^2 Z + \beta(\nabla v \cdot \mathbf{e}_2)Z - \beta\mathcal{T}(Z),$$

where

$$Z(\ell) = -\frac{\zeta(\theta)}{2\sqrt{r^2 + \dot r^2}}.$$

Proof. For small $|\tau|$, set

$$Q(\theta,\tau) = q(\theta) + \tau\zeta(\theta), \quad R(\theta,\tau) = \sqrt{1 + q(\theta) + \tau\zeta(\theta)}.$$

Let $\Theta(\tau)$ be the continuous function satisfying that $R(\Theta(\tau),\tau)(\cos\Theta(\tau), \sin\Theta(\tau))$ is on the normal line of $\Gamma(q)$ at $r(\theta)(\cos\theta, \sin\theta)$, and that $\Theta(0) = \theta$. Then we have

$$\left.\frac{d}{d\tau}\right|_{\tau=0} R(\Theta(\tau),\tau)(\cos\Theta(\tau), \sin\Theta(\tau)) \cdot \mathbf{e}_1 = 0,$$

which implies that

$$\frac{d\Theta}{d\tau}(0) = \frac{\dot r}{r\sqrt{r^2 + \dot r^2}} Z(\ell)$$

and

$$\left.\frac{d}{d\tau}\right|_{\tau=0} R(\Theta(\tau),\tau)(\cos\Theta(\tau), \sin\Theta(\tau)) = Z\mathbf{e}_2.$$

Hence Z is the normal velocity of the evolution of curves $\tau \mapsto \Gamma(Q(\cdot,\tau))$.

We see that

$$M'_\beta(q)\zeta = L_s(q,\dot q,\ddot q)\ddot\zeta + L_p(q,\dot q,\ddot q)\dot\zeta + L_t(q,\dot q,\ddot q)\zeta + \frac{\beta}{2}\frac{\nabla v\cdot\omega}{r}\zeta + \frac{\beta}{2}\int_{S^1} G(x,r(\hat\omega)\hat\omega)\zeta(\hat\omega)\,d\hat\omega\Big|_{x=r(\cos\theta,\sin\theta)}.$$

First we have

$$\begin{aligned}\nabla v\cdot\omega &= (\omega\cdot\mathbf{e}_1)(\nabla v\cdot\mathbf{e}_1) + (\omega\cdot\mathbf{e}_2)(\nabla v\cdot\mathbf{e}_2)\\ &= \frac{\dot r}{\sqrt{r^2+\dot r^2}}(\nabla v\cdot\mathbf{e}_1) - \frac{r}{\sqrt{r^2+\dot r^2}}(\nabla v\cdot\mathbf{e}_2),\end{aligned}$$

$$\nabla v\cdot\mathbf{e}_1 = \frac{\partial v}{\partial \ell} = -\frac{1}{\beta}\frac{d\kappa}{d\ell} = -\frac{1}{\beta}\frac{d\kappa}{d\theta}\frac{1}{\sqrt{r^2+\dot r^2}},$$

and thus

$$\begin{aligned}\frac{\beta}{2}\frac{\nabla v\cdot\omega}{r}\zeta &= \frac{\beta}{2}\frac{\dot r}{r\sqrt{r^2+\dot r^2}}(\nabla v\cdot\mathbf{e}_1)\zeta - \frac{\beta}{2}\frac{\nabla v\cdot\mathbf{e}_2}{\sqrt{r^2+\dot r^2}}\zeta\\ &= \frac{\dot r}{r\sqrt{r^2+\dot r^2}}\frac{d\kappa}{d\theta}Z - \frac{\beta}{2}\frac{\nabla v\cdot\mathbf{e}_2}{\sqrt{r^2+\dot r^2}}\zeta\\ &= \frac{\dot r}{r\sqrt{r^2+\dot r^2}}\frac{d\kappa}{d\theta}Z + \beta(\nabla v\cdot\mathbf{e}_2)Z.\end{aligned}$$

Hence

$$M'_\beta(q)\zeta = I_1 + I_2 + I_3,$$

where

$$\begin{aligned}I_1 &:= L_s\ddot\zeta + L_p\dot\zeta + L_t\zeta + \frac{\dot r}{r\sqrt{r^2+\dot r^2}}\frac{d\kappa}{d\theta}Z,\\ I_2 &:= \beta(\nabla v\cdot\mathbf{e}_2)Z,\\ I_3 &:= \frac{\beta}{2}\int_{S^1} G(x,r(\hat\omega)\hat\omega)\zeta(\hat\omega)\,d\hat\omega\Big|_{x=r(\cos\theta,\sin\theta)}.\end{aligned}$$

Then by change of variable, we have

$$\frac{\beta}{2}\int_{S^1} G(x,r(\hat\omega)\hat\omega)\zeta(\hat\omega)\,d\hat\omega = -\beta\int_{\Gamma^*} G(x,y)Z(y)\,ds_y$$

for $x\in\Gamma^*$, and thus $I_3 = -\beta\mathcal{T}(Z)$.

Next using

$$\begin{aligned}L_s\ddot\zeta + L_p\dot\zeta + L_t\zeta &= \frac{\partial}{\partial\tau}\Big|_{\tau=0} L(q+\tau\zeta,\dot q+\tau\dot\zeta,\ddot q+\tau\ddot\zeta)\\ &= \frac{\partial}{\partial\tau}\Big|_{\tau=0}\frac{R^2+2\dot R^2 - R\ddot R}{(R^2+\dot R^2)^{3/2}}\end{aligned}$$

and

$$\left.\frac{\partial R}{\partial \tau}\right|_{\tau=0} = -\frac{\sqrt{r^2+\dot r^2}}{r} Z,$$

we obtain

$$I_1 = L_s\ddot\zeta + L_p\dot\zeta + L_t\zeta + \frac{\dot r}{r\sqrt{r^2+\dot r^2}}\frac{d\kappa}{d\theta} Z = \frac{d^2 Z}{d\ell^2} + \kappa^2 Z \tag{B.1}$$

after lengthy computations. The claim of Proposition B.1 now follows from (B.1) and $I_3 = -\beta\mathcal{T}(Z)$. Note that

$$\begin{aligned}
&\left.\frac{d}{d\tau}\right|_{\tau=0} L(q(\Theta(\tau)) + \tau\zeta(\Theta(\tau)), \dot q(\Theta(\tau)) + \tau\dot\zeta(\Theta(\tau)), \ddot q(\Theta(\tau)) + \tau\ddot\zeta(\Theta(\tau)))\\
&\quad = L_s\ddot\zeta + L_p\dot\zeta + L_t\zeta + (L_s\dddot q + L_p\ddot q + L_t\dot q)\frac{d\Theta}{d\tau}(0)\\
&\quad = L_s\ddot\zeta + L_p\dot\zeta + L_t\zeta + \frac{d}{d\theta}L(q(\theta), \dot q(\theta), \ddot q(\theta))\frac{d\Theta}{d\tau}(0)\\
&\quad = L_s\ddot\zeta + L_p\dot\zeta + L_t\zeta + \frac{\dot r}{r\sqrt{r^2+\dot r^2}}\frac{d\kappa}{d\theta} Z.
\end{aligned}$$

Hence the left-hand side of (B.1) is the derivative of the curvature at the point along the normal direction. □

PROPOSITION B.2. *There exists a constant $C > 0$ such that*

$$\|(\mathbf{Id} - P)\tilde{\mathcal{L}} Z\|_{C^\alpha} \ge C\|Z\|_{C^{2,\alpha}}$$

for any $Z \in C^{2,\alpha}(\Gamma^*)$ with $\int_{\Gamma^*} Z\, ds = 0$.

Proof. Suppose not. Then there exists a sequence $Z_n \in C^{2,\alpha}(\Gamma^*)$ with $\int_{\Gamma^*} Z_n\, ds = 0$, $\|Z_n\|_{C^{2,\alpha}} = 1$, and $\|(\mathbf{Id} - P)\tilde{\mathcal{L}} Z\|_{C^\alpha} \to 0$. Then taking a subsequence if necessary, we may assume that $Z_n \to Z$ in C^1. Then

$$\frac{d^2 Z_n}{d\ell^2} \to (\mathbf{Id} - P)(-\kappa^2 Z - \beta(\nabla v \cdot \mathbf{e}_2)Z + \beta\mathcal{T}(Z))$$

in C^α and thus uniformly. Hence $Z_n \to Z$ in $C^{2,\alpha}$ and $(\mathbf{Id} - P)\tilde{\mathcal{L}} Z = 0$. Setting $\zeta(\theta) = -2\sqrt{r^2+\dot r^2} Z(\ell)$, it follows from Proposition B.1 that $\Pi_0^\perp M'_\beta(q)\zeta = 0$. Then by the nondegeneracy, we have $\zeta = 0$ and thus $Z = 0$. It follows that $0 = \|Z\|_{C^{2,\alpha}} = \lim_{n\to\infty}\|Z_n\|_{C^{2,\alpha}} = 1$. This is a contradiction, and the proof is complete. □

It follows from Proposition B.2 that the linear operator $(\mathbf{Id} - P)\tilde{\mathcal{L}}$ has a bounded inverse from $\{Z \in C^\alpha(\Gamma^*)\,;\, P(Z) = 0\}$ to $\{Z \in C^{2,\alpha}(\Gamma^*)\,;\, P(Z) = 0\}$. Therefore (A5) of [12] is satisfied and Theorem 1.2 of [12] can be applied for (1.4) with $g(u, v) = u - m$ and $\mu = 1$.

REFERENCES

[1] J. BYEON AND Y. OSHITA, *Existence of Multi-bump standing waves with a critical frequency for nonlinear Schrödinger equations*, Comm. Partial Differential Equations, 29 (2004), pp. 1877–1904.

[2] L. A. CAFFARELLI AND A. FRIEDMAN, *Convexity of solutions of semilinear elliptic equations*, Duke Math. J., 52 (1985), pp. 431–456.

[3] X. Chen, D. Hilhorst, and E. Logak, *Asymptotic behavior of solutions of an Allen–Cahn equation with a nonlocal term*, Nonlinear Anal., 28 (1997), pp. 1283–1298.
[4] X. Chen and Y. Oshita, *Periodicity and uniqueness of global minimizers of an energy functional containing a long-range interaction*, SIAM J. Math. Anal., 37 (2006), pp. 1299–1332.
[5] X. Chen and M. Taniguchi, *Instability of spherical interfaces in a nonlinear free boundary problem*, Adv. Differential Equations, 5 (2000), pp. 747–772.
[6] M. A. del Pino, *Radially symmetric internal layers in a semilinear elliptic system*, Trans. Amer. Math. Soc., 347 (1995), pp. 4807–4837.
[7] T. Kolokolnikov, M. Titcombe, and M. Ward, *Optimizing the fundamental Neumann eigenvalue for the Laplacian in a domain with small traps*, European J. Appl. Math., 16 (2005), pp. 161–200.
[8] Y. Y. Li, *On a singularly perturbed elliptic equation*, Adv. Differential Equations, 2 (1997), pp. 955–980.
[9] Y. Y. Li, *On a singularly perturbed equation with Neumann boundary condition*, Comm. Partial Differential Equations, 23 (1998), pp. 487–545.
[10] Y. Y. Li and L. Nirenberg, *The Dirichlet problem for singularly perturbed elliptic equations*, Comm. Pure Appl. Math., 51 (1998), pp. 1445–1490.
[11] Y. Nishiura and H. Suzuki, *Nonexistence of higher-dimensional stable Turing patterns in the singular limit*, SIAM J. Math. Anal., 29 (1998), pp. 1087–1105.
[12] Y. Nishiura and H. Suzuki, *Higher dimensional SLEP equation and applications to morphological stability in polymer problems*, SIAM J. Math. Anal., 36 (2004/05), pp. 916–966.
[13] Y. Oshita, *On stable nonconstant stationary solutions and mesoscopic patterns for FitzHugh–Nagumo equations in higher dimensions*, J. Differential Equations, 188 (2003), pp. 110–134.
[14] Y. Oshita, *Stable stationary patterns and interfaces arising in reaction-diffusion systems*, SIAM J. Math. Anal., 36 (2004), pp. 479–497.
[15] K. Sakamoto and H. Suzuki, *Spherically symmetric internal layers for activator-inhibitor systems.* I. *Existence by a Lyapunov-Schmidt reduction*, J. Differential Equations, 204 (2004), pp. 56–92.
[16] K. Sakamoto and H. Suzuki, *Spherically symmetric internal layers for activator-inhibitor systems.* II. *Stability and symmetry breaking bifurcations*, J. Differential Equations, 204 (2004), pp. 93–122.
[17] P. Soravia and P. E. Souganidis, *Phase-field theory for FitzHugh-Nagumo-type systems*, SIAM J. Math. Anal., 27 (1996), pp. 1341–1359.
[18] M. Taniguchi, *Multiple existence and linear stability of equilibrium balls in a nonlinear free boundary problem*, Quart. Appl. Math., 58 (2000), pp. 283–302.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1912–1931

STRONG INSTABILITY OF STANDING WAVES FOR THE NONLINEAR KLEIN–GORDON EQUATION AND THE KLEIN–GORDON–ZAKHAROV SYSTEM*

MASAHITO OHTA† AND GROZDENA TODOROVA‡

Abstract. The orbital instability of ground state standing waves $e^{i\omega t}\phi_\omega(x)$ for the nonlinear Klein–Gordon equation has been known in the domain of all frequencies ω for the supercritical case and for frequencies strictly less than a critical frequency ω_c in the subcritical case. We prove the strong instability of ground state standing waves for the entire domain above. For the case when the frequency is equal to the critical frequency ω_c we prove strong instability for all radially symmetric standing waves $e^{i\omega_c t}\varphi(x)$. We prove similar strong instability results for the Klein–Gordon–Zakharov system.

Key words. nonlinear Klein–Gordon equations, standing waves, instability, Klein–Gordon–Zakharov system

AMS subject classifications. 35L70, 35B35, 35A15

DOI. 10.1137/050643015

1. Introduction and main results. We study the strong instability of standing wave solutions $e^{i\omega t}\varphi(x)$ for the nonlinear Klein–Gordon equation of the form

$$\partial_t^2 u - \Delta u + u = |u|^{p-1}u, \quad (t,x) \in \mathbb{R}\times\mathbb{R}^N, \tag{1.1}$$

where $N \geq 2$, $1 < p < 1 + 4/(N-2)$, $-1 < \omega < 1$, and $\varphi \in H^1(\mathbb{R}^N)$ is a nontrivial solution of

$$-\Delta\varphi + (1-\omega^2)\varphi - |\varphi|^{p-1}\varphi = 0, \quad x \in \mathbb{R}^N. \tag{1.2}$$

We also study the same problem for the Klein–Gordon–Zakharov (KGZ) system

$$\partial_t^2 u - \Delta u + u + nu = 0, \quad (t,x) \in \mathbb{R}\times\mathbb{R}^N, \tag{1.3}$$

$$c_0^{-2}\partial_t^2 n - \Delta n = \Delta(|u|^2), \quad (t,x) \in \mathbb{R}\times\mathbb{R}^N, \tag{1.4}$$

where $N = 2, 3$, and $c_0 > 0$ is a constant. The system (1.3)–(1.4) describes the interaction of a Langumiur wave and an ion acoustic wave in a plasma. The complex valued function u denotes the fast time scale component of an electric field raised by electrons, and the real valued function n denotes the deviation of ion density (see [34, 4, 8]).

From the result of Ginibre and Velo [10], the Cauchy problem for (1.1) is locally well-posed in the energy space $X := H^1(\mathbb{R}^N) \times L^2(\mathbb{R}^N)$. Thus, for any $(u_0, u_1) \in X$ there exists a unique solution $\vec{u} := (u, \partial_t u) \in C([0, T_{\max}); X)$ of (1.1) with $\vec{u}(0) = (u_0, u_1)$ such that either $T_{\max} = \infty$ (global existence) or $T_{\max} < \infty$ and $\lim_{t\to T_{\max}}$

*Received by the editors October 18, 2005; accepted for publication (in revised form) September 27, 2006; published electronically March 22, 2007.
http://www.siam.org/journals/sima/38-6/64301.html
†Department of Mathematics, Faculty of Science, Saitama University, Saitama 338-8570, Japan (mohta@rimath.saitama-u.ac.jp).
‡Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1300 (todorova@math.utk.edu). The research of the second author was supported in part by NSF grant DMS-0245578.

$\|\vec{u}(t)\|_X = \infty$ (finite time blowup). Moreover, the solution $u(t)$ satisfies the conservation laws of energy and charge:

$$E(\vec{u}(t)) = E(u_0, u_1), \quad Q(\vec{u}(t)) = Q(u_0, u_1), \quad t \in [0, T_{\max}),$$

where

$$E(u, v) = \frac{1}{2}\|v\|_2^2 + \frac{1}{2}\|\nabla u\|_2^2 + \frac{1}{2}\|u\|_2^2 - \frac{1}{p+1}\|u\|_{p+1}^{p+1}, \tag{1.5}$$

$$Q(u, v) = \operatorname{Im} \int_{\mathbb{R}^N} \overline{u} v \, dx. \tag{1.6}$$

Let $\phi_\omega \in H^1(\mathbb{R}^N)$ be the ground state (the least energy solution) of (1.2). We refer to [2, 30] for the existence of ϕ_ω, and to [13] for the uniqueness of ϕ_ω. The stability of standing waves $e^{i\omega t}\phi_\omega$ for (1.1) has been studied by many authors. First, we consider the orbital stability of $e^{i\omega t}\phi_\omega$. Shatah [27] proves that $e^{i\omega t}\phi_\omega$ is orbitally stable if $p < 1 + 4/N$ and $\omega_c < |\omega| < 1$, where

$$\omega_c = \sqrt{\frac{p-1}{4 - (N-1)(p-1)}}. \tag{1.7}$$

Shatah and Strauss [29] prove that $e^{i\omega t}\phi_\omega$ is orbitally unstable when $p < 1 + 4/N$ and $|\omega| < \omega_c$ or when $p \geq 1 + 4/N$ and $|\omega| < 1$. Here, we say that a standing wave solution $e^{i\omega t}\varphi$ is orbitally stable for (1.1) if for any $\varepsilon > 0$ there exists $\delta > 0$ such that if $(u_0, u_1) \in X$ satisfies $\|(u_0, u_1) - (\varphi, i\omega\varphi)\|_X < \delta$, then the solution $u(t)$ of (1.1) with $\vec{u}(0) = (u_0, u_1)$ exists globally and satisfies

$$\sup_{t \geq 0} \inf_{\theta \in \mathbb{R}, y \in \mathbb{R}^N} \|\vec{u}(t) - e^{i\theta}(\varphi(\cdot + y), i\omega\varphi(\cdot + y))\|_X < \varepsilon.$$

Otherwise, $e^{i\omega t}\varphi$ is said to be orbitally unstable.

Next, we consider instability of $e^{i\omega t}\phi_\omega$ in a stronger sense. Berestycki and Cazenave [1] prove that the ground state standing wave $e^{i\omega t}\phi_\omega$ for (1.1) is very strongly unstable (see Definition 1 below) when the frequency $\omega = 0$ (see also [14, 26]). Shatah [28] proves that the ground state standing wave $e^{i\omega t}\phi_\omega$ for nonlinear Klein–Gordon equations with general nonlinearity is strongly unstable (see Definition 2 below) when $\omega = 0$ and $N \geq 3$. Recently, the authors [23] proved that the ground state standing wave $e^{i\omega t}\phi_\omega$ for (1.1) is very strongly unstable when $|\omega| \leq \sqrt{(p-1)/(p+3)}$ and $N \geq 3$. Here, we give the definitions of very strong instability and strong instability.

Definition 1 (very strong instability). *We say that $e^{i\omega t}\varphi$ is very strongly unstable for (1.1) if for any $\varepsilon > 0$ there exists $(u_0, u_1) \in X$ such that $\|(u_0, u_1) - (\varphi, i\omega\varphi)\|_X < \varepsilon$ and the solution $u(t)$ of (1.1) with $\vec{u}(0) = (u_0, u_1)$ blows up in finite time.*

Definition 2 (strong instability). *We say that $e^{i\omega t}\varphi$ is strongly unstable for (1.1) if for any $\varepsilon > 0$ there exists $(u_0, u_1) \in X$ such that $\|(u_0, u_1) - (\varphi, i\omega\varphi)\|_X < \varepsilon$ and the solution $u(t)$ of (1.1) with $\vec{u}(0) = (u_0, u_1)$ either blows up in finite time or exists globally and satisfies $\limsup_{t\to\infty} \|\vec{u}(t)\|_X = \infty$.*

Note that, by the definitions, if $e^{i\omega t}\varphi$ is very strongly unstable, then it is strongly unstable, and that if $e^{i\omega t}\varphi$ is strongly unstable, then it is orbitally unstable.

Before stating our main results, we recall instability results for the nonlinear Schrödinger equation

$$i\partial_t u + \Delta u + |u|^{p-1}u = 0, \quad (t, x) \in \mathbb{R} \times \mathbb{R}^N. \tag{1.8}$$

Let $\omega > 0$ and $\phi_\omega \in H^1(\mathbb{R}^N)$ be the ground state of

$$-\Delta\varphi + \omega\varphi - |\varphi|^{p-1}\varphi = 0, \quad x \in \mathbb{R}^N. \tag{1.9}$$

It is known that for any $\omega > 0$ the standing wave solution $e^{i\omega t}\phi_\omega$ for (1.8) is orbitally stable when $1 < p < 1 + 4/N$, and it is very strongly unstable when $1 + 4/N < p < 1 + 4/(N-2)$ (see [1, 7, 33]). Moreover, for the critical case $p = 1 + 4/N$, for any $\omega > 0$ and any nontrivial solution $\varphi \in H^1(\mathbb{R}^N)$ of (1.9), it is known that the standing wave $e^{i\omega t}\varphi$ is very strongly unstable for (1.8) (see [32]). For general theory of orbital stability and instability of solitary waves, we refer to Grillakis, Shatah, and Strauss [11, 12]. We state our main results.

THEOREM 1. *Let $N \geq 2$, $1 < p < 1+4/(N-2)$, $\omega \in (-1, 1)$, and ϕ_ω be the ground state of* (1.2). *Assume that $|\omega| \leq \omega_c$ if $p < 1 + 4/N$, where the critical frequency ω_c is given by* (1.7). *Then, the standing wave $e^{i\omega t}\phi_\omega$ for the nonlinear Klein–Gordon equation* (1.1) *is strongly unstable in the sense of Definition* 2.

Can we refine further this instability result? Namely, can we prove in certain cases that the standing wave $e^{i\omega t}\phi_\omega$ for (1.1) is very strongly unstable in the sense of Definition 1? The result of Cazenave [5] answers this question for the restricted range for the exponent p of nonlinearity $1 < p \leq 5$ for $N = 2$ and $1 < p \leq N/(N-2)$ for $N \geq 3$. Cazenave proves that any global solution $u(t)$ of (1.1) is uniformly bounded in X, i.e., $\sup_{t\geq 0} \|\vec{u}(t)\|_X < \infty$, if $1 < p \leq 5$ and $N = 2$, and if $1 < p \leq N/(N-2)$ and $N \geq 3$. Therefore, for this range of the exponent p, Theorem 1 together with the result of Cazenave gives us a very strong instability result in the sense of Definition 1 for ground state standing waves $e^{i\omega t}\phi_\omega$ of (1.1). Using an argument in Merle and Zaag [18], we can extend the result of Cazenave and prove the uniform boundedness of global solutions of (1.1) in X when $1 < p < 1+4/(N-1)$ and $N \geq 2$. The following lemma holds.

LEMMA 2. *Let $N \geq 2$ and $1 < p < 1 + 4/(N-1)$. If $\vec{u} \in C([0,\infty), X)$ is a global solution of* (1.1), *then $\sup_{t\geq 0} \|\vec{u}(t)\|_X < \infty$.*

Therefore, from Theorem 1 and Lemma 2, we deduce the following.

COROLLARY 3. *In addition to the assumptions in Theorem* 1, *let $1 < p \leq 1 + 4/(N-1)$ if $N = 2, 3$, and that $1 < p < 1+4/(N-1)$ if $N \geq 4$. Then, the ground state standing wave $e^{i\omega t}\phi_\omega$ for* (1.1) *is very strongly unstable in the sense of Definition* 1.

Remark. Let us mention that when the exponent p of nonlinearity is in the range $1 + 4/(N-1) < p < 1 + 4/(N-2)$ we are unable to give better instability results than those in Theorem 1 for ground state standing waves $e^{i\omega t}\phi_\omega$ of (1.1) for large frequencies $|\omega| > \sqrt{(p-1)/(p+3)}$. The very strong instability result for small frequencies $|\omega| \leq \sqrt{(p-1)/(p+3)}$ and $N \geq 3$ is given in [23]. The following theorem is an important contribution of Kenji Nakanishi on the very strong instability in this area for large p and large frequencies ω.

THEOREM A (due to Kenji Nakanishi). *Let $N \geq 2$, $1+4/N \leq p < 1+4/(N-2)$, $|\omega| < 1$, and ϕ_ω be the ground state of* (1.2). *Then, the standing wave $e^{i\omega t}\phi_\omega$ for the nonlinear Klein–Gordon equation* (1.1) *is very strongly unstable in the sense of Definition* 1.

This way, we have the entire picture for the very strong instability of ground state standing waves.

For the critical frequency $\omega = \omega_c$ in the case $1 < p < 1 + 4/N$, we can prove a much more general instability result for standing waves which are not necessarily related to the ground state.

THEOREM 4. *Let $N \geq 2$, $1 < p < 1 + 4/N$, and $\varphi \in H^1(\mathbb{R}^N)$ be any nontrivial, radially symmetric solution of (1.2) with $\omega = \omega_c$. Then, the standing wave solution $e^{i\omega_c t}\varphi$ of (1.1) is very strongly unstable in the sense of Definition 1. The same assertion is true for $\omega = -\omega_c$.*

For the existence of infinitely many radially symmetric solutions of (1.2), we refer to [3]. As mentioned above, a similar result of Theorem 4 is known for the nonlinear Schrödinger equation (1.8) in the critical case $p = 1+4/N$ without assuming the radial symmetry of the solution of (1.9) and the restriction on space dimensions $N \geq 2$ (see [32]).

The proofs of Theorems 1 and 4 are based on using local versions of the virial-type identities. To prove strong instability of the ground state for the case when $\omega = 0$ and $N \geq 3$, Shatah in [28] considers a local version of the following identity:

$$\frac{d}{dt}\,\mathrm{Re}\int_{\mathbb{R}^N} x\cdot\nabla u\partial_t\bar{u}\,dx = NK_1(\vec{u}(t)), \tag{1.10}$$
$$K_1(u,v) := -\frac{1}{2}\|v\|_2^2 + \left(\frac{1}{2}-\frac{1}{N}\right)\|\nabla u\|_2^2 + \frac{1}{2}\|u\|_2^2 - \frac{1}{p+1}\|u\|_{p+1}^{p+1}.$$

Since the integral in the left-hand side of (1.10) is not well defined on the energy space X, one needs to approximate the weight function x in (1.10) by suitable bounded functions. To control error terms by the approximation, initial perturbations are restricted to being radially symmetric, and the decay estimate for radially symmetric functions in $H^1(\mathbb{R}^N)$,

$$\|w\|_{L^\infty(|x|\geq m)} \leq Cm^{-(N-1)/2}\|w\|_{H^1}, \tag{1.11}$$

(see [30]) is employed. The assumption $N \geq 2$ is needed here. In the case $N = 1$, we expect similar very strong instability results for the standing waves. This kind of approach has been also used for blowup problems of the nonlinear Schrödinger equation (1.8) (see, e.g., [21, 22, 15, 16, 17, 19, 20]).

In the proof of Theorem 1 for the case $p \geq 1 + 4/N$, we use a local version of the virial identity

$$-\frac{d}{dt}\,\mathrm{Re}\int_{\mathbb{R}^N}\{2x\cdot\nabla u + Nu\}\partial_t\bar{u}\,dx = P(u(t)), \tag{1.12}$$

where

$$P(u) := 2\|\nabla u\|_2^2 - \frac{N(p-1)}{p+1}\|u\|_{p+1}^{p+1}. \tag{1.13}$$

Namely, instead of the left-hand side of (1.12), which is not well defined in the energy space X, we use (2.6) with conveniently chosen weights (see the beginning of section 2).

Note that (1.12) follows from (1.10) and

$$\frac{1}{2}\frac{d^2}{dt^2}\|u(t)\|_2^2 = \frac{d}{dt}\,\mathrm{Re}\int_{\mathbb{R}^N} u\partial_t\bar{u}\,dx = -K_2(\vec{u}(t)), \tag{1.14}$$
$$K_2(u,v) = -\|v\|_2^2 + \|\nabla u\|_2^2 + \|u\|_2^2 - \|u\|_{p+1}^{p+1},$$

and that the functional P appears in the virial identity for the nonlinear Schrödinger equation (1.8):

$$\frac{d^2}{dt^2}\|xu(t)\|_2^2 = 4P(u(t)). \tag{1.15}$$

The case $p < 1+4/N$ is more delicate. Here we use a local version of the identity

$$-\frac{d}{dt}\,\mathrm{Re}\int_{\mathbb{R}^N}\{2x\cdot\nabla u+(N+\alpha)u\}\partial_t\bar{u}\,dx = K(\vec{u}(t)), \tag{1.16}$$

where $\alpha := 4/(p-1)-N$ and

$$K(u,v) := -\alpha\|v\|_2^2+\alpha\|u\|_2^2+(\alpha+2)\left\{\|\nabla u\|_2^2-\frac{2}{p+1}\|u\|_{p+1}^{p+1}\right\} \tag{1.17}$$

(cf. [29, page 185]). Note that

$$\begin{aligned} K(u,v) &= P(u)+\alpha K_2(u,v)\\ &= -2(\alpha+1)\|v-i\omega u\|_2^2+2(\alpha+2)(E-\omega Q)(u,v)\\ &\qquad -2\alpha\omega Q(u,v)-2\{1-(\alpha+1)\omega^2\}\|u\|_2^2, \end{aligned} \tag{1.18}$$

and that $1-(\alpha+1)\omega^2>0$ if $|\omega|>\omega_c$, and correspondingly $1-(\alpha+1)\omega^2=0$ if $|\omega|=\omega_c$. Again instead of the left-hand side of (1.16), we use (2.7) with conveniently chosen weights.

Next, we consider the KGZ system (1.3)–(1.4). The well-posedness of the Cauchy problem for (1.3)–(1.4) in the energy space is studied by Ozawa, Tsutaya, and Tsutsumi [25]. Here, the energy space Y is defined by $Y = H^1(\mathbb{R}^N)\times L^2(\mathbb{R}^N)\times L^2(\mathbb{R}^N)\times \dot{H}^{-1}(\mathbb{R}^N)$. When $N=3$ and $c_0\neq 1$, it is proved in [25] that for any $(u_0,u_1,n_0,n_1)\in Y$ there exists a unique solution $\mathbf{u} := (u,\partial_t u,n,\partial_t n)\in C([0,T_{\max});Y)$ of (1.3)–(1.4) with initial data $\mathbf{u}(0)=(u_0,u_1,n_0,n_1)$ satisfying the conservation laws of the energy $H(\mathbf{u}(t))=H(\mathbf{u}(0))$ and the charge $Q(\mathbf{u}(t))=Q(\mathbf{u}(0))$ for all $t\in[0,T_{\max})$, where Q is defined by (1.6) and

$$\begin{aligned} H(u,v,n,\nu) = \frac{1}{2}\|v\|_2^2+\frac{1}{4c_0^2}\|\nu\|_{\dot{H}^{-1}}^2\\ +\frac{1}{2}\|\nabla u\|_2^2+\frac{1}{2}\|u\|_2^2+\frac{1}{4}\|n\|_2^2+\frac{1}{2}\int_{\mathbb{R}^N}|u|^2n\,dx. \end{aligned} \tag{1.19}$$

The case when $N=3$ and $c_0=1$ is treated in [24, 31], where the global small data solutions result is presented. For the case $N=2$, by using the idea of the paper of Ozawa, Tsutaya, and Tsutsumi [25] we can prove the local well-posedness of the Klein–Gordon–Zakharov system (1.3)–(1.4) in the energy space Y for all $c_0>0$.

We study instability of standing wave solutions

$$(u_\omega(t,x),n_\omega(t,x)) = (e^{i\omega t}\phi_\omega(x),-|\phi_\omega(x)|^2)$$

for (1.3)–(1.4), where $-1<\omega<1$, $N=2,3$, and $\phi_\omega\in H^1(\mathbb{R}^N)$ is the ground state of

$$-\Delta\varphi+(1-\omega^2)\varphi-|\varphi|^2\varphi=0,\quad x\in\mathbb{R}^N. \tag{1.20}$$

By a similar method as in the proof of Theorem 1 for the case $p\geq 1+4/N$ together with an argument in Merle [17] for the Zakharov system, we have the following.

THEOREM 5. *Let $N=2,3$, $\omega\in(-1,1)$, ϕ_ω be the ground state of* (1.20)*, and $c_0\neq 1$ if $N=3$. Then, the standing wave $(e^{i\omega t}\phi_\omega,-|\phi_\omega|^2)$ of KGZ system* (1.3)–(1.4) *is strongly unstable in the following sense. For any $\lambda>1$, the solution $\mathbf{u}(t)$ of* (1.3)–(1.4) *with initial data $\mathbf{u}(0)=(\lambda\phi_\omega,\lambda i\omega\phi_\omega,-\lambda^2|\phi_\omega|^2,0)$ either blows up in finite time or exists globally and satisfies $\limsup_{t\to\infty}\|\mathbf{u}(t)\|_Y=\infty$.*

Remark. It is known (see [4, Theorem 3]) that the negative initial energy $H(\mathbf{u}(0))$ implies that the solution $\mathbf{u}(t)$ of (1.3)–(1.4) either blows up in finite time or blows up in infinite time, namely the solution exists globally and satisfies the asymptotic condition $\limsup_{t\to\infty} \|\mathbf{u}(t)\|_Y = \infty$. Since the energy

$$H(\lambda\phi_\omega, \lambda i\omega\phi_\omega, -\lambda^2|\phi_\omega|^2, 0) > 0$$

for λ close to 1, the result in [4] is not applicable to Theorem 5.

Next, we consider the very strong instability of $(e^{i\omega t}\phi_\omega, -|\phi_\omega|^2)$ for (1.3)–(1.4). Since (1.4) of the KGZ system is massless, it seems difficult to obtain the uniform boundedness of global solutions for (1.3)–(1.4) similar to Lemma 2. Therefore, for the standing wave $(e^{i\omega t}\phi_\omega, -|\phi_\omega|^2)$ we do not deduce a very strong instability similar to the instability result in Corollary 3 of Theorem 1. However, using the method in our previous paper [23], we obtain the following very strong instability result for small frequencies.

THEOREM 6. *Let $N = 3$, $c_0 \neq 1$, $|\omega| < 1/\sqrt{3}$, and ϕ_ω be the ground state of* (1.20). *Then, the standing wave $(e^{i\omega t}\phi_\omega, -|\phi_\omega|^2)$ of the KGZ system* (1.3)–(1.4) *is very strongly unstable in the following sense. For any $\lambda > 1$, the solution $\mathbf{u}(t)$ of* (1.3)–(1.4) *with the initial data $\mathbf{u}(0) = (\lambda\phi_\omega, \lambda i\omega\phi_\omega, -\lambda^2|\phi_\omega|^2, 0)$ blows up in finite time.*

Remark. In Theorem 6, the case $\omega = 0$ is proved by Gan and Zhang [9].

The plan of this paper is as follows. In section 2, we prove Theorems 1 and 4 and Lemma 2 for the nonlinear Klein–Gordon equation (1.1). The proof of Theorem A is given at the end of section 2. Section 3 is devoted to applications to the KGZ system (1.3)–(1.4), and we prove Theorems 5 and 6.

2. Proof of theorems for the nonlinear Klein–Gordon equation (1.1). In this section, we first prove Theorems 1 and 4.

We start with a convenient choice of the weight functions, as follows. Let $\Phi \in C^2([0,\infty))$ be a nonnegative function such that

$$\Phi(r) = \begin{cases} N & \text{for} \quad 0 \le r \le 1, \\ 0 & \text{for} \quad r \ge 2, \end{cases} \qquad \Phi'(r) \le 0 \ \text{ for } \ 1 \le r \le 2.$$

For $m > 0$, we put

$$\Phi_m(r) = \Phi\left(\frac{r}{m}\right), \quad \Psi_m(r) = \frac{1}{r^{N-1}} \int_0^r s^{N-1}\Phi_m(s)\, ds. \tag{2.1}$$

Then, Φ_m and Ψ_m satisfy the following properties.

LEMMA 7. *For $m > 0$, we have*

$$\Phi_m(r) = N, \quad \Psi_m(r) = r, \quad 0 \le r \le m, \tag{2.2}$$

$$\Psi_m'(r) + \frac{N-1}{r}\Psi_m(r) = \Phi_m(r), \quad r \ge 0, \tag{2.3}$$

$$|\Phi_m^{(k)}(r)| \le \frac{C}{m^k}, \quad r \ge 0, \ k = 0, 1, 2, \tag{2.4}$$

$$\Psi_m'(r) \le 1, \quad r \ge 0. \tag{2.5}$$

Proof. Properties (2.2)–(2.4) follow from definition (2.1). We show (2.5). Integrating by parts implies

$$Nr^{N-1}\Psi_m(r) = \int_0^r Ns^{N-1}\Phi_m(s)\, ds = r^N\Phi_m(r) - \int_0^r s^N\Phi_m'(s)\, ds.$$

Thus, by (2.3), we have

$$\Psi_m'(r) = \Phi_m(r) - \frac{N-1}{r}\Psi_m(r) = \frac{1}{N}\Phi_m(r) + \frac{N-1}{Nr^N}\int_0^r s^N \Phi_m'(s)\,ds.$$

Since $\Phi_m(r) \le N$ and $\Phi_m'(r) \le 0$ for $r \ge 0$, we have (2.5). □

LEMMA 8. *Let $u(t)$ be a radially symmetric solution of (1.1), and put*

$$I_m^1(t) = 2\,\mathrm{Re}\int_{\mathbb{R}^N} \Psi_m \partial_r u \partial_t \overline{u}\,dx + \mathrm{Re}\int_{\mathbb{R}^N} \Phi_m u \partial_t \overline{u}\,dx, \tag{2.6}$$

$$I_m^2(t) = I_m^1(t) + \alpha\,\mathrm{Re}\int_{\mathbb{R}^N} u\partial_t \bar{u}\,dx, \tag{2.7}$$

where $\alpha := 4/(p-1) - N$. Then, there exists a constant $C_0 > 0$ independent of m such that

$$-\frac{d}{dt}I_m^1(t) \le P(u(t)) + \frac{N(p-1)}{p+1}\int_{|x|\ge m} |u(t,x)|^{p+1}\,dx + \frac{C_0}{m^2}\|u(t)\|_2^2, \tag{2.8}$$

$$-\frac{d}{dt}I_m^2(t) \le K(\vec{u}(t)) + \frac{N(p-1)}{p+1}\int_{|x|\ge m} |u(t,x)|^{p+1}\,dx + \frac{C_0}{m^2}\|u(t)\|_2^2 \tag{2.9}$$

for all $t \in [0, T_{\max})$.

Proof. We multiply (1.1) by $\Psi_m\overline{\partial_r u}$ and by $\Phi_m \overline{u}$, respectively, and have

$$\begin{aligned}-\frac{d}{dt}2\,\mathrm{Re}\int_{\mathbb{R}^N}\Psi_m\partial_r u\partial_t\overline{u}\,dx &= \int_{\mathbb{R}^N}\left(\Psi_m' + \frac{N-1}{r}\Psi_m\right)|\partial_t u|^2\,dx \\ &+ \int_{\mathbb{R}^N}\left(\Psi_m' - \frac{N-1}{r}\Psi_m\right)|\nabla u|^2\,dx - \int_{\mathbb{R}^N}\left(\Psi_m' + \frac{N-1}{r}\Psi_m\right)|u|^2\,dx \\ &+ \frac{2}{p+1}\int_{\mathbb{R}^N}\left(\Psi_m' + \frac{N-1}{r}\Psi_m\right)|u|^{p+1}\,dx\end{aligned}$$

and

$$\begin{aligned}-\frac{d}{dt}\,\mathrm{Re}\int_{\mathbb{R}^N}\Phi_m u\partial_t\overline{u}\,dx &= -\int_{\mathbb{R}^N}\Phi_m|\partial_t u|^2\,dx - \frac{1}{2}\int_{\mathbb{R}^N}\Delta\Phi_m|u|^2\,dx \\ &+ \int_{\mathbb{R}^N}\Phi_m|\nabla u|^2\,dx + \int_{\mathbb{R}^N}\Phi_m|u|^2\,dx - \int_{\mathbb{R}^N}\Phi_m|u|^{p+1}\,dx.\end{aligned}$$

By (2.3) in Lemma 7, we have the identity

$$-\frac{d}{dt}I_m^1(t) = 2\int_{\mathbb{R}^N}\Psi_m'|\nabla u|^2\,dx - \frac{p-1}{p+1}\int_{\mathbb{R}^N}\Phi_m|u|^{p+1}\,dx - \frac{1}{2}\int_{\mathbb{R}^N}\Delta\Phi_m|u|^2\,dx.$$

The inequality (2.8) follows from Lemma 7. Finally, (2.9) follows from (2.8), (1.14), and (1.18). □

First, we consider the case $p \ge 1 + 4/N$. We define the functional

$$J_\omega(u) = \frac{1}{2}\|\nabla u\|_2^2 + \frac{1-\omega^2}{2}\|u\|_2^2 - \frac{1}{p+1}\|u\|_{p+1}^{p+1}, \tag{2.10}$$

and consider the constrained minimization problem

$$d_\omega^1 = \inf\{J_\omega(u) : u \in H^1(\mathbb{R}^N)\setminus\{0\},\ P(u) = 0\} \tag{2.11}$$

and the set

$$\mathcal{R}^1_\omega = \{(u,v) \in X : (E-\omega Q)(u,v) < d^1_\omega,\ P(u) < 0\}, \tag{2.12}$$

where E and Q are the energy and the charge, respectively, and the functional P is defined by (1.13).

Note that

$$(E-\omega Q)(u,v) = J_\omega(u) + \frac{1}{2}\|v - i\omega u\|_2^2, \tag{2.13}$$

$$P(u) = 2\partial_\lambda J_\omega(\lambda^{N/2}u(\lambda\cdot))|_{\lambda=1}. \tag{2.14}$$

LEMMA 9. *Let $N \geq 2$, $1 + 4/N \leq p < 1 + 4/(N-2)$, and $\omega \in (-1,1)$. Then, we have the following:*

(i) $J_\omega(u) - \frac{1}{N(p-1)}P(u) > d^1_\omega$ *for all* $u \in H^1(\mathbb{R}^N)$ *satisfying* $P(u) < 0$.

(ii) *The minimization problem* (2.11) *is attained at the ground state* ϕ_ω *of* (1.2).

(iii) $\lambda(\phi_\omega, i\omega\phi_\omega) \in \mathcal{R}^1_\omega$ *for all* $\lambda > 1$.

Proof. (i) We put

$$\begin{aligned} J^1_\omega(u) &:= J_\omega(u) - \frac{1}{N(p-1)}P(u) \\ &= \left\{\frac{1}{2} - \frac{2}{N(p-1)}\right\}\|\nabla u\|_2^2 + \frac{1-\omega^2}{2}\|u\|_2^2. \end{aligned} \tag{2.15}$$

Note that $1/2 - 2/N(p-1) \geq 0$ by the assumption $p \geq 1 + 4/N$. Let $u \in H^1(\mathbb{R}^N)$ satisfy $P(u) < 0$. Then, we have $u \neq 0$, and there exists $\lambda_1 \in (0,1)$ such that $P(\lambda_1 u) = 0$. By (2.11), we have $d^1_\omega \leq J_\omega(\lambda_1 u) = J^1_\omega(\lambda_1 u) < J^1_\omega(u)$.

(ii) For the case $p > 1 + 4/N$, see [6, Proposition 8.2.4], and for $p = 1 + 4/N$, see [19, Proposition 2.5].

(iii) By (2.13), we have

$$\begin{aligned} &(E-\omega Q)(\lambda(\phi_\omega, i\omega\phi_\omega)) = J_\omega(\lambda\phi_\omega) \\ &= \lambda^2\left(\frac{1}{2}\|\nabla\phi_\omega\|_2^2 + \frac{1-\omega^2}{2}\|\phi_\omega\|_2^2\right) - \frac{\lambda^{p+1}}{p+1}\|\phi_\omega\|_{p+1}^{p+1}. \end{aligned}$$

Since $J_\omega(\phi_\omega) = d^1_\omega$, $\partial_\lambda J_\omega(\lambda\phi_\omega)|_{\lambda=1} = 0$, and $\partial^2_\lambda J_\omega(\lambda\phi_\omega)|_{\lambda=1} < 0$, we have $(E - \omega Q)(\lambda(\phi_\omega, i\omega\phi_\omega)) < d^1_\omega$ for all $\lambda > 1$. Similarly, we have $P(\lambda\phi_\omega) < 0$ for all $\lambda > 1$. Hence, we have $\lambda(\phi_\omega, i\omega\phi_\omega) \in \mathcal{R}^1_\omega$ for all $\lambda > 1$. □

LEMMA 10. *Suppose that $N \geq 2$, $1 + 4/N \leq p < 1 + 4/(N-2)$, and $\omega \in (-1,1)$. If $(u_0, u_1) \in \mathcal{R}^1_\omega$, then the solution $u(t)$ of* (1.1) *with $\vec{u}(0) = (u_0, u_1)$ satisfies*

$$-\frac{1}{N(p-1)}P(u(t)) > d^1_\omega - (E-\omega Q)(u_0,u_1), \quad t \in [0, T_{\max}). \tag{2.16}$$

Proof. First, we show that $P(u(t)) < 0$ for all $t \in [0, T_{\max})$. Suppose that there exists $t_1 \in (0, T_{\max})$ such that $P(u(t_1)) = 0$ and $P(u(t)) < 0$ for $t \in [0, t_1)$. Then, by Lemma 9(i) and (2.15), we have

$$\left\{\frac{1}{2} - \frac{2}{N(p-1)}\right\}\|\nabla u\|_2^2 + \frac{1-\omega^2}{2}\|u(t)\|_2^2 > d^1_\omega > 0, \quad t \in [0, t_1).$$

Thus, we have $u(t_1) \neq 0$. Therefore, by (2.11), we have $d^1_\omega \leq J_\omega(u(t_1))$. Meanwhile, since $(u_0, u_1) \in \mathcal{R}^1_\omega$, E and Q are conserved, and by (2.13), we have $J_\omega(u(t_1)) \leq$

$(E-\omega Q)(\vec{u}(t_1)) < d_\omega^1$. This is a contradiction. Hence, we have $P(u(t)) < 0$ for all $t \in [0, T_{\max})$. From this fact, Lemma 9(i) and (2.13), we obtain (2.16). □

Proof of Theorem 1 *for the case* $p \geq 1 + 4/N$. Let $\lambda > 1$ be fixed and denote

$$\delta := \frac{N(p-1)}{2}\{d_\omega^1 - (E-\omega Q)(\lambda(\phi_\omega, i\omega\phi_\omega))\}.$$

Then, by Lemma 9(iii), we have $\delta > 0$. Suppose that the solution $u(t)$ of (1.1) with $\vec{u}(0) = \lambda(\phi_\omega, i\omega\phi_\omega)$ exists for all $t \in [0, \infty)$ and is uniformly bounded in X, i.e.,

$$M_1 := \sup_{t\geq 0} \|\vec{u}(t)\|_X < \infty. \tag{2.17}$$

Since $u(t)$ is radially symmetric in x for all $t \geq 0$, we define $I_m^1(t)$ for $u(t)$ by (2.6). By (1.11) and (2.17), we have

$$\begin{aligned}&\int_{|x|\geq m} |u(t,x)|^{p+1}\,dx \leq \|u(t)\|_{L^\infty(|x|\geq m)}^{p-1}\|u(t)\|_2^2\\ &\leq Cm^{-(N-1)(p-1)/2}\|u(t)\|_{H^1}^{p+1} \leq CM_1^{p+1}m^{-(N-1)(p-1)/2}\end{aligned}$$

for all $t \geq 0$ and $m > 0$. Note that we assume $N \geq 2$. Thus, there exists $m_0 > 0$ such that

$$\sup_{t\geq 0}\left(\frac{N(p-1)}{p+1}\int_{|x|\geq m_0} |u(t,x)|^{p+1}\,dx + \frac{C_0}{m_0^2}\|u(t)\|_2^2\right) < \delta.$$

Thus, by Lemmas 8 and 10, we have

$$\begin{aligned}&\frac{d}{dt}I_{m_0}^1(t)\\ &\geq -P(u(t)) - \left(\frac{N(p-1)}{p+1}\int_{|x|\geq m_0} |u(t,x)|^{p+1}\,dx + \frac{C_0}{m_0^2}\|u(t)\|_2^2\right)\\ &\geq 2\delta - \delta = \delta\end{aligned}$$

for all $t \geq 0$. Therefore, we have $\lim_{t\to\infty} I_{m_0}^1(t) = \infty$. On the other hand, there exists a constant $C = C(m_0) > 0$ such that $I_{m_0}^1(t) \leq C\|\vec{u}(t)\|_X^2 \leq CM_1^2$ for all $t \geq 0$. This is a contradiction. Hence, for any $\lambda > 1$, the solution $u(t)$ of (1.1) with $\vec{u}(0) = \lambda(\phi_\omega, i\omega\phi_\omega)$ either blows up in finite time or exists for all $t \geq 0$ and $\limsup_{t\to\infty} \|\vec{u}(t)\|_X = \infty$. This completes the proof of Theorem 1 for the case $p \geq 1 + 4/N$. □

Next, we consider the case $p < 1 + 4/N$. For this case, we need a different variational characterization of the ground state ϕ_ω of (1.2) from that for the case $p \geq 1 + 4/N$. We define the functional

$$K_\omega^0(u) = \alpha(1-\omega^2)\|u\|_2^2 + (\alpha+2)\left\{\|\nabla u\|_2^2 - \frac{2}{p+1}\|u\|_{p+1}^{p+1}\right\},$$

and consider the constrained minimization problem

$$d_\omega^0 = \inf\{J_\omega(u) : u \in H^1(\mathbb{R}^N) \setminus \{0\},\ K_\omega^0(u) = 0\} \tag{2.18}$$

and the set

$$\mathcal{R}_\omega^0 = \{(u,v) \in X : (E-\omega Q)(u,v) < d_\omega^0,\ K_\omega^0(u) < 0\}, \tag{2.19}$$

where $\alpha = 4/(p-1) - N > 0$. Note that

$$K_\omega^0(u) = 2\partial_\lambda J_\omega(\lambda^\beta u(\lambda\cdot))|_{\lambda=1}, \quad \beta = \frac{\alpha+N}{2} = \frac{2}{p-1}. \tag{2.20}$$

LEMMA 11. *Let $N \geq 2$, $1 < p < 1 + 4/N$, and $\omega \in (-1,1)$. Then, we have the following:*

(i) $\frac{1-\omega^2}{\alpha+2}\|u\|_2^2 > d_\omega^0$ *for all* $u \in H^1(\mathbb{R}^N)$ *satisfying* $K_\omega^0(u) < 0$.

(ii) *The minimization problem* (2.18) *is attained at the ground state* ϕ_ω *of* (1.2).

(iii) $\lambda(\phi_\omega, i\omega\phi_\omega) \in \mathcal{R}_\omega^0$ *for all* $\lambda > 1$.

Proof. First, we note that

$$J_\omega(u) - \frac{1}{2(\alpha+2)}K_\omega^0(u) = \frac{1-\omega^2}{\alpha+2}\|u\|_2^2, \tag{2.21}$$

$$d_\omega^0 = \inf\left\{\frac{1-\omega^2}{\alpha+2}\|u\|_2^2 : u \in H^1(\mathbb{R}^N)\setminus\{0\},\ K_\omega^0(u) = 0\right\}. \tag{2.22}$$

(i) Let $u \in H^1(\mathbb{R}^N)$ satisfy $K_\omega^0(u) < 0$. Then, we have $u \neq 0$, and there exists $\lambda_1 \in (0,1)$ such that $K_\omega^0(\lambda_1 u) = 0$. By (2.18), we have

$$d_\omega^0 \leq \frac{1-\omega^2}{\alpha+2}\|\lambda_1 u\|_2^2 < \frac{1-\omega^2}{\alpha+2}\|u\|_2^2.$$

(ii) Note that $d_\omega^0 \geq 0$ by (2.22). Let $\{u_j\} \subset H^1(\mathbb{R}^N)$ be a minimizing sequence for (2.18). By considering the Schwarz symmetrization of u_j, we can assume that $\{u_j\} \subset H_{rad}^1(\mathbb{R}^N)$. We refer to [2, Appendix A.III] for the definition and basic properties of the Schwarz symmetrization. By (2.22), we see that $\{u_j\}$ is bounded in $L^2(\mathbb{R}^N)$. Moreover, by $K_\omega^0(u_j) = 0$ and the Gagliardo–Nirenberg inequality, we have

$$\begin{aligned}&(\alpha+2)\|\nabla u_j\|_2^2 + \alpha(1-\omega^2)\|u_j\|_2^2\\ &= \frac{2(\alpha+2)}{p+1}\|u_j\|_{p+1}^{p+1} \leq C\|u_j\|_2^{p+1-\theta}\|\nabla u_j\|_2^\theta,\end{aligned}$$

where $\theta = (p-1)N/2$. Since $p < 1+4/N$, we see that $\theta < 2$ and that $\{u_j\}$ is bounded in $H^1(\mathbb{R}^N)$. Therefore, there exist a subsequence of $\{u_j\}$ (we still denote it by the same letter) and $w \in H_{rad}^1(\mathbb{R}^N)$ such that $u_j \rightharpoonup w$ weakly in $H^1(\mathbb{R}^N)$ and $u_j \to w$ strongly in $L^{p+1}(\mathbb{R}^N)$. Here, we used the fact that the embedding $H_{rad}^1(\mathbb{R}^N) \hookrightarrow L_{rad}^q(\mathbb{R}^N)$ is compact for $2 < q < 2 + 4/(N-2)$ (see [30]). Next, we show that $w \neq 0$. Suppose that $w = 0$. Then, by $K_\omega^0(u_j) = 0$ and the strong convergence $u_j \to 0$ in $L^{p+1}(\mathbb{R}^N)$, we see that $u_j \to 0$ in $H^1(\mathbb{R}^N)$. On the other hand, by $K_\omega^0(u_j) = 0$ and the Sobolev inequality, we have

$$\begin{aligned}&(\alpha+2)\|\nabla u_j\|_2^2 + \alpha(1-\omega^2)\|u_j\|_2^2 = \frac{2(\alpha+2)}{p+1}\|u_j\|_{p+1}^{p+1}\\ &\leq C\{(\alpha+2)\|\nabla u_j\|_2^2 + \alpha(1-\omega^2)\|u_j\|_2^2\}^{(p+1)/2}.\end{aligned}$$

Since $u_j \neq 0$, we have $\|u_j\|_{H^1} \geq C$ for some $C > 0$. This is a contradiction. Thus, we see that $w \in H^1(\mathbb{R}^N)\setminus\{0\}$. Therefore, by (2.21) and (2.22), we have

$$d_\omega^0 \leq \frac{1-\omega^2}{\alpha+2}\|w\|_2^2 \leq \liminf_{j\to\infty}\frac{1-\omega^2}{\alpha+2}\|u_j\|_2^2 = \liminf_{j\to\infty} J_\omega(u_j) = d_\omega^0,$$

and $K_\omega^0(w) \le \liminf_{j\to\infty} K_\omega^0(u_j) = 0$. Moreover, by (i), we have $K_\omega^0(w) = 0$. Therefore, w attains (2.22) and (2.18). Since w attains (2.18), there exists a Lagrange multiplier $\eta \in \mathbb{R}$ such that

$$J'_\omega(w) = \frac{\eta}{2(\alpha+2)}(K_\omega^0)'(w). \tag{2.23}$$

That is, w satisfies

$$-(1-\eta)\Delta w + (1-\omega^2)\left(1 - \frac{\alpha}{\alpha+2}\eta\right) w - (1-\eta)|w|^{p-1}w = 0 \tag{2.24}$$

in $H^{-1}(\mathbb{R}^N)$. First, we show that $\eta < 1$. Suppose that $\eta \ge 1$. Then, by (2.24) and $K_\omega^0(w) = 0$, we have

$$\begin{aligned}
0 &= (1-\eta)\|\nabla w\|_2^2 + (1-\omega^2)\left(1 - \frac{\alpha}{\alpha+2}\eta\right)\|w\|_2^2 - (1-\eta)\|w\|_{p+1}^{p+1} \\
&= \frac{(\eta-1)(p-1)}{2}\|\nabla w\|_2^2 + \frac{\alpha(p-1)(1-\omega^2)}{2(\alpha+2)}\left\{\eta - 1 + \frac{4}{\alpha(p-1)}\right\}\|w\|_2^2 \\
&\ge \frac{2(1-\omega^2)}{\alpha+2}\|w\|_2^2 > 0.
\end{aligned}$$

This is a contradiction. Thus, we have $\eta < 1$. Since we have

$$1 - \eta > 0, \quad (1-\omega^2)\left(1 - \frac{\alpha}{\alpha+2}\eta\right) > 0$$

in (2.24), by [6, Theorem 8.1.1], we have $x \cdot \nabla w \in H^1(\mathbb{R}^N)$. Therefore, by (2.23), we have

$$\begin{aligned}
0 = K_\omega^0(w) &= 2\partial_\lambda J_\omega(\lambda^\beta w(\lambda\cdot))|_{\lambda=1} = 2\langle J'_\omega(w), x\cdot\nabla w + \beta w\rangle \\
&= \frac{\eta}{\alpha+2}\langle (K_\omega^0)'(w), x\cdot\nabla w + \beta w\rangle = \frac{\eta}{\alpha+2}\partial_\lambda K_\omega^0(\lambda^\beta w(\lambda\cdot))|_{\lambda=1},
\end{aligned}$$

where $\beta = (\alpha+N)/2$. Moreover, by $K_\omega^0(w) = 0$, we have

$$\begin{aligned}
&\partial_\lambda K_\omega^0(\lambda^\beta w(\lambda\cdot))|_{\lambda=1} \\
&= \alpha^2(1-\omega^2)\|w\|_2^2 + (\alpha+2)^2\left\{\|\nabla w\|_2^2 - \frac{2}{p+1}\|w\|_{p+1}^{p+1}\right\} \\
&= -2\alpha(1-\omega^2)\|w\|_2^2 < 0.
\end{aligned}$$

Thus, we have $\eta = 0$. Therefore, w satisfies $J'(w) = 0$ and $K_\omega^2(w) = 0$, where

$$K_\omega^2(u) := \langle J'_\omega(u), u\rangle = \|\nabla u\|_2^2 + (1-\omega^2)\|u\|_2^2 - \|u\|_{p+1}^{p+1}.$$

Since ϕ_ω attains

$$\inf\{J_\omega(u) : u \in H^1(\mathbb{R}^N)\setminus\{0\},\ K_\omega^2(u) = 0\}$$

(see, e.g., [23, Lemma 3]), we have $J_\omega(\phi_\omega) \le J_\omega(w)$. On the other hand, ϕ_ω satisfies $K_\omega^0(\phi_\omega) = 0$, and we have $d_\omega^0 = J_\omega(w) \le J_\omega(\phi_\omega)$. Hence, ϕ_ω attains (2.18).

(iii) The proof is similar to that of Lemma 9(iii), and we omit it. □

LEMMA 12. *Suppose that* $N \geq 2$, $1 < p < 1 + 4/N$, *and* $\omega \in (-1, 1)$. *If* $(u_0, u_1) \in \mathcal{R}^0_\omega$, *then the solution* $u(t)$ *of* (1.1) *with* $\vec{u}(0) = (u_0, u_1)$ *satisfies*

$$\frac{1-\omega^2}{\alpha+2}\|u(t)\|_2^2 > d^0_\omega, \quad t \in [0, T_{\max}).$$

Proof. The proof is similar to that for Lemma 10. We omit the details. □

Proof of Theorem 1 *for the case* $p < 1 + 4/N$. Let $\lambda > 1$ be fixed and define

$$\delta_1 = (\alpha+2)\{d^0_\omega - (E - \omega Q)(\lambda(\phi_\omega, i\omega\phi_\omega))\},$$
$$\delta_2 = \alpha\left\{\omega Q(\lambda(\phi_\omega, i\omega\phi_\omega)) - \frac{\omega^2(\alpha+2)}{1-\omega^2}d^0_\omega\right\},$$

and $\delta = \delta_1 + \delta_2$. Then, by Lemma 11(iii), we have $\delta_1 > 0$. Moreover, by Lemma 11(ii) and (2.22), we have

$$\frac{\omega^2(\alpha+2)}{1-\omega^2}d^0_\omega = \omega^2\|\phi_\omega\|_2^2 < \lambda^2\omega^2\|\phi_\omega\|_2^2 = \omega Q(\lambda(\phi_\omega, i\omega\phi_\omega)).$$

Thus, we have $\delta_2 > 0$ and $\delta > 0$. Suppose that the solution $u(t)$ of (1.1) with $\vec{u}(0) = \lambda(\phi_\omega, i\omega\phi_\omega)$ exists for all $t \in [0, \infty)$ and is uniformly bounded in X. Since $u(t)$ is radially symmetric in x for all $t \geq 0$, we define $I^2_m(t)$ for $u(t)$ by (2.7). As in the proof of Theorem 1 for the case $p \geq 1 + 4/N$, there exists $m_0 > 0$ such that

$$\sup_{t\geq 0}\left(\frac{N(p-1)}{p+1}\int_{|x|\geq m_0}|u(t,x)|^{p+1}\,dx + \frac{C_0}{m_0^2}\|u(t)\|_2^2\right) < \delta.$$

Thus, by Lemma 8, we have

$$\frac{d}{dt}I^2_{m_0}(t) \geq -K(\vec{u}(t)) - \delta, \quad t \geq 0.$$

Here, recall that we assume $|\omega| \leq \omega_c$, so we have $1 - (\alpha+1)\omega^2 \geq 0$. Thus, by (1.18) and Lemma 12, we have

$$\begin{aligned}
&-K(\vec{u}(t)) \\
&\geq -2(\alpha+2)(E-\omega Q)(\vec{u}(t)) + 2\alpha\omega Q(\vec{u}(t)) + 2\{1-(\alpha+1)\omega^2\}\|u(t)\|_2^2 \\
&\geq -2(\alpha+2)(E-\omega Q)(\vec{u}(0)) + 2\alpha\omega Q(\vec{u}(0)) + 2\{1-\omega^2-\alpha\omega^2\}\frac{\alpha+2}{1-\omega^2}d^0_\omega \\
&= 2\delta
\end{aligned}$$

for all $t \geq 0$. Therefore, we have $(d/dt)I^2_{m_0}(t) \geq \delta$ for all $t \geq 0$, and $\lim_{t\to\infty} I^2_{m_0}(t) = \infty$. The rest of the proof is the same as in the proof of Theorem 1 for the case $p \geq 1 + 4/N$, and we omit the details. □

Proof of Theorem 4. Let us first note that identity (1.18) contains the reason that in Theorem 4 we can allow any radially symmetric solutions of (1.2), unlike the case of Theorem 1 where we can treat only the ground state of (1.2). Namely, when $\omega = \omega_c$ we have $1 - (\alpha+1)\omega_c^2 = 0$, and therefore the identity (1.18) does not contain the norm $\|u\|_2^2$. Let us recall that in Theorem 1 we control this norm by using the variational characterization of the ground state.

Let $\varphi \in H^1(\mathbb{R}^N) \setminus \{0\}$ be a radially symmetric solution of (1.2) with $\omega = \omega_c$. Let $\lambda > 1$ and put

$$\delta = \alpha\omega_c Q(\lambda(\varphi, i\omega_c\varphi)) - (\alpha+2)(E - \omega_c Q)(\lambda(\varphi, i\omega_c\varphi)).$$

Since $J'_{\omega_c}(\varphi) = 0$, we have $(E - \omega_c Q)(\lambda(\varphi, i\omega_c\varphi)) = J_{\omega_c}(\lambda\varphi) < J_{\omega_c}(\varphi)$ for $\lambda > 1$. Moreover, we have $\omega_c Q(\lambda(\varphi, i\omega_c\varphi)) = \omega_c^2\lambda^2\|\varphi\|_2^2 > \omega_c^2\|\varphi\|_2^2$ for $\lambda > 1$. Thus, we have

$$\delta > \alpha\omega_c^2\|\varphi\|_2^2 - (\alpha+2)J_{\omega_c}(\varphi) = -\frac{1}{2}K^0_{\omega_c}(\varphi) - \{1 - (\alpha+1)\omega_c^2\}\|\varphi\|_2^2.$$

By [6, Theorem 8.1.1], we have $x \cdot \nabla\varphi \in H^1(\mathbb{R}^N)$. Therefore, by (2.20) and by $J'_{\omega_c}(\varphi) = 0$, we have

$$K^0_{\omega_c}(\varphi) = 2\langle J'_{\omega_c}(\varphi), x \cdot \nabla\varphi + \beta\varphi\rangle = 0.$$

Moreover, since $(\alpha+1)\omega_c^2 = 1$, we have $\delta > 0$. Suppose that the solution $u(t)$ of (1.1) with $\vec{u}(0) = \lambda(\varphi, i\omega_c\varphi)$ exists for all $t \in [0, \infty)$ and is uniformly bounded in X. Since $u(t)$ is radially symmetric in x for all $t \geq 0$, we define $I^2_m(t)$ for $u(t)$ by (2.7). As in the proof of Theorem 1 for the case $p \geq 1 + 4/N$, there exists $m_0 > 0$ such that

$$\sup_{t\geq 0}\left(\frac{N(p-1)}{p+1}\int_{|x|\geq m_0}|u(t,x)|^{p+1}\,dx + \frac{C_0}{m_0^2}\|u(t)\|_2^2\right) < \delta.$$

Thus, by Lemma 8, we have

$$\frac{d}{dt}I^2_{m_0}(t) \geq -K(\vec{u}(t)) - \delta, \quad t \geq 0.$$

Moreover, by (1.18) and $(\alpha+1)\omega_c^2 = 1$, we have

$$\begin{aligned}
&-K(\vec{u}(t))\\
&\geq -2(\alpha+2)(E - \omega_c Q)(\vec{u}(t)) + 2\alpha\omega_c Q(\vec{u}(t)) + 2\{1 - (\alpha+1)\omega_c^2\}\|u(t)\|_2^2\\
&\geq -2(\alpha+2)(E - \omega_c Q)(\vec{u}(0)) + 2\alpha\omega_c Q(\vec{u}(0)) = 2\delta
\end{aligned}$$

for all $t \geq 0$. Therefore, we have $(d/dt)I^2_{m_0}(t) \geq \delta$ for all $t \geq 0$, and $\lim_{t\to\infty} I^2_{m_0}(t) = \infty$. On the other hand, there exists a constant $C = C(m_0) > 0$ such that $I^2_{m_0}(t) \leq C\|\vec{u}(t)\|_X^2 \leq C$ for all $t \geq 0$. This is a contradiction. Therefore, for any $\lambda > 1$, the solution $u(t)$ of (1.1) with $\vec{u}(0) = \lambda(\varphi, i\omega_c\varphi)$ either blows up in finite time or exists for all $t \geq 0$ and $\limsup_{t\to\infty}\|\vec{u}(t)\|_X = \infty$. Finally, by Lemma 2, if $u(t)$ exists for all $t \geq 0$, then $\sup_{t\geq 0}\|\vec{u}(t)\|_X < \infty$. Hence, $u(t)$ blows up in finite time. This completes the proof. □

Proof of Lemma 2. By Proposition 3.1 and Lemma 3.5 in [5], we have

$$\sup_{t\geq 0}\|u(t)\|_2 < \infty, \tag{2.25}$$

$$\sup_{t\geq 0}\int_t^{t+1}\|\vec{u}(s)\|_X^2\,ds < \infty. \tag{2.26}$$

By (2.26) and the conservation of energy E, we have

$$C_1 := \sup_{t\geq 0}\int_t^{t+1}\|u(s)\|_{p+1}^{p+1}\,ds < \infty. \tag{2.27}$$

Note that the estimates (2.25), (2.26), and (2.27) hold true for $1 < p < 1 + 4/(N-2)$. In what follows, we use an argument in Merle and Zaag [18]. First, for $r = (p+3)/2$, we show

$$\sup_{t\geq 0}\|u(t)\|_r < \infty. \tag{2.28}$$

Indeed, by (2.27) and the mean value theorem, for any $t \geq 0$ there exists $\tau(t) \in [t, t+1]$ such that

$$\|u(\tau(t))\|_{p+1}^{p+1} = \int_t^{t+1} \|u(s)\|_{p+1}^{p+1}\, ds \leq C_1. \tag{2.29}$$

Since $2 < r < p+1$, it follows from (2.25) and (2.29) that $\sup_{t\geq 0} \|u(\tau(t))\|_r < \infty$. Moreover, for any $t \geq 0$, we have

$$\begin{aligned}
&\|u(t)\|_r^r - \|u(\tau(t))\|_r^r = \int_{\tau(t)}^t \frac{d}{ds}\|u(s)\|_r^r\, ds \\
&\leq C \int_t^{t+1} \int_{\mathbb{R}^N} |u(s,x)|^{r-1} |\partial_s u(s,x)|\, dx\, ds \\
&\leq C \int_t^{t+1} \left(\|u(s)\|_{2(r-1)}^{2(r-1)} + \|\partial_s u(s)\|_2^2 \right) ds.
\end{aligned}$$

Since $2(r-1) = p+1$, by (2.26), (2.27), and $\sup_{t\geq 0} \|u(\tau(t))\|_r < \infty$, we have (2.28). Next, by the Gagliardo–Nirenberg inequality, we have

$$\|u(t)\|_{p+1} \leq C \|u(t)\|_r^{1-\theta} \|\nabla u(t)\|_2^{\theta},$$

where

$$\frac{1}{p+1} = \theta \left(\frac{1}{2} - \frac{1}{N} \right) + \frac{1-\theta}{r}.$$

Since we assume $p < 1 + 4/(N-1)$, we have $(p+1)\theta < 2$. Thus, by (2.28), there exists a constant $C_2 > 0$ such that

$$\frac{2}{p+1} \|u(t)\|_{p+1}^{p+1} \leq C_2 + \frac{1}{2} \|\nabla u(t)\|_2^2, \quad t \geq 0.$$

Moreover, by the conservation of energy E, for any $t \geq 0$ we have

$$\begin{aligned}
\|\vec{u}(t)\|_X^2 &= 2E(\vec{u}(0)) + \frac{2}{p+1} \|u(t)\|_{p+1}^{p+1} \\
&\leq 2E(\vec{u}(0)) + C_2 + \frac{1}{2} \|\nabla u(t)\|_2^2,
\end{aligned}$$

which implies $\|\vec{u}(t)\|_X^2 \leq 4E(\vec{u}(0)) + 2C_2$. This completes the proof. □

We conclude this section with the proof of Theorem A.

Proof of Theorem A (due to Kenji Nakanishi). Following the proof of Theorem 1, take the radially symmetric solution $u(t,r)$ $(r = |x|)$ starting from $(u(0), \partial_t u(0)) = \lambda(\phi_\omega, i\omega\phi_\omega)$ with $\lambda > 1$, and assume by contradiction that it exists for all $t \geq 0$. Then Cazenave's estimate (2.26) implies that there exists $M < \infty$ such that for all $T > 0$

$$\int_T^{T+1} \int_{\mathbb{R}^N} |\partial_t u|^2 + |\nabla u|^2 + |u|^2 dx dt \leq M. \tag{2.30}$$

Hence for any positive integer j, there exists $T_j \in [j-1, j]$ such that

$$\int_{\mathbb{R}^N} |\partial_t u|^2 + |\nabla u|^2 + |u|^2 dx|_{t=T_j} \leq M.$$

By Lemmas 8, 9, and 10, there exists $\delta > 0$ such that for any $m > 1$ and $t > 0$ we have

$$\frac{d}{dt} I_m^1(t) \geq 2\delta - R_m(t), \quad R_m(t) := \frac{N(p-1)}{p+1} \int_{|x|\geq m} |u|^{p+1} dx + \frac{C}{m^2} \|u(t)\|_2^2,$$

where I_m^1 is defined by (2.6). Here and below, C is a positive constant, which may depend only on p and N. Integrating in t, we get

$$I_m^1(T_{j+2}) - I_m^1(T_j) \geq 2\delta - \int_{T_j}^{T_{j+2}} R_m(t) dt,$$

since $T_{j+2} - T_j \geq 1$. Notice that (2.30) is enough to control the error term R_m uniformly in j. To see this, let $\chi(t,r) \in C^\infty(\mathbb{R}^2)$ satisfy $\chi(t,r) = 1$ when $|t| \leq 2$ and $|r| \geq 1$, and $\chi(t,r) = 0$ if $|t| \geq 4$ or $|r| \leq 1/2$. For any $m > 1$ and $T > 4$, let $v(t,r) = \chi(t - T, r/m)u(t,|r|)$. Then we have

$$\int_{\mathbb{R}^2} |\partial_t v|^2 + |\partial_r v|^2 + |v|^2 dr dt \leq Cm^{1-N} \int_{T-4}^{T+4} \int_{\mathbb{R}^N} |\partial_t u|^2 + |\nabla u|^2 + |u|^2 dx dt \leq 8Cm^{1-N} M.$$

Hence the Sobolev embedding $H^1(\mathbb{R}^2) \subset L^{p+1}(\mathbb{R}^2)$ implies that

$$\int_{T-2}^{T+2} \int_{|x|\geq m} |u|^{p+1} dx dt \leq C \sum_{j=0}^{\infty} \int_{T-2}^{T+2} (2^j m)^{N-1} \int_{r \geq 2^j m} |u|^{p+1} dr dt \leq Cm^{-(p-1)(N-1)/2} M^{(p+1)/2}.$$

Therefore choosing m sufficiently large, we obtain

$$I_m^1(T_{j+2}) - I_m^1(T_j) \geq \delta$$

for all $j \geq 4$, which contradicts the global bound

$$I_m^1(T_j) \leq Cm \int_{\mathbb{R}^N} |\partial_t u|^2 + |\partial_r u|^2 + |u|^2 dx|_{t=T_j} \leq CmM.$$

This completes the proof. □

3. Proof of theorems for the KGZ system. In this section, we prove Theorems 5 and 6.

Proof of Theorem 5. Let $\lambda > 1$ and put

$$\tilde{d}_\omega = (H - \omega Q)(\phi_\omega, i\omega\phi_\omega, -|\phi_\omega|^2, 0),$$
$$\delta = N\{\tilde{d}_\omega - (H - \omega Q)(\lambda\phi_\omega, \lambda i\omega\phi_\omega, -\lambda^2|\phi_\omega|^2, 0)\},$$

where H and Q are defined by (1.19) and (1.6), respectively. In the same way as in Lemma 9(iii), we see that $\delta > 0$. Suppose that the solution $\mathbf{u}(t)$ of (1.3)–(1.4) with $\mathbf{u}(0) = (\lambda\phi_\omega, \lambda i\omega\phi_\omega, -\lambda^2|\phi_\omega|^2, 0)$ exists globally and satisfies $M := \sup_{t\geq 0} \|\mathbf{u}(t)\|_Y < \infty$. Note that since the initial data is radially symmetric, the solution $\mathbf{u}(t)$ is also

radially symmetric for all $t \geq 0$. Following Merle [17], we introduce the function $w(t) := -(-\Delta)^{-1}\partial_t n(t)$, and for $m > 0$ we consider the function

$$\tilde{I}_m(t) = I_m^1(t) + \frac{1}{c_0^2}\int_{\mathbb{R}^N} \Psi_m n(t)\partial_r w(t)\,dx,$$

where $I_m^1(t)$ is defined by (2.6) and Φ_m and Ψ_m are given by (2.1). Note that since $\partial_t n(t) \in \dot{H}^{-1}(\mathbb{R}^N)$, we see that $w(t) \in \dot{H}^1(\mathbb{R}^N)$ and $\|\partial_t n\|_{\dot{H}^{-1}} = \|\nabla w\|_2$. By the same computations as in Lemma 8, we have

$$\begin{aligned}-\frac{d}{dt}\tilde{I}_m(t) &= 2\int_{\mathbb{R}^N} \Psi_m'|\nabla u|^2\,dx + \frac{1}{2}\int_{\mathbb{R}^N}\Phi_m(n^2 + 2|u|^2 n)\,dx \\ &\quad -\frac{1}{2}\int_{\mathbb{R}^N}\Delta\Phi_m|u|^2\,dx + \frac{1}{2c_0^2}\int_{\mathbb{R}^N}\left(\Psi_m' - \frac{N-1}{r}\Psi_m\right)|\nabla w|^2\,dx.\end{aligned}$$

By Lemma 7, we have

$$\begin{aligned}&\int_{\mathbb{R}^N}\Psi_m'|\nabla u|^2\,dx \leq \|\nabla u(t)\|_2^2, \\ &-\frac{1}{2}\int_{\mathbb{R}^N}\Delta\Phi_m|u|^2\,dx \leq \frac{C_1}{m^2}\|u(t)\|_2^2 \leq \frac{C_1 M^2}{m^2}, \\ &\int_{\mathbb{R}^N}\left(\Psi_m' - \frac{N-1}{r}\Psi_m\right)|\nabla w|^2\,dx \leq \|\nabla w(t)\|_2^2 = \|\partial_t n(t)\|_{\dot{H}^{-1}}.\end{aligned}$$

Moreover, we have

$$\begin{aligned}&\int_{\mathbb{R}^N}\Phi_m(n^2 + 2|u|^2 n)\,dx \\ &= \int_{\mathbb{R}^N}\Phi_m(n + |u|^2)^2\,dx - \int_{\mathbb{R}^n} N|u|^4\,dx + \int_{\mathbb{R}^n}(N - \Phi_m)|u|^4\,dx \\ &\leq N\|n + |u|^2\|_2^2 - N\|u\|_4^4 + \int_{|x|\geq m}(N - \Phi_m)|u|^4\,dx,\end{aligned}$$

and by (1.11) we have

$$\begin{aligned}&\frac{1}{2}\int_{|x|\geq m}(N - \Phi_m)|u|^4\,dx \leq C\|u(t)\|_{L^\infty(|x|\geq m)}^2\|u(t)\|_2^2 \\ &\leq \frac{C_2}{m^{N-1}}\|u(t)\|_{H^1}^4 \leq \frac{C_2 M^4}{m^{N-1}}.\end{aligned}$$

Therefore, we have

$$-\frac{d}{dt}\tilde{I}_m(t) \leq \tilde{P}(\mathbf{u}(t)) + \frac{C_1 M^2}{m^2} + \frac{C_2 M^4}{m^{N-1}} \tag{3.1}$$

for all $t \geq 0$, where we put

$$\tilde{P}(u, v, n, \nu) = 2\|\nabla u\|_2^2 - \frac{N}{2}\|u\|_4^4 + \frac{N}{2}\|n + |u|^2\|_2^2 + \frac{1}{2c_0^2}\|\nu\|_{\dot{H}^{-1}}^2.$$

Note that

$$\begin{aligned}(H-\omega Q)(u,v,n,\nu)&-\frac{1}{2N}\tilde{P}(u,v,n,\nu)\\&=\frac{1}{2}\|v-i\omega u\|_2^2+\left(1-\frac{1}{N}\right)\frac{1}{4c_0^2}\|\nu\|_{\dot{H}^{-1}}^2+\left(\frac{1}{2}-\frac{1}{N}\right)\|\nabla u\|_2^2+\frac{1-\omega^2}{2}\|u\|_2^2\\&\geq\left(\frac{1}{2}-\frac{1}{N}\right)\|\nabla u\|_2^2+\frac{1-\omega^2}{2}\|u\|_2^2.\end{aligned}$$

Using this inequality, in the same way as in Lemmas 9 and 10, we see that

$$-\tilde{P}(\mathbf{u}(t))\geq 2N\{\tilde{d}_\omega-(H-\omega Q)(\mathbf{u}(0))\}=2\delta \tag{3.2}$$

holds for all $t\geq 0$. Therefore, taking $m_1>0$ such that

$$\frac{C_1M^2}{m_1^2}+\frac{C_2M^4}{m_1^{N-1}}<\delta,$$

by (3.1) and (3.2), we have $(d/dt)\tilde{I}_{m_1}(t)\geq\delta$ for all $t\geq 0$, and $\lim_{t\to\infty}\tilde{I}_{m_1}(t)=\infty$. The rest of the proof is the same as in the proof of Theorem 1 for the case $p\geq 1+4/N$, and we omit the details. □

Proof of Theorem 6. Let $\lambda>1$. Suppose that the solution $\mathbf{u}(t)$ of (1.3)–(1.4) with $\mathbf{u}(0)=(\lambda\phi_\omega,\lambda i\omega\phi_\omega,-\lambda^2|\phi_\omega|^2,0)$ exists globally. By the assumption $|\omega|<1/\sqrt{3}$, we can take α such that $2\omega^2/(1-\omega^2)<\alpha<1$. For such an α, we consider a function defined by

$$I_\alpha(t)=\frac{1}{2}\left\{\|u(t)\|_2^2+\frac{\alpha}{c_0^2}\|n(t)\|_{\dot{H}^{-1}}^2\right\}.$$

Note that since $n(0)=-\lambda^2|\phi_\omega|^2\in L^1(\mathbb{R}^3)\cap L^2(\mathbb{R}^3)\subset\dot{H}^{-1}(\mathbb{R}^3)$ and $\partial_t n\in C([0,\infty);\dot{H}^{-1}(\mathbb{R}^3))$, we see that $n\in C^1([0,\infty);\dot{H}^{-1}(\mathbb{R}^3)\cap L^2(\mathbb{R}^3))$. Then, we have

$$\begin{aligned}\frac{d}{dt}I_\alpha(t)&=\mathrm{Re}\langle u(t),\partial_t u(t)\rangle_{L^2}+\frac{\alpha}{c_0^2}\langle n(t),\partial_t n(t)\rangle_{\dot{H}^{-1}}\\&=\mathrm{Re}\langle u(t),\partial_t u(t)-i\omega u(t)\rangle_{L^2}+\frac{\alpha}{c_0^2}\langle n(t),\partial_t n(t)\rangle_{\dot{H}^{-1}}\end{aligned}$$

and

$$\begin{aligned}\frac{d^2}{dt^2}I_\alpha(t)=\|\partial_t u(t)\|_2^2+\frac{\alpha}{c_0^2}\|\partial_t n(t)\|_{\dot{H}^{-1}}^2-\|\nabla u(t)\|_2^2-\|u(t)\|_2^2\\-\alpha\|n(t)\|_2^2-(1+\alpha)\int_{\mathbb{R}^3}|u(t,x)|^2n(t,x)\,dx.\end{aligned}$$

Thus, we have

$$\begin{aligned}&\frac{d^2}{dt^2}I_\alpha(t)+2(1+\alpha)(H-\omega Q)(\mathbf{u}(0))-2\omega Q(\mathbf{u}(0))\\&=(2+\alpha)\|\partial_t u(t)-i\omega u(t)\|_2^2+\left(2+\frac{1-\alpha}{2\alpha}\right)\frac{\alpha}{c_0^2}\|\partial_t n(t)\|_{\dot{H}^{-1}}^2\\&\qquad+K_{\omega,\alpha}(u(t),n(t)),\end{aligned}$$

where we put

$$K_{\omega,\alpha}(u,n)=\alpha\left\{\|\nabla u\|_2^2+\left(1-\omega^2-\frac{2}{\alpha}\omega^2\right)\|u\|_2^2+\frac{1-\alpha}{2\alpha}\|n\|_2^2\right\}.$$

Here, we define

$$\begin{aligned}
&J_\omega(u,n)=\frac{1}{2}\|\nabla u\|_2^2+\frac{1-\omega^2}{2}\|u\|_2^2+\frac{1}{4}\|n\|_2^2+\frac{1}{2}\int_{\mathbb{R}^3}|u(x)|^2n(x)\,dx,\\
&K^1_{\omega,\alpha}(u,n)=\partial_\lambda J_\omega(\lambda u,\lambda^{2\alpha}n)|_{\lambda=1}\\
&=\|\nabla u\|_2^2+(1-\omega^2)\|u\|_2^2+\alpha\|n\|_2^2+(1+\alpha)\int_{\mathbb{R}^3}|u|^2n\,dx,\\
&K^2_{\omega,\alpha}(u,n)=2\partial_\lambda J_\omega(\lambda^{(1-\alpha)/\alpha}u(\cdot/\lambda),n(\cdot/\lambda))|_{\lambda=1}\\
&=\frac{2-\alpha}{\alpha}\|\nabla u\|_2^2+\frac{2+\alpha}{\alpha}(1-\omega^2)\|u\|_2^2\\
&\qquad\qquad+\frac{3}{2}\|n\|_2^2+\frac{2+\alpha}{\alpha}\int_{\mathbb{R}^3}|u|^2n\,dx,
\end{aligned}$$

and put

$$\begin{aligned}
&J^1_{\omega,\alpha}(u,n)=J_\omega(u,n)-\frac{1}{2(1+\alpha)}K^1_{\omega,\alpha}(u,n)\\
&=\frac{\alpha}{1+\alpha}\left\{\frac{1}{2}\|\nabla u\|_2^2+\frac{1-\omega^2}{2}\|u\|_2^2+\frac{1-\alpha}{4\alpha}\|n\|_2^2\right\},\\
&J^2_{\omega,\alpha}(u,n)=J_\omega(u,n)-\frac{\alpha}{2(2+\alpha)}K^2_{\omega,\alpha}(u,n)\\
&=\frac{\alpha}{2+\alpha}\left\{\|\nabla u\|_2^2+\frac{1-\alpha}{2\alpha}\|n\|_2^2\right\},\\
&\theta=1-\frac{2\omega^2}{(1-\omega^2)\alpha}.
\end{aligned}$$

Then, we have $0<\theta<1$ and

$$K_{\omega,\alpha}(u,n)=2(1+\alpha)\theta J^1_{\omega,\alpha}(u,n)+(2+\alpha)(1-\theta)J^2_{\omega,\alpha}(u,n).$$

Moreover, in a similar way as in Lemmas 3 and 4 in [23], we can prove that $J^j_{\omega,\alpha}(u(t),n(t))\geq\tilde{d}_\omega$ for all $t\geq0$ and $j=1,2$. Therefore, we have

$$\begin{aligned}
K_{\omega,\alpha}(u(t),n(t))&\geq\{2(1+\alpha)\theta+(2+\alpha)(1-\theta)\}\,\tilde{d}_\omega\\
&=2\left(1+\alpha-\frac{\omega^2}{1-\omega^2}\right)\tilde{d}_\omega
\end{aligned}$$

for all $t\geq0$. Moreover, since we have $\tilde{d}_\omega=(1-\omega^2)\|\phi_\omega\|_2^2$, putting $\beta=\min\{2+\alpha,2+(1-\alpha)/2\alpha\}$, we have

$$\begin{aligned}
\frac{d^2}{dt^2}I_\alpha(t)&\geq\beta\left\{\|\partial_tu(t)-i\omega u(t)\|_2^2+\frac{\alpha}{c_0^2}\|\partial_tn(t)\|^2_{\dot{H}^{-1}}\right\}\\
&\quad+2(1+\alpha)\{\tilde{d}_\omega-(H-\omega Q)(\mathbf{u}(0))\}+2\omega Q(\mathbf{u}(0))-2\omega^2\|\phi_\omega\|_2^2
\end{aligned}$$

for all $t \geq 0$. Since $\beta > 2$, $(H - \omega Q)(\mathbf{u}(0)) < \tilde{d}_\omega$, and $\omega Q(\mathbf{u}(0)) > \omega^2 \|\phi_\omega\|_2^2$ for all $\lambda > 1$, by the standard concavity argument, we see that there exists $T_1 \in (0, \infty)$ such that $\lim_{t \to T_1 - 0} I_\alpha(t) = \infty$. This is a contradiction. Hence, for all $\lambda > 1$, the solution $\mathbf{u}(t)$ of (1.3)–(1.4) with $\mathbf{u}(0) = (\lambda\phi_\omega, \lambda i\omega\phi_\omega, -\lambda^2|\phi_\omega|^2, 0)$ blows up in finite time. This completes the proof. □

Acknowledgments. The first author thanks the Department of Mathematics, University of Tennessee for its hospitality. The authors acknowledge Kenji Nakanishi who made an important contribution by proving Theorem A. Finally, the authors thank the referees for their useful comments.

REFERENCES

[1] H. Berestycki and T. Cazenave, *Instabilité des états stationnaires dans les équations de Schrödinger et de Klein–Gordon nonlinéaires*, C. R. Acad. Sci. Paris, 293 (1981), pp. 489–492.

[2] H. Berestycki and P. L. Lions, *Nonlinear scalar field equations* I, Arch. Ration. Mech. Anal., 82 (1983), pp. 313–345.

[3] H. Berestycki and P. L. Lions, *Nonlinear scalar field equations* II, Arch. Ration. Mech. Anal., 82 (1983), pp. 347–375.

[4] L. Bergé, B. Bidégaray, and T. Colin, *A perturbative analysis of the time-envelope approximation in strong Langmuir turbulence*, Phys. D, 95 (1996), pp. 351–379.

[5] T. Cazenave, *Uniform estimates for solutions of nonlinear Klein–Gordon equations*, J. Funct. Anal., 60 (1985), pp. 36–55.

[6] T. Cazenave, *Semilinear Schrödinger Equations*, Courant Lect. Notes in Math. 10, New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2003.

[7] T. Cazenave and P. L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equations*, Comm. Math. Phys., 85 (1982), pp. 549–561.

[8] R. O. Dendy, *Plasma Dynamics*, Oxford University Press, Oxford, UK, 1990.

[9] Z. Gan and J. Zhang, *Instability of standing waves for Klein–Gordon–Zakharov equations with different propagation speeds in three space dimesions*, J. Math. Anal. Appl., 307 (2005), pp. 219–231.

[10] J. Ginibre and G. Velo, *The global Cauchy problem for the nonlinear Klein–Gordon equation*, Math. Z., 189 (1985), pp. 487–505.

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

[12] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry*, II, J. Funct. Anal., 94 (1990), pp. 308–348.

[13] M. K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* $\mathbb{R}^n$, Arch. Ration. Mech. Anal., 105 (1989), pp. 234–266.

[14] H. A. Levine, *Instability and nonexistence of global solutions to nonlinear wave equations of the form* $Pu_{tt} = -Au + F(u)$, Trans. Amer. Math. Soc., 192 (1974), pp. 1–21.

[15] F. Merle, *On uniqueness and continuation properties after blow-up time of self-similar solutions of nonlinear Schrödinger equation with critical exponent and critical mass*, Comm. Pure Appl. Math., 45 (1992), pp. 203–254.

[16] F. Merle, *Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equations with critical power*, Duke Math. J., 69 (1993), pp. 427–454.

[17] F. Merle, *Blow-up results of virial type for Zakharov equations*, Comm. Math. Phys., 175 (1996), pp. 433–455.

[18] F. Merle and H. Zaag, *Determination of the blow-up rate for the semilinear wave equation*, Amer. J. Math., 125 (2003), pp. 1147–1164.

[19] H. Nawa, *Asymptotic profiles of blow-up solutions of the nonlinear Schrödinger equation with critical power nonlinearity*, J. Math. Soc. Japan, 46 (1994), pp. 557–586.

[20] H. Nawa, *Asymptotic and limiting profiles of blowup solutions of the nonlinear Schrödinger equation with critical power*, Comm. Pure Appl. Math., 52 (1999), pp. 193–270.

[21] T. Ogawa and Y. Tsutsumi, *Blow-up of* H^1 *solution for the nonlinear Schrödinger equation*, J. Differential Equations, 92 (1991), pp. 317–330.

[22] T. Ogawa and Y. Tsutsumi, *Blow-up of* H^1 *solutions for the one-dimensional nonlinear Schrödinger equation with critical power nonlinearity*, Proc. Amer. Math. Soc., 111 (1991), pp. 487–496.

[23] M. Ohta and G. Todorova, *Strong instability of standing waves for nonlinear Klein–Gordon equations*, Discrete Contin. Dyn. Syst., 12 (2005), pp. 315–322.
[24] T. Ozawa, K. Tsutaya, and Y. Tsutsumi, *Normal form and global solutions for the Klein–Gordon–Zakharov equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 459–503.
[25] T. Ozawa, K. Tsutaya, and Y. Tsutsumi, *Well-posedness in energy space for the Cauchy problem of the Klein–Gordon–Zakharov equations with different propagation speeds in three space dimensions*, Math. Ann., 313 (1999), pp. 127–140.
[26] L. E. Payne and D. H. Sattinger, *Saddle points and instability of nonlinear hyperbolic equations*, Israel J. Math., 22 (1975), pp. 273–303.
[27] J. Shatah, *Stable standing waves of nonlinear Klein–Gordon equations*, Comm. Math. Phys., 91 (1983), pp. 313–327.
[28] J. Shatah, *Unstable ground state of nonlinear Klein–Gordon equations*, Trans. Amer. Math. Soc., 290 (1985), pp. 701–710.
[29] J. Shatah and W. Strauss, *Instability of nonlinear bound states*, Comm. Math. Phys., 100 (1985), pp. 173–190.
[30] W. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.
[31] K. Tsutaya, *Global existence of small amplitude solutions for the Klein–Gordon–Zakharov equations*, Nonlinear Anal., 27 (1996), pp. 1373–1380.
[32] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1983), pp. 567–576.
[33] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.
[34] V. E. Zakharov, *Collapse of Langmuir waves*, Soviet Phys. JETP, 35 (1972), pp. 908–914.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1932–1966

POISSON–NERNST–PLANCK SYSTEMS FOR ION CHANNELS WITH PERMANENT CHARGES*

BOB EISENBERG† AND WEISHI LIU‡

Abstract. Ionic channels and semiconductor devices use atomic scale structures to control macroscopic flows from one reservoir to another. The one-dimensional steady-state Poisson-Nernst-Planck (PNP) system is a useful representation of these devices, but experience shows that describing the reservoirs as boundary conditions is difficult. We study the PNP system for two types of ions with three regions of piecewise constant permanent charge, assuming the Debye number is large, because the electric field is so strong compared to diffusion. Reservoirs are represented by the outer regions with permanent charge zero. If the reciprocal of the Debye number is viewed as a singular parameter, the PNP system can be treated as a singularly perturbed system that has two limiting systems: inner and outer systems (termed fast and slow systems in geometric singular perturbation theory). A complete set of integrals for the inner system is presented that provides information for boundary and internal layers. Application of the exchange lemma from geometric singular perturbation theory gives rise to the existence and (local) uniqueness of the solution of the singular boundary value problem near each singular orbit. A set of simultaneous equations appears in the construction of singular orbits. Multiple solutions of such equations in this or similar problems might explain a variety of multiple valued phenomena seen in biological channels, for example, some forms of gating, and might be involved in other more complex behaviors, for example, some kinds of active transport.

Key words. singular perturbation, boundary layers, internal layers

AMS subject classifications. 34A26, 34B16, 34D15, 37D10, 92C35

DOI. 10.1137/060657480

1. Introduction.

Electrodiffusion, the diffusion and migration of electric charge, plays a central role in a wide range of our technology and science [53, 11, 54, 14, 15, 67, 41]: semiconductor technology controls the migration and diffusion of quasi-particles of charge in transistors and integrated circuits [75, 62, 71], chemical sciences deal with charged molecules in water [11, 19, 8, 26, 9, 10], all of biology occurs in plasmas of ions and charged organic molecules in water [3, 16, 33, 72]. It is no coincidence that the physics of electrodiffusion is of such general importance: systems of moving charge have a richness of behavior that can be sometimes easily controlled by boundary conditions [67, 71], and the goal of technology (and much of physical science) is to control systems to allow useful behavior.

Control is important to the medical and biological sciences as well. Medicine seeks to control disease and help life. Evolution controls life by selecting those organisms that successfully reproduce. Organisms control their internal environment and external behavior to make reproduction possible, often using electrodiffusion for the mechanism of control [72, 33]. Whatever the reason, it is a fact that nearly all biology occurs in ultrafiltrates of blood called plasmas, in which ions move much as they move in gaseous plasmas, or as quasi-particles move in semiconductors [21, 22, 23, 24].

*Received by the editors April 18, 2006; accepted for publication (in revised form) September 27, 2006; published electronically March 30, 2007.
http://www.siam.org/journals/sima/38-6/65748.html
†Department of Molecular Biophysics and Physiology, Rush Medical Center, 1759 Harrison St., Chicago, IL 60612 (beisenbe@rush.edu). This author was partially supported by NIH grant NIGMS-076013-01.
‡Department of Mathematics, University of Kansas, 1460 Jayhawk Blvd., Room 405, Lawrence, KS 66045 (wliu@math.ku.edu). This author was partially supported by NSF grant DMS-0406998.

The pun between the medical and physical meanings of "plasma" is useful and surprisingly precise.

In semiconductor and biological devices, macroscopic flows of charges are driven through tiny (atomic scale) channels that link one macroscopic reservoir to another. The reservoirs are macroscopic regions in which the concentration of charges is nearly constant (because the dimensions of the reservoirs are macroscopic and so the total number of charges is hardly changed by the flows) and electrical potentials are nearly constant too. The electrical resistance of the macroscopic region is so small that only a tiny electrical potential gradient is needed to drive significant flow of charge in the reservoir. The electric field is strong throughout these systems and only a few charges (ions) are needed to create significant electrical potentials, compared to the enormous number of ions (10^{23}, Avogadro's number) needed to create chemical potentials (and diffusion). That is why the Debye number is so large (see systems (1) and (5)). Semiconductors and evolution take advantage of the strength of the electric field. Engineers and biophysicists control flow by setting the electric potential at the boundaries called terminals, contacts, or baths.

The flow through the atomic scale channel is affected by other variables besides the applied boundary potentials, namely, by the shape of the pore in the channel (through which permanently charged ions flow) and the distribution of permanent and induced (i.e., polarization) charge on the wall of the channel as well as the mobility of ions [49, 25, 17, 40, 32]. A precise description and understanding of flow on an atomic scale is daunting. Enormous numbers of variables are needed to describe atomic scale trajectories that have a fundamental time scale 10^{-16} sec and length scale 10^{-10} m compared to biological function that is typically much slower than 10^{-5} sec. It is not clear what to do with this number of variables and trajectories even if they could be computed accurately or with known inaccuracies.

We are fortunate that description on the atomic time scale is unnecessary. What is needed in fact is a reduced description that focuses attention on the properties that control function in technology and biology. This reduced description needs to describe channel structure on the atomic scale of distance, in all likelihood, but it needs to describe flows and reservoirs only on the macroscopic scale.

Reduced descriptions of this type are familiar in engineering where they are called device equations. Semiconductor manufacturers produce the device behaviors they need by choosing particular structures of permanent charge, using as little atomic structure as possible, so cost is minimized. Device behaviors are described by device equations. It is device equations that we seek as we try to understand and control ion channels (and molecular machines of biology in general).

Device equations are most useful when they predict complex behaviors realistically while using only a few parameters with fixed values (that do not need to be changed to describe the complex behaviors). Fortunately, electrodiffusion allows rich behavior with simple device equations and a fixed set of parameters. Remarkably, the diverse (technologically important) behavior of transistors can be described by simple conservation laws and constitutive relations, the Poisson–Nernst–Planck (PNP) equations using fixed values of parameters. A single transistor can behave as many different devices, each with its own device equation, and this rich behavior can be described quite well by the PNP equations with a fixed set of parameters. Different values of the boundary potentials (i.e., power supply voltages) move the solution of the equations into different domains, each with its own device equation.

The PNP system of equations has been analyzed mathematically to some extent, but the equations have been simulated and computed to a much larger extent [18, 6,

16, 38, 49, 4, 17, 66, 36, 37, 40, 2, 20, 31, 55, 56, 1, 13, 70, 65, 29, 30, 43, 7]. Computational and experimental experience with a variety of PNP-like systems shows that the existing mathematical analysis is unsatisfactory. It is clear from these simulations that macroscopic reservoirs must be included in the mathematical formulation to describe the actual behavior of channels (or useful transistors) [60, 32, 31, 12, 59, 57, 58, 34, 29]. Macroscopic boundary conditions that describe such reservoirs introduce boundary layers of concentration and charge. If those boundary layers reach into the part of the device performing atomic control, they dramatically affect its behavior. Boundary layers of charge are particularly likely to create artifacts over long distances because the electric field spreads a long way. Indeed, transistors, channels, transporters, and receptors are actually built so that the contacts, electrodes, and control systems that maintain the reservoirs are quite distant and distinct from the channel.

In this paper, we construct and analyze the minimal model that includes reservoirs and channels and start the study of its mathematical properties. We begin with simple setups and conditions using geometric singular perturbation theory to extract powerful results. In particular, we consider three regions, two of which are reservoirs, and one of which is the narrow channel (with permanent charge, i.e., doping). And we consider only two species of current carriers. Nonetheless, we find quite complex behavior showing clearly that the reservoirs are inextricably linked to the channel and cannot be replaced by simple boundary conditions. We find general properties of the system and hints that somewhat more complicated systems (with several regions of permanent charge of different density and/or sign) carrying multiple ionic species (with different valence, i.e., with different permanent charge on each type of ion) may have quite rich behavior. Such rich behavior is apparent in biology where channels switch ("gate") between different values of current (one value nearly zero) and where transporters couple the flow of different types of ions in an extremely important, quite robust, but nearly unknown way.

The rest of the paper is organized as follows. In section 2, we begin with a description of a three-dimensional PNP system as the model for ion flow through an ion channel and discuss a one-dimensional reduction as the maximal radius of cross-sections of the channel approaching zero. We then identify the problem to be studied in this paper: steady-states of boundary value problems of the one-dimensional PNP system. In section 3, we cast our problem in the language of geometric singular perturbation theory. By introducing new dependent variables, we write the PNP system as a singularly perturbed system of first order equations. Making use of the inner and outer limiting systems, we then construct singular orbits for the PNP boundary value problem. In section 4, we apply geometric singular perturbation theory to show that, for small $\epsilon > 0$, there is a true solution shadowing each singular orbit. We conclude the paper by a general remark in section 5.

2. Three-dimensional model PNP system and a one-dimensional reduction. We now briefly describe the model PNP system of equations. As discussed above, the key features of an ion channel are the shape of its pore and the distribution of the permanent charge along its interior wall. As a first approximation, we consider a special ion channel modeled by

$$\Omega_\mu = \{(x, y, z) : 0 < x < 1,\ y^2 + z^2 < g^2(x, \mu)\},$$

where g is a smooth function satisfying

$$g(x, 0) = 0 \ \text{ and } \ g_0(x) = \frac{\partial g}{\partial \mu}(x, 0) > 0 \ \text{ for } \ x \in [0, 1].$$

The small parameter μ measures the maximal radius of cross-sections of the channel. The boundary $\partial\Omega_\mu$ of Ω_μ consists of three portions:

$$\begin{aligned}\mathcal{L}_\mu &= \{(x,y,z) \in \partial\Omega_\mu : x = 0\},\\ \mathcal{R}_\mu &= \{(x,y,z) \in \partial\Omega_\mu : x = 1\},\\ \mathcal{M}_\mu &= \{(x,y,z) \in \partial\Omega_\mu : y^2 + z^2 = g^2(x,\mu)\}.\end{aligned}$$

Here, $\mathcal{L}_\mu$ and $\mathcal{R}_\mu$ are viewed as the two ends of the reservoirs and $\mathcal{M}_\mu$ the wall of the channel and the reservoirs.

Then the model employed for flow through the channel is the PNP system (see [5] for a derivation from Boltzmann transport equation; see [66] for a derivation including correlations from coupled Langevin–Poisson equations; see [11, p. 773, eq. 26.64] for the classical description of the system at thermodynamic equilibrium, when all fluxes are zero):

$$\begin{aligned}\Delta\phi &= -\lambda\left(\sum_{i=1}^{n}\alpha_i c_i + Q\right),\\ \frac{\partial c_i}{\partial t} &= D_i\nabla\cdot(\nabla c_i + \alpha_i c_i\nabla\phi),\end{aligned} \tag{1}$$

where ϕ is the electric potential; c_i's are the concentrations of the n species, and α_i's are the valences, i.e., charge on one ion; D_i's are the diffusion constants; λ is the Debye number; and Q is the distribution of the permanent charge along the interior wall of the channel.

As mentioned in the introduction, the concentrations of the ions and the electrical potential in the reservoirs are nearly constants, and the wall of the channel is assumed to be perfectly insulated. We thus assume the following boundary conditions:

$$\begin{gathered}\phi|_{\mathcal{L}_\mu} = \nu_0, \quad \phi|_{\mathcal{R}_\mu} = 0, \quad c_i|_{\mathcal{L}_\mu} = L_i, \quad c_i|_{\mathcal{R}_\mu} = R_i,\\ \frac{\partial\phi}{\partial\mathbf{n}}\Big|_{\mathcal{M}_\mu} = \frac{\partial c_k}{\partial\mathbf{n}}\Big|_{\mathcal{M}_\mu} = 0,\end{gathered} \tag{2}$$

where ν_0, L_i, R_i are constants and $\mathbf{n}$ is the outward unit normal vector to $\mathcal{M}_\mu$.

We remark that, typically in the reservoirs, one imposes electroneutrality conditions: $\alpha L_1 - \beta L_2 = 0$ and $\alpha R_1 - \beta R_2 = 0$. In this case, there will be no boundary layers at the two ends although there will be internal layers *where* the permanent charge Q jumps. For mathematical interest, we use the slightly more general boundary conditions.

In [52], for $n = 2$ with $Q = 0$, we obtained the following limiting one-dimensional PNP system as $\mu \to 0$:

$$\begin{aligned}\frac{1}{g_0^2}\frac{\partial}{\partial x}\left(g_0^2\frac{\partial}{\partial x}\phi\right) &= -\lambda(\alpha_1 c_1 + \alpha_2 c_2),\\ \frac{\partial c_1}{\partial t} &= \frac{D_1}{g_0^2}\frac{\partial}{\partial x}\left(g_0^2\frac{\partial}{\partial x}c_1 + \alpha_1 c_1 g_0^2\frac{\partial}{\partial x}\phi\right),\\ \frac{\partial c_2}{\partial t} &= \frac{D_2}{g_0^2}\frac{\partial}{\partial x}\left(g_0^2\frac{\partial}{\partial x}c_2 + \alpha_2 c_2 g_0^2\frac{\partial}{\partial x}\phi\right)\end{aligned} \tag{3}$$

on $x \in (0,1)$ with the boundary conditions

$$\phi(t,0) = \nu_0, \quad \phi(t,1) = 0, \quad c_i(t,0) = L_i, \quad c_i(t,1) = R_i. \tag{4}$$

In particular, we showed that the attractors $\mathcal{A}_\mu$ of (1) and (2) are upper semicontinuous at $\mu = 0$ to the attractor $\mathcal{A}_0$ of (3) and (4). One-dimensional PNP systems of the form (3) also arise in treatments based on the density functional theory of statistical mechanics [31]. The motivation for such a mathematical treatment is that, first of all, the one-dimensional system is much simpler; second, if the one-dimensional limiting system is structurally stable (i.e., if the global dynamics is robust), then the dynamics for the system on the three-dimensional domain with small μ is essentially the same as that of the limiting one-dimensional system. There is a well-established framework for verification of structural stability although it is by no means trivial. A key step is to understand the behavior of the steady-state of the limiting one-dimensional system.

In light of the above result and discussion, we will then study steady-states of the one-dimensional PNP system for two species of current carriers with valences $\alpha > 0$ and $-\beta < 0$, including now a permanent charge:

$$\begin{aligned} \epsilon^2 h^{-1}(x)\frac{d}{dx}\left(h(x)\frac{d}{dx}\phi\right) &= -(\alpha c_1 - \beta c_2 + Q(x)), \quad \frac{dJ_i}{dx} = 0, \\ h(x)\frac{dc_1}{dx} + \alpha c_1 h(x)\frac{d\phi}{dx} &= -J_1, \\ h(x)\frac{dc_2}{dx} - \beta c_2 h(x)\frac{d\phi}{dx} &= -J_2, \end{aligned} \tag{5}$$

with the boundary conditions

$$\phi(0) = \nu_0, \quad c_i(0) = L_i; \quad \phi(1) = 0, \quad c_i(1) = R_i. \tag{6}$$

Here J_i is the total flux of the ith ion, $Q(x)$ is the permanent charge along the channel, $h(x) = g_0^2(x)$, and ϵ is related to λ via $\lambda = \epsilon^{-2}$.

Many mathematical papers have been written about the existence and uniqueness of solutions of the boundary value problems, and numerical algorithms have been developed to approximate solutions even for high-dimensional systems (see, e.g., [39, 42, 61, 44]). Under the assumption that $\epsilon \ll 1$, the problem can be viewed as a singular perturbation one. In particular, for $\alpha = \beta = 1$, $h(x) = 1$, and $Q(x) = 0$, the boundary value problem for the one-dimensional PNP system (5) was studied in [7] using the method of matched asymptotic expansions as well as numerical simulations, which provide a good quantitative understanding of the problem with one region without permanent charge. In [51], assuming $\epsilon \ll 1$ but for general α, β, $h(x) = 1$ and $Q(x) = 0$, the boundary value problem was treated using geometric theory for singularly perturbed problems (see, e.g., [27, 45, 47, 50]).

We use the geometric framework in paper [51] to investigate PNP systems with multiple regions of permanent charge and with multiple ions. A major difference of the model studied in this paper from those previously studied is the inclusion of multiple regions of permanent charge. The focus will be on the simple case of two ions and two reservoirs (i.e., two regions without permanent charge). The idea is to construct singular orbits for the boundary value problem and apply geometric singular perturbation theory to obtain, for $\epsilon > 0$ small, solutions near singular orbits. Issues of the existence and multiplicity of singular orbits are reduced to the properties of a set of nonlinear algebraic equations (43). To our surprise, for the simple case we study, multiple solutions for the boundary value problem are shown to exist. This contrasts to what was suspected in some early works (see, for example, [63, 64]) which expressed the (entirely reasonable) opinion that multiple solutions cannot occur for the simple

structure of permanent charge considered here. The set of equations (43) governs the multiplicity of solutions to the boundary value problem. We will thoroughly examine the set of algebraic equations in the future.

3. A dynamical system framework and a construction of singular orbits. We will rewrite the PNP system into a standard form for singularly perturbed systems and convert the boundary value problem to a connecting problem.

Denote the derivative with respect to x by overdot and introduce $u = \epsilon\dot{\phi}$ and $\tau = x$. System (5) becomes

$$\begin{aligned} &\epsilon\dot{\phi} = u, \quad \epsilon\dot{u} = \beta c_2 - \alpha c_1 - Q(\tau) - \epsilon\frac{h'(\tau)}{h(\tau)}u, \\ &\epsilon\dot{c}_1 = -\alpha c_1 u - \epsilon h^{-1}(\tau)J_1, \\ &\epsilon\dot{c}_2 = \beta c_2 u - \epsilon h^{-1}(\tau)J_2, \\ &\dot{J}_1 = \dot{J}_2 = 0, \quad \dot{\tau} = 1. \end{aligned} \tag{7}$$

We will treat system (7) as a dynamical system of phase space $\boldsymbol{R}^7$ with state variables $(\phi, u, c_1, c_2, J_1, J_2, \tau)$. The introduction of the extra state variable $\tau = x$ and the τ-equation seems to add complications to the problem, but this has a great advantage that we will explain shortly.

For $\epsilon > 0$, the rescaling $x = \epsilon\xi$ of the independent variable x gives rise to an equivalent system

$$\begin{aligned} &\phi' = u, \quad u' = \beta c_2 - \alpha c_1 - Q(\tau) - \epsilon\frac{h'(\tau)}{h(\tau)}u, \\ &c_1' = -\alpha c_1 u - \epsilon h^{-1}(\tau)J_1, \\ &c_2' = \beta c_2 u - \epsilon h^{-1}(\tau)J_2, \\ &J_1' = J_2' = 0, \quad \tau' = \epsilon, \end{aligned} \tag{8}$$

where prime denotes the derivative with respect to the variable ξ.

For $\epsilon > 0$, systems (7) and (8) have exactly the same phase portrait. But their limits at $\epsilon = 0$ are different and, very often, the two limiting systems provide complementary information on state variables. Therefore, the main task of singularly perturbed problems is to patch the limiting information together to form a solution for the entire $\epsilon > 0$ system. In terms of asymptotic expansions, system (7) and its limit at $\epsilon = 0$ will be used to study *outer or regular layer* solutions. We will call this system the *outer system* and its limit at $\epsilon = 0$ the *outer limit system*. System (8) and its limit at $\epsilon = 0$ will be used to study *inner or singular layer* solutions, and we call the system the *inner system* and its limit system at $\epsilon = 0$ the *inner limit system*. By *a singular orbit*, we mean a continuous and piecewise smooth curve in $\boldsymbol{R}^7$ that is a union of finitely many orbits of the outer limit system or inner limit system. In the theory of geometric singular perturbations, viewing the independent variables x and ξ as slow and fast time variables, the outer system is called the *slow system*, the inner system is called the *fast system*, and a singular orbit is a union of *slow* and *fast* orbits.

Let B_L and B_R be the subsets of the phase space $\boldsymbol{R}^7$ defined by

$$B_L = \{(\nu_0, u, L_1, L_2, J_1, J_2, 0) \in \boldsymbol{R}^7 : \text{ arbitrary } u, J_1, J_2\}, \tag{9}$$

$$B_R = \{(0, u, R_1, R_2, J_1, J_2, 1) \in \boldsymbol{R}^7 : \text{ arbitrary } u, J_1, J_2\}. \tag{10}$$

Then the boundary value problem is equivalent to a *connecting problem*, namely, finding a solution of (7) or (8) from B_L to B_R. To see this, suppose that $(\phi, u, c_1, c_2, J_1, J_2, \tau)$ is an orbit starting at a point on B_L and ending at a point on B_R. Due to the definitions of B_L and B_R, the starting point automatically has $x = \tau = 0$ with the assigned values for ϕ, c_1, and c_2 at $x = 0$, and the ending point has $x = \tau = 1$ with the assigned values for ϕ, c_1, and c_2 at $x = 1$. This solution $(\phi, u, c_1, c_2, J_1, J_2, \tau)$ satisfies the boundary condition automatically. Most importantly, when we *arbitrarily* rescale the independent variable x, the phase portrait will remain the same. Therefore, in searching for a solution from B_L to B_R, we can apply any rescaling of the independent variable x, even a rescaling that depends on each individual solution. (We will use a rescaling that is different for each solution when we derive the system (34) from system (33).) This is the significant advantage of introducing $\tau = x$ and $\dot{\tau} = 1$ promised earlier. The idea of converting a boundary value problem to a connecting one is now rather standard in applied dynamical systems.

In this paper, we will consider the case where the outer regions are reservoirs and the permanent charge is constant along the channel; that is,

$$Q(x) = \begin{cases} 0 & \text{for } 0 < x < a, \\ Q & \text{for } a < x < b, \\ 0 & \text{for } b < x < 1, \end{cases}$$

where Q is a constant. The intervals $[0, a]$ and $[b, 1]$ are the reservoirs, and the interval $[a, b]$ is the channel.

We will be interested in solutions of the connecting problem for system (7) or (8) from B_L to B_R defined in (9) and (10). In view of the jump of Q at $x = a$ and $x = b$, the best one can hope is that the solution is continuous and piecewise differentiable. We therefore require our solutions to be *continuous and piecewise differentiable*. The continuity of u implies that ϕ, c_1, and c_2 are differentiable. Our requirement is motivated by two considerations: (i) the dissipation present in the full PNP system (that includes time evolution) improves the regularity of solutions; in particular, the attractor contains regular solutions. Steady-state solutions, being in the attractor, should have the regularity imposed; (ii) if the requirement is relaxed, say, only requiring ϕ, c_1, c_2 to be piecewise differentiable, then one can preassign any value for (ϕ, c_1, c_2) at any partition points $0 < x_1 < x_2 < \cdots < x_k < 1$ and construct solutions over each subinterval and piece them together to create a solution over $[0, 1]$ with the preassigned values for (ϕ, c_1, c_2) at the partition points. (This assertion follows from the work in [7, 51].) It is clear that the only relevant solutions are those in which ϕ, c_1, c_2 are differentiable.

Our construction of a solution involves two main steps: the first step is to construct a singular orbit to the connecting problem, and the second step is to apply geometric singular perturbation theory to show that there is a unique solution near the singular orbit for $\epsilon > 0$ and small. Here we will give a detailed explanation for the first step and leave the explanation of the second step to section 4.

To construct a singular orbit, we first construct a singular orbit on each of the subinterval $[0, a]$, $[a, b]$, and $[b, 1]$. The reason to split the interval $[0, 1]$ into three subintervals is simply because the permanent charge $Q(x)$ has jumps at $x = a$ and $x = b$. To be able to construct a singular orbit on each subinterval, we need to preassign the values of ϕ, c_1, and c_2 at $x = a$ and $x = b$. Suppose, for the moment, that $\phi = \phi^a$, $c_1 = c_1^a$, and $c_2 = c_2^a$ at $x = a$, and that $\phi = \phi^b$, $c_1 = c_1^b$, and $c_2 = c_2^b$

at $x = b$. Those *six* unknown values

$$\phi^a,\ c_1^a,\ c_2^a;\ \phi^b,\ c_1^b,\ c_2^b \tag{11}$$

will be determined along our construction of a singular orbit on the whole interval $[0, 1]$.

1. On the left subinterval $[0, a]$ where $Q = 0$ or there is no permanent charge, we construct a singular orbit for the boundary value problem with (ϕ, c_1, c_2, τ) being

$$(\nu_0, L_1, L_2, 0) \text{ at } x = 0 \text{ and } (\phi^a, c_1^a, c_2^a, a) \text{ at } x = a.$$

   The orbit consists of two boundary layers Γ_l^0 and Γ_l^a and one regular layer Λ_l. In particular, given (ϕ^a, c_1^a, c_2^a), the flux densities J_1^l, J_2^l and the value $u_l(a)$ are uniquely determined (see section 3.1).
2. On the middle subinterval $[a, b]$, we construct a singular orbit for the boundary value problem with (ϕ, c_1, c_2, τ) being

$$(\phi^a, c_1^a, c_2^a, a) \text{ at } x = a \text{ and } (\phi^b, c_1^b, c_2^b, b) \text{ at } x = b.$$

   The orbit consists of two boundary layers Γ_m^a and Γ_m^b and one regular layer Λ_m. In particular, given (ϕ^a, c_1^a, c_2^a) and (ϕ^b, c_1^b, c_2^b), the flux densities J_1^m, J_2^m and the values $u_m(a)$ and $u_m(b)$ are uniquely determined (see section 3.2).
3. On the right subinterval $[b, 1]$, we construct a singular orbit for the boundary value problem with (ϕ, c_1, c_2, τ) being

$$(\phi^b, c_1^b, c_2^b, b) \text{ at } x = b \text{ and } (0, R_1, R_2, 1) \text{ at } x = 1.$$

   The orbit again consists of two boundary layers Γ_r^b and Γ_r^1 and one regular layer Λ_r. In particular, given (ϕ^b, c_1^b, c_2^b), the flux densities J_1^r, J_2^r and the value $u_r(b)$ are uniquely determined (see section 3.3).
4. Finally, for a singular orbit on the whole interval $[0, 1]$, we require that

$$J_1^l = J_1^m = J_1^r, \quad J_2^l = J_2^m = J_2^r, \quad u_l(a) = u_m(a), \quad u_m(b) = u_r(b).$$

   This consists of *six* conditions. *The number of conditions is exactly the same as the number of unknown values in* (11) (see section 3.4).
   The qualitative properties of these six equations and conditions are of great importance. It turns out that they can have multiple solutions. Different solutions yield different amounts of current for otherwise identical conditions, suggesting that each level might correspond to a different functional state of a transporter, or a different gating state of a channel. Indeed, it seems likely that more complex systems than those considered here would be described by similar systems of equations with multiple solutions. Interesting and very important properties of channels and transporters—each corresponding to a quite distinct device with a quite distinct input-output relation and device equation—might arise this way in systems including Ca^{2+} or in systems with multiple regions of nonzero permanent charge, or in systems with branched, Y-shaped, or adjacent interacting channels.

Remark 3.1. We call Γ_l^a, Γ_m^a, Γ_m^b, and Γ_r^b boundary layers because, relative to each subinterval, they are boundary layers. But, relative to the whole interval $[0, 1]$, they should be termed *internal layers*.

3.1. Singular orbit on $[0, a]$ where $Q(x) = 0$. We consider the case with zero permanent charge on the subinterval $[0, a]$ because $[0, a]$ is viewed as one of the reservoirs. The nonzero Q over the subinterval $[a, b]$ will affect the solution on $[0, a]$ and on $[b, 1]$. This effect will show up when we impose *matching* conditions on ϕ^a, c_1^a, and c_2^a to construct the singular orbit over the *whole* interval $[0, 1]$.

Following the discussion above, we set $\phi(a) = \phi^a$, $c_1(a) = c_1^a$, and $c_2(a) = c_2^a$, where ϕ^a, c_i^a are unknown values to be determined later on. Now let

$$B_a = \{(\phi^a, u, c_1^a, c_2^a, J_1, J_2, a) \in \boldsymbol{R}^7 : u, J_i \text{ arbitrary}\}.$$

In this part, we will construct a singular orbit that connects B_L to B_a. Two boundary layers will be constructed in section 2.1.1 followed by the construction of the regular layer in section 2.1.2. The permanent charge Q is zero in both constructions.

If we set $\epsilon = 0$ in system (7) with $Q(x) = 0$, we get the outer limit system and, in particular, $u = 0$ and $\alpha c_1 = \beta c_2$. The set

$$\mathcal{Z}_l = \{u = 0, \alpha c_1 = \beta c_2\}$$

will be called *the outer manifold*. In the theory of geometric singular perturbations, $\mathcal{Z}_l$ is called *the slow manifold* because if x and ξ are viewed as time variables, the evolution on $\mathcal{Z}_l$ is characterized by the time variable ξ, which is slow.

Remark 3.2. In systems (7) and (8), there appear to be four fast equations and three slow equations. Typically, one would expect a three-dimensional slow manifold. But, in this specific problem, the slow manifold is five-dimensional. This fact indicates some degeneracy of the slow flow, which is reflected in sections 3.1.2 and 3.2.2. The exchange lemma applied in the proof of Theorem 4.1 in section 4 is still valid. In fact, it applies to singular perturbation problems of more general forms than standard ones (see, e.g., [46, p. 562, Remark 1]).

The geometric method for a construction of singular orbits on each of the subintervals $[0, a]$, $[a, b]$, and $[b, 1]$ is the same. Let us explain the approach for constructing the singular orbit that connects B_L to B_a on $[0, a]$ (see Figure 1). Generally, the outer manifold $\mathcal{Z}_l$ will not intersect B_L and B_a. Since every outer or regular layer orbit lies entirely on the outer manifold $\mathcal{Z}_l$, it will not intersect B_L and B_a; that is, it cannot satisfy the boundary conditions. Two boundary or inner layers need to be introduced to connect boundaries B_L and B_a with the outer layer solution on $\mathcal{Z}_l$. These boundary layers should satisfy the inner limit system. The boundary layer orbit Γ_l^0 at $x = 0$ will connect B_L to $\mathcal{Z}_l$. It must lie on the stable manifold $W^s(\mathcal{Z}_l)$; that is, it belongs to the intersection $M_L \cap W^s(\mathcal{Z}_l)$, where M_L is the collection of orbits starting from points on B_L. Similarly, the boundary layer Γ_l^a at $x = a$ will connect $\mathcal{Z}_l$ to B_a and it must lie on the unstable manifold $W^u(\mathcal{Z}_l)$; that is, it belongs to the intersection $M_l^a \cap W^u(\mathcal{Z}_l)$, where M_l^a is the collection of orbits starting from points on B_l^a.

The first step in the construction examines the stability of the outer manifold $\mathcal{Z}_l$ by linearizing along $\mathcal{Z}_l$. ($\mathcal{Z}_l$ is the set of equilibria of the inner limit system.) It turns out that the outer manifold $\mathcal{Z}_l$ has a stable manifold $W^s(\mathcal{Z}_l)$ and an unstable manifold $W^u(\mathcal{Z}_l)$. The next step is to check whether $W^s(\mathcal{Z}_l)$ intersects B_L and whether $W^u(\mathcal{Z}_l)$ intersects B_a. This requires concrete knowledge of the *global* behavior of $W^s(\mathcal{Z}_l)$ and $W^u(\mathcal{Z}_l)$, and the information from the linearization is not enough. Neither is abstract dynamical systems theory (since the inner limit system is *nonlinear*). Luckily, we discovered a complete set of integrals for the inner limit system (see Proposition 3.2). The set of integrals reflects the intrinsic *mathematical*

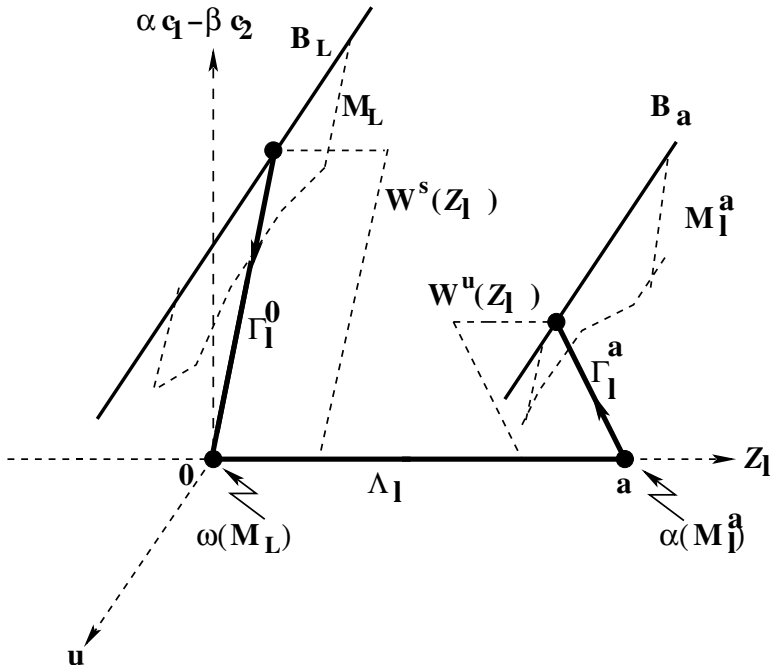


FIG. 1. *Schematic picture of the singular orbit (solid curves) on* $[0,a]$*: one left boundary layer* Γ_l^0*, one regular layer* Λ_l*, and one right boundary layer* Γ_l^a*.*

structure of this particular electrodiffusion system, the channel problem. This mathematical special structure implies particular specific physical and chemical properties of the ion channel. It is irresistible, albeit speculative, to suspect that the special mathematical structure produces biologically important properties of the channel. In that sense, the mathematical structure of the problem provides one possible "device equation" for the channel system.

It is this set of integrals in Proposition 3.2 that allows us to give a complete, global description of the inner limit dynamics; in particular, we are able to establish the required intersections $M_L \cap W^s(\mathcal{Z}_l)$ and $M_l^a \cap W^u(\mathcal{Z}_l)$ and are also able to identify the so-called ω-limit set $\omega(M_L \cap W^s(\mathcal{Z}_l))$ and the α-limit set $\alpha(M_l^a \cap W^u(\mathcal{Z}_l))$ of the intersections. The intersections give the set of candidates for the boundary layers (consisting of two parameter families of inner orbits parameterized by J_1 and J_2). The foot points $\omega(M_L \cap W^s(\mathcal{Z}_l))$ and $\alpha(M_l^a \cap W^u(\mathcal{Z}_l))$ (each parameterized by J_1 and J_2 also) on $\mathcal{Z}_l$ provide the (reduced) boundary conditions for the outer solutions. It turns out there is only one outer orbit Λ_l that connects $\omega(M_L \cap W^s(\mathcal{Z}_l))$ to $\alpha(M_l^a \cap W^u(\mathcal{Z}_l))$ and also determines the pair (J_1, J_2) uniquely. The desired singular orbit connecting B_L to B_a on $[0,a]$ is formed by this outer orbit Λ_l together with the two boundary layers Γ_l^0 and Γ_l^a that are uniquely determined by the pair (J_1, J_2).

We remind the reader that the singular orbit to be constructed on this subinterval with zero permanent charge will *not* be complete *until* the unknowns in (11) are determined through matching conditions implicitly posed by the permanent charge Q on the *whole* interval $[0,1]$, including the channel region where the permanent charge is not zero. The entire system is coupled and must be solved together, suggesting the source of difficulties with earlier treatments, which tried to replace the reservoirs with

boundary conditions. The importance of the coupling of different intervals suggests that the shapes of antechambers commonly found in biological channels may be important to their function. It is interesting that synthetic nanochannels acquire some properties of biological channels when they are built with antechambers of specific shape [68, 69].

3.1.1. Inner dynamics on $[0, a]$: Boundary layers or inner solutions. We start with the examination of boundary layers on the interval $[0, a]$ where $Q = 0$. These will be studied using the inner limit system obtained by setting $\epsilon = 0$ in (8):

$$\begin{aligned} &\phi' = u, \quad u' = \beta c_2 - \alpha c_1, \\ &c_1' = -\alpha c_1 u, \\ &c_2' = \beta c_2 u, \\ &J_1' = J_2' = 0, \quad \tau' = 0. \end{aligned} \tag{12}$$

This inner limit system describes what a chemist would call (thermodynamic) equilibrium. The reader should be warned that the word "equilibrium" is used widely, albeit informally, in computational electronics to describe a system *not* at thermodynamic equilibrium, namely, a system in which the distribution of velocities is a displaced Maxwellian, with displacement given by the flux (in appropriate units). Only when the flux of every species is zero is the "equilibrium" of computational electronics a thermodynamic equilibrium.

The set of equilibria of (12), that is, the set of points at which the vector field of (12) vanishes, is precisely $\mathcal{Z}_l = \{u = 0, \alpha c_1 = \beta c_2\}$. The linearization at points $(\phi, 0, c_1, c_2, J_1, J_2, \tau) \in \mathcal{Z}_l$ is

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\alpha & \beta & 0 & 0 & 0 \\ 0 & -\alpha c_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta c_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

This linearization is similar to the Green–Kubo expansion used by physical chemists to describe a nonequilibrium system close to equilibrium [11, 48, 74, 76]. Of course, such a linearization is useful only around some specific (operating) point; here the thermodynamic operating point with zero fluxes. To study nonlinear behavior far from the thermodynamic operating point, one must do a linearization around other points, at which fluxes are not zero. Such analyses have not been attempted, as far as we know for the PNP system, or in physical chemistry in general, perhaps because the locations and properties of operating points other than the thermodynamic one are hard to specify simply. Linearization around general *nonequilibrium* operating points is a crucial method in electrical engineering and has been used to design nonlinear circuits since the invention of electron valves—i.e., vacuum tubes—in the 1930s.

The linearized system has five zero eigenvalues whose generalized eigenspace is the tangent space of the five-dimensional outer manifold $\mathcal{Z}_l$ of equilibria. The two other eigenvalues are $\pm\sqrt{(\alpha + \beta)\alpha c_1} \neq 0$ whose eigenvectors are not tangent to $\mathcal{Z}_l$.

In this sense, $\mathcal{Z}_l$ is called *normally hyperbolic*. The theory of normally hyperbolic invariant manifolds (e.g., [27]) states that

(i) there is a six-dimensional stable manifold $W^s(\mathcal{Z}_l)$ of $\mathcal{Z}_l$ that consists of points approaching $\mathcal{Z}_l$ in forward time;

(ii) there is a six-dimensional unstable manifold $W^u(\mathcal{Z}_l)$ of $\mathcal{Z}_l$ that consists of points approaching $\mathcal{Z}_l$ in backward time;

(iii) $\mathcal{Z}_l$ as well as $W^s(\mathcal{Z}_l)$ and $W^u(\mathcal{Z}_l)$ persists for $\epsilon > 0$ small; that is, for $\epsilon > 0$ small, there exist invariant manifolds $\mathcal{Z}_l^\epsilon$, $W^s(\mathcal{Z}_l^\epsilon)$, and $W^u(\mathcal{Z}_l^\epsilon)$, close to their counterparts.

What this result suggests is that, for a singular orbit connecting B_L to B_a, the boundary layer at $x = 0$ must lie in $M_L \cap W^s(\mathcal{Z}_l)$ and the boundary layer at $x = a$ must lie in $M_l^a \cap W^u(\mathcal{Z}_l)$, where M_L is the collection of orbits from B_L in forward time under the flow of system (12) and M_l^a is the collection of orbits from B_a in backward time under the flow of system (12). This is precisely what we will show.

DEFINITION 3.1. *A smooth function $H : \boldsymbol{R}^n \to \boldsymbol{R}$ is called* an integral *of system $\frac{d}{dt}z = f(z)$, $z \in \boldsymbol{R}^n$, if $\frac{d}{dt}[H(z(t))] = 0$ whenever $z(t)$ is a solution.*

For a system on $\boldsymbol{R}^n$, if there are $(n-1)$ (independent) integrals, then any orbits can be theoretically determined by the intersections of $(n-1)$ level sets of the integrals.

PROPOSITION 3.2. *System* (12) *has the following six integrals:*

$$H_1 = e^{\alpha\phi}c_1, \quad H_2 = e^{-\beta\phi}c_2, \quad H_3 = c_1 + c_2 - \frac{1}{2}u^2,$$
$$H_4 = J_1, \quad H_5 = J_2, \quad H_6 = \tau.$$

Proof. The proof can be verified directly. ∎

The reader seeking physical insight is reminded that α is the valence (i.e., charge) of the ions with number density c_1; $(-\beta)$ is the charge of the ions with number density c_2, $u = \epsilon\dot{\phi}$, $\tau = x$; and ϵ is the Debye length.

These integrals allow one to completely understand the boundary layers (at $x = 0, a$) and characterize landing points of boundary layers on the outer manifold $\mathcal{Z}_l$. The information on landing points is crucial because it provides the boundary conditions that allow the regular layer to connect boundary layers.

COROLLARY 3.3. (i) *Let $\phi = \phi^L$ be the unique solution of*

$$\alpha L_1 e^{\alpha(\nu_0-\phi)} - \beta L_2 e^{-\beta(\nu_0-\phi)} = 0, \quad \textit{that is, } \phi^L = \nu_0 - \frac{1}{\alpha+\beta}\ln\frac{\beta L_2}{\alpha L_1},$$

and let

$$c_1^L = \frac{1}{\alpha}(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^L = \frac{1}{\beta}(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}}.$$

The stable manifold $W^s(\mathcal{Z}_l)$ intersects B_L transversally at points with

$$\begin{aligned} u_0 &= [sgn(\phi^L - \nu_0)]\sqrt{2(L_1+L_2) - 2(L_1 e^{\alpha(\nu_0-\phi^L)} + L_2 e^{-\beta(\nu_0-\phi^L)})} \\ &= [sgn(\alpha L_1 - \beta L_2)]\sqrt{2\left(L_1 + L_2 - \frac{\alpha+\beta}{\alpha\beta}(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}}\right)} \end{aligned} \tag{13}$$

and arbitrary J_i's, where sgn is the sign function (see Figure 1*).*

Let $\phi = \phi^{a,l}$ be the unique solution of

$$\alpha c_1^a e^{\alpha(\phi^a-\phi)} - \beta c_2^a e^{-\beta(\phi^a-\phi)} = 0, \quad \textit{that is, } \phi^{a,l} = \phi^a - \frac{1}{\alpha+\beta}\ln\frac{\beta c_2^a}{\alpha c_1^a},$$

and let

$$c_1^{a,l} = \frac{1}{\alpha}(\alpha c_1^a)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^a)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^{a,l} = \frac{1}{\beta}(\alpha c_1^a)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^a)^{\frac{\alpha}{\alpha+\beta}}.$$

The unstable manifold $W^u(\mathcal{Z}_l)$ intersects B_a transversally at points with

$$\begin{aligned} u_l(a) &= [sgn(\phi^a - \phi^{a,l})]\sqrt{2(c_1^a + c_2^a) - 2(c_1^a e^{\alpha(\phi^a-\phi^{a,l})} + c_2^a e^{-\beta(\phi^a-\phi^{a,l})})} \\ &= [sgn(\beta c_2^a - \alpha c_1^a)]\sqrt{2\left(c_1^a + c_2^a - \frac{\alpha+\beta}{\alpha\beta}(\alpha c_1^a)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^a)^{\frac{\alpha}{\alpha+\beta}}\right)} \end{aligned} \tag{14}$$

and arbitrary J_i's (see Figure 1*).*

(ii) *Potential boundary layers Γ_l^0 at $x = 0$ are determined up to (J_1, J_2) as follows: the ϕ-component satisfies the Hamiltonian system*

$$\phi'' + \alpha L_1 e^{\alpha(\nu_0-\phi)} - \beta L_2 e^{-\beta(\nu_0-\phi)} = 0,$$

together with $\phi(0) = \nu_0$ and $\phi(\xi) \to \phi^L$ as $\xi \to \infty$, $u(\xi) = \phi'(\xi)$, and

$$c_1(\xi) = L_1 e^{\alpha(\nu_0-\phi(\xi))}, \quad c_2(\xi) = L_2 e^{-\beta(\nu_0-\phi(\xi))}.$$

Similarly, potential boundary layers Γ_l^a at $x = a$ are determined in the following way: the ϕ-component satisfies the Hamiltonian system

$$\phi'' + \alpha c_1^a e^{\alpha(\phi^a-\phi)} - \beta c_2^a e^{-\beta(\phi^a-\phi)} = 0,$$

together with $\phi(0) = \phi^a$ and $\phi(\xi) \to \phi^{a,l}$ as $\xi \to -\infty$, $u(\xi) = \phi'(\xi)$, and

$$c_1(\xi) = c_1^a e^{\alpha(\phi^a-\phi(\xi))}, \quad c_2(\xi) = c_2^a e^{-\beta(\phi^a-\phi(\xi))}.$$

(iii) *Let $N_L = M_L \cap W^s(\mathcal{Z}_l)$ and $N_l^a = M_l^a \cap W^u(\mathcal{Z}_l)$. Then,*

$$\omega(N_L) = \left\{\left(\phi^L, 0, c_1^L, c_2^L, J_1, J_2, 0\right) : \text{ all } J_1, J_2\right\},$$

$$\alpha(N_l^a) = \left\{\left(\phi^{a,l}, 0, c_1^{a,l}, c_2^{a,l}, J_1, J_2, a\right) : \text{ all } J_1, J_2\right\},$$

where ϕ^L, c_1^L, c_2^L, $\phi^{a,l}$, $c_1^{a,l}$, and $c_2^{a,l}$ are given explicitly as in part (i)*.*

Proof. We provide a proof for the first part that is related to the boundary layer on the left in each statement.

Let $z(\xi) = (\phi(\xi), u(\xi), c_1(\xi), c_2(\xi), J_1(\xi), J_2(\xi), \tau(\xi))$ be a solution of system (12) with $z(0) \in B_L$ and $z(\xi) \in W^s(\mathcal{Z}_l)$. Then, $J_i(\xi) = J_i$, $\tau(\xi) = 0$ for all ξ, $z(\xi) \to z(\infty) = (\phi^L, 0, c_1^L, c_2^L, J_1, J_2, 0) \in \mathcal{Z}_l$ for some ϕ^L and c_i^L with $\alpha c_1^L = \beta c_2^L$, and

$$\phi(0) = \nu_0, \quad c_1(0) = L_1, \quad c_2(0) = L_2.$$

Using the integrals H_1 and H_2, we have

$$e^{\alpha\phi} c_1 = e^{\alpha\nu_0} L_1, \quad e^{-\beta\phi} c_2 = e^{-\beta\nu_0} L_2.$$

Therefore,

$$c_1 = L_1 e^{\alpha(\nu_0-\phi)}, \quad c_2 = L_2 e^{-\beta(\nu_0-\phi)}. \tag{15}$$

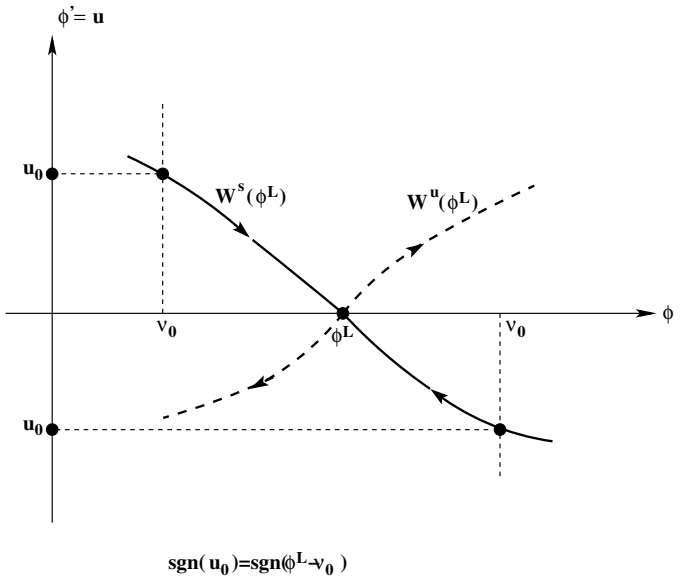


FIG. 2. *The stable manifold $W^s(\phi^L)$ of the equilibrium $(\phi^L, 0)$ is the solid curve, and the unstable manifold $W^u(\phi^L)$ is the dashed curve. The left branch of $W^s(\phi^L)$ has positive u-coordinates, and the right branch has negative u-coordinates; e.g., if $(\phi, u) \in W^s(\phi^L)$, then sign$[u]$ = sign$[\phi^L - \phi]$.*

Taking the limit as $\xi \to \infty$, we have

$$c_1^L = L_1 e^{\alpha(\nu_0 - \phi^L)}, \quad c_2^L = L_2 e^{-\beta(\nu_0 - \phi^L)}.$$

In view of the relation $\alpha c_1^L = \beta c_2^L$, one has

$$\alpha L_1 e^{\alpha(\nu_0 - \phi^L)} = \beta L_2 e^{-\beta(\nu_0 - \phi^L)} \text{ or } \phi^L = \nu_0 - \frac{1}{\alpha + \beta} \ln \frac{\beta L_2}{\alpha L_1}.$$

Hence,

$$c_1^L = \frac{1}{\alpha} (\alpha L_1)^{\frac{\beta}{\alpha+\beta}} (\beta L_2)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^L = \frac{1}{\beta} (\alpha L_1)^{\frac{\beta}{\alpha+\beta}} (\beta L_2)^{\frac{\alpha}{\alpha+\beta}}.$$

Since $\phi'' = \beta c_2 - \alpha c_1$, (15) implies that ϕ satisfies the Hamiltonian equation

$$\phi'' + \alpha L_1 e^{\alpha(\nu_0 - \phi)} - \beta L_2 e^{-\beta(\nu_0 - \phi)} = 0$$

with $\phi(0) = \nu_0$ and $\phi(\xi) \to \phi^L$ as $\xi \to \infty$. The Hamiltonian is

$$H(\phi, u) = \frac{u^2}{2} - L_1 e^{\alpha(\nu_0 - \phi)} + L_2 e^{-\beta(\nu_0 - \phi)}.$$

In terms of ϕ and $u = \phi'$, the equation becomes

$$\phi' = u, \quad u' = \beta L_2 e^{-\beta(\nu_0 - \phi)} - \alpha L_1 e^{\alpha(\nu_0 - \phi)}. \tag{16}$$

The Hamiltonian system has a unique equilibrium $(\phi^L, 0)$ with ϕ^L given above. If $W^s(\phi^L)$ is the stable manifold of $(\phi^L, 0)$, then it is the restriction of $W^s(\mathcal{Z}_l)$ to the (ϕ, u)-plane. In order to have $(\nu_0, u_0) \in W^s(\phi^L)$ (see Figure 2), $H(\phi^L, 0) = H(\nu_0, u_0)$

and one has the expression for u_0 in (13). To determine the sign of u_0, note that the left branch of the stable manifold $W^s(\phi^L)$ lies above the ϕ-axis and hence that $\nu_0 < \phi^L$ implies $u_0 > 0$; similarly, if $\nu_0 > \phi^L$, then $u_0 < 0$. □

Remark 3.3. We claim that the quantities under the square root in the displays (13) and (14) are nonnegative. In fact, quite interestingly, the nonnegativeness is equivalent to Young's inequality

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab \text{ for } a, b \geq 0, \ \frac{1}{p} + \frac{1}{q} = 1; \text{ “ = ” holds if and only if } a^p = b^q.$$

Take (13) for example. If we set

$$a = (\alpha L_1)^{\frac{\beta}{\alpha+\beta}}, \quad b = (\beta L_2)^{\frac{\alpha}{\alpha+\beta}}, \quad p = \frac{\alpha+\beta}{\beta}, \quad q = \frac{\alpha+\beta}{\alpha},$$

then

$$L_1 + L_2 - \frac{\alpha+\beta}{\alpha\beta}(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}} = \frac{\alpha+\beta}{\alpha\beta}\left(\frac{a^p}{p} + \frac{b^q}{q} - ab\right).$$

Thus, the quantity is always nonnegative and it is zero if and only if $\alpha L_1 = \beta L_2$.

3.1.2. Outer dynamics on $[0, a]$: Regular layers or outer solutions. We now construct regular layers or outer solutions on $\mathcal{Z}_l$ that connect $\omega(N_L)$ to $\alpha(N_l^a)$. We find that the outer flow on $\mathcal{Z}_l$ is itself a singular perturbation problem. To see this, we zoom in on an $O(\epsilon)$-neighborhood of $\mathcal{Z}_l$ by blowing up the u and $\alpha c_1 - \beta c_2$ coordinates; that is, we make a scaling $u = \epsilon p$ and $\beta c_2 - \alpha c_1 = \epsilon q$. System (7) becomes

$$\begin{aligned}
&\dot{\phi} = p, \quad \epsilon\dot{p} = q - \epsilon\frac{h'(\tau)}{h(\tau)}p,\\
&\epsilon\dot{q} = (\alpha(\alpha+\beta)c_1 + \epsilon\beta q)p - h^{-1}(\tau)(\beta J_2 - \alpha J_1),\\
&\dot{c}_1 = -\alpha c_1 p - h^{-1}(\tau)J_1,\\
&\dot{J}_i = 0, \quad \dot{\tau} = 1,
\end{aligned} \tag{17}$$

which is indeed a singular perturbation problem due to the factor ϵ in front of $\dot{p}$ and $\dot{q}$. Its limit, as $\epsilon \to 0$, is

$$\begin{aligned}
&\dot{\phi} = p, \quad 0 = q,\\
&0 = \alpha(\alpha+\beta)c_1 p - h^{-1}(\tau)(\beta J_2 - \alpha J_1),\\
&\dot{c}_1 = -\alpha c_1 p - h^{-1}(\tau)J_1,\\
&\dot{J}_i = 0, \quad \dot{\tau} = 1.
\end{aligned} \tag{18}$$

For this system, the outer manifold is

$$\mathcal{S}_l = \left\{ p = \frac{\beta J_2 - \alpha J_1}{\alpha(\alpha+\beta)h(\tau)c_1}, \ q = 0 \right\}.$$

The outer limit dynamics on $\mathcal{S}_l$ is

$$\begin{aligned}
&\dot{\phi} = \frac{\beta J_2 - \alpha J_1}{\alpha(\alpha+\beta)h(\tau)c_1},\\
&\dot{c}_1 = -\frac{\beta(J_1 + J_2)}{(\alpha+\beta)h(\tau)},\\
&\dot{J}_i = 0, \quad \dot{\tau} = 1.
\end{aligned} \tag{19}$$

Remark 3.4. Following the suggestion of one of the referees, we give a sketch of an alternative and more standard way of deriving the outer limit dynamics (19).

Introduce $\hat{q} = \beta c_2 - \alpha c_1$. In terms of the variables $(\phi, u, \hat{q}, c_1, J_i, \tau)$, system (8) (with $Q = 0$) becomes

$$\begin{aligned} &\phi' = u, \quad u' = \hat{q} - \epsilon \frac{h'(\tau)}{h(\tau)} u, \\ &\hat{q}' = (\alpha(\alpha+\beta)c_1 + \beta\hat{q})u - \epsilon h^{-1}(\tau)(\beta J_2 - \alpha J_1), \\ &c_1' = -\alpha c_1 u - \epsilon h^{-1}(\tau) J_1, \\ &J_i' = 0, \quad \tau' = \epsilon. \end{aligned} \tag{20}$$

For $\epsilon = 0$, the set $\{u = \hat{q} = 0\}$ is a normally hyperbolic invariant manifold consisting of equilibria. By Fenichel's theory, the manifold persists for $\epsilon > 0$ small and is given by

$$u = \epsilon A(\phi, c_1, J_i, \tau) + O(\epsilon^2), \quad \hat{q} = \epsilon B(\phi, c_1, J_i, \tau) + O(\epsilon^2).$$

Using the invariance of the manifold and substituting the above expressions for u and $\hat{q}$ into system (20), one obtains

$$B = O(\epsilon), \quad A = \frac{\beta J_2 - \alpha J_1}{\alpha(\alpha+\beta)h(\tau)c_1} + O(\epsilon).$$

System (20) on the perturbed invariant manifold can be obtained by substituting the expression of u and $\hat{q}$ with the approximations of A and B above. It reads as follows:

$$\begin{aligned} &\phi' = \epsilon \frac{\beta J_2 - \alpha J_1}{\alpha(\alpha+\beta)h(\tau)c_1} + O(\epsilon^2), \\ &c_1' = -\epsilon \frac{\beta(J_1 + J_2)}{(\alpha+\beta)h(\tau)} + O(\epsilon^2), \\ &J_i' = 0, \quad \tau' = \epsilon. \end{aligned} \tag{21}$$

The corresponding outer dynamics is

$$\begin{aligned} &\dot{\phi} = \frac{\beta J_2 - \alpha J_1}{\alpha(\alpha+\beta)h(\tau)c_1} + O(\epsilon), \\ &\dot{c}_1 = -\frac{\beta(J_1 + J_2)}{(\alpha+\beta)h(\tau)} + O(\epsilon), \\ &\dot{J}_i = 0, \quad \dot{\tau} = 1. \end{aligned} \tag{22}$$

Its limiting dynamics at $\epsilon = 0$ is exactly system (19).

The outer limit dynamics (33) in section 3.2.2 can also be derived this way.

The solution of (19) with the initial condition $(\phi^L, c_1^L, J_1, J_2, 0)$ that corresponds to the point $(\phi^L, 0, c_1^L, c_2^L, J_1, J_2, 0) \in \omega(N_L)$ is

$$\begin{aligned} \tau(x) &= x, \quad c_1(x) = c_1^L - \frac{\beta(J_1+J_2)}{\alpha+\beta} \int_0^x h^{-1}(s) ds, \\ \phi(x) &= \phi^L + \frac{\beta J_2 - \alpha J_1}{\alpha(\alpha+\beta)} \int_0^x h^{-1}(s) c_1^{-1}(s) ds \\ &= \phi^L - \frac{\beta J_2 - \alpha J_1}{\alpha\beta(J_1+J_2)} \int_0^x \frac{\dot{c}_1(s)}{c_1(s)} ds \quad ((19) \text{ is used here}) \\ &= \phi^L - \frac{\beta J_2 - \alpha J_1}{\alpha\beta(J_1+J_2)} \ln \frac{c_1(x)}{c_1^L}. \end{aligned}$$

Recall that we are looking for solutions that belong to $\alpha(N_l^a)$ when $\tau = a$. Evaluating the solution at $\tau = x = a$, we have

$$c_1^{a,l} = c_1^L - \frac{\beta(J_1+J_2)}{\alpha+\beta}\int_0^a h^{-1}(s)ds,$$
$$\phi^{a,l} = \phi^L - \frac{\beta J_2 - \alpha J_1}{\alpha\beta(J_1+J_2)}\ln\frac{c_1^{a,l}}{c_1^L};$$

in particular,

$$J_1 = \frac{(c_1^L - c_1^{a,l})}{\int_0^a h^{-1}(s)ds}\left(1 + \frac{\alpha(\phi^L - \phi^{a,l})}{\ln c_1^L - \ln c_1^{a,l}}\right),$$
$$J_2 = \frac{(c_2^L - c_2^{a,l})}{\int_0^a h^{-1}(s)ds}\left(1 - \frac{\beta(\phi^L - \phi^{a,l})}{\ln c_2^L - \ln c_2^{a,l}}\right). \tag{23}$$

We have used the relations $\alpha c_1^L = \beta c_2^L$ and $\alpha c_1^{a,l} = \beta c_2^{a,l}$ to get this more symmetric form for J_2.

The regular layer Λ_l is given by

$$\begin{aligned}
&\phi(x) = \phi^L - \frac{\beta J_2 - \alpha J_1}{\alpha\beta(J_1+J_2)}\ln\frac{c_1(x)}{c_1^L},\\
&u(x) = 0, \quad \alpha c_1(x) = \beta c_2(x),\\
&c_1(x) = c_1^L - \frac{\beta(J_1+J_2)}{\alpha+\beta}\int_0^x h^{-1}(s)ds,\\
&\tau(x) = x
\end{aligned} \tag{24}$$

with J_1 and J_2 determined by (23).

To summarize, for given values (ϕ^a, c_1^a, c_2^a), we have constructed a unique singular orbit on the left subinterval $[0,a]$ that connects B_L to B_a. It consists of two boundary layer orbits Γ_l^0 from the point $(\nu_0, u_0, L_1, L_2, J_1, J_2, 0) \in B_L$ to the point $(\phi^L, 0, c_1^L, c_2^L, J_1, J_2, 0) \in \omega(N_L) \subset \mathcal{Z}_l$ and Γ_l^a from the point $(\phi^{a,l}, 0, c_1^{a,l}, c_2^{a,l}, J_1, J_2, a) \in \alpha(N_l^a) \subset \mathcal{Z}_l$ to the point $(\phi^a, u_l(a), c_1^a, c_2^a, J_1, J_2, a) \in B_a$, and a regular layer Λ_l on $\mathcal{Z}_l$ that connects the two foot points $(\phi^L, 0, c_1^L, c_2^L, J_1, J_2, 0) \in \omega(N_L)$ and $(\phi^{a,l}, 0, c_1^{a,l}, c_2^{a,l}, J_1, J_2, a) \in \alpha(N_l^a)$ of the two boundary layers.

3.2. Singular orbits on $[a,b]$ with $Q(x) = Q$. We now construct a singular orbit on the subinterval $[a,b]$ viewed as the channel where the permanent charge $Q(x) = Q$ is a nonzero constant. The construction is nearly the same as that for singular orbits on $[0,a]$.

We set $\phi(b) = \phi^b$, $c_1(b) = c_1^b$, and $c_2(b) = c_2^b$, where ϕ^b, c_i^b are unknowns to be determined later. Let

$$B_b = \{(\phi^b, u, c_1^b, c_2^b, J_1, J_2, b) \in \boldsymbol{R}^7 : \text{ arbitrary } u, J_1, J_2\}.$$

The singular orbit to be constructed will be a connecting orbit from B_a to B_b over $[a,b]$.

3.2.1. Inner dynamics on $[a,b]$: Boundary layers or inner solutions. By setting $\epsilon = 0$ in system (7) with $Q(x) = Q$, we get $u = 0$ and $\alpha c_1 + Q = \beta c_2$. The outer manifold is

$$\mathcal{Z}_m = \{u = 0, \alpha c_1 + Q = \beta c_2\}.$$

In terms of ξ, we obtain the inner system of (7):

$$\begin{aligned}
&\phi' = u, \quad u' = \beta c_2 - \alpha c_1 - Q - \epsilon \frac{h'(\tau)}{h(\tau)} u,\\
&c_1' = -\alpha c_1 u - \epsilon h^{-1}(\tau) J_1,\\
&c_2' = \beta c_2 u - \epsilon h^{-1}(\tau) J_2,\\
&J_1' = J_2' = 0, \quad \tau' = \epsilon.
\end{aligned} \tag{25}$$

The limiting system at $\epsilon = 0$ is

$$\begin{aligned}
&\phi' = u, \quad u' = \beta c_2 - \alpha c_1 - Q,\\
&c_1' = -\alpha c_1 u,\\
&c_2' = \beta c_2 u,\\
&J_1' = J_2' = 0, \quad \tau' = 0.
\end{aligned} \tag{26}$$

The set of equilibria of (26) is precisely $\mathcal{Z}_m$, and $\mathcal{Z}_m$ is normally hyperbolic with a six-dimensional stable manifold $W^s(\mathcal{Z}_m)$ and a six-dimensional unstable manifold $W^u(\mathcal{Z}_m)$. The manifolds $\mathcal{Z}_m$, $W^s(\mathcal{Z}_m)$, and $W^s(\mathcal{Z}_m)$ persist for $\epsilon > 0$ small.

PROPOSITION 3.4. (i) *System* (26) *has the following six integrals:*

$$\begin{aligned}
&H_1 = e^{\alpha\phi} c_1, \quad H_2 = e^{-\beta\phi} c_2, \quad H_3 = c_1 + c_2 - \frac{1}{2}u^2 - Q\phi,\\
&H_4 = J_1, \quad H_5 = J_2, \quad H_6 = \tau.
\end{aligned}$$

(ii) *Let $\phi = \phi^{a,m}$ be the unique solution of*

$$\alpha c_1^a e^{\alpha(\phi^a - \phi)} - \beta c_2^a e^{-\beta(\phi^a - \phi)} + Q = 0, \tag{27}$$

and let

$$c_1^{a,m} = e^{\alpha(\phi^a - \phi^{a,m})} c_1^a, \quad c_2^{a,m} = e^{-\beta(\phi^a - \phi^{a,m})} c_2^a.$$

The stable manifold $W^s(\mathcal{Z}_m)$ intersects B_a transversally at points with

(28)

$$\begin{aligned}
&u_m(a)\\
&= [sgn(\phi^{a,m} - \phi^a)] \sqrt{2c_1^a(1 - e^{\alpha(\phi^a - \phi^{a,m})}) + 2c_2^a(1 - e^{-\beta(\phi^a - \phi^{a,m})}) - 2Q(\phi^a - \phi^{a,m})}.
\end{aligned}$$

and arbitrary J_i's.

Let $\phi = \phi^{b,m}$ be the unique solution of

$$\alpha c_1^b e^{\alpha(\phi^b - \phi)} - \beta c_2^b e^{-\beta(\phi^b - \phi)} + Q = 0, \tag{29}$$

and let

$$c_1^{b,m} = e^{\alpha(\phi^b - \phi^{b,m})} c_1^b, \quad c_2^{b,m} = e^{-\beta(\phi^b - \phi^{b,m})} c_2^b.$$

The unstable manifold $W^u(\mathcal{Z}_m)$ intersects B_b transversally at points with

(30)

$$u_m(b) = [sgn(\phi^b - \phi^{b,m})]\sqrt{2c_1^b(1 - e^{\alpha(\phi^b-\phi^{b,m})}) + 2c_2^b(1 - e^{-\beta(\phi^b-\phi^{b,m})}) - 2Q(\phi^b - \phi^{b,m})}.$$

and arbitrary J_i's.

(iii) *Potential boundary layers Γ_m^a at $x = a$ can be determined in the following way: the ϕ-component satisfies the Hamiltonian system*

$$\phi'' + \alpha c_1^a e^{\alpha(\phi^a-\phi)} - \beta c_2^a e^{-\beta(\phi^a-\phi)} + Q = 0,$$

together with $\phi(0) = \phi^a$ and $\phi(\xi) \to \phi^{a,m}$ as $\xi \to \infty$, $u(\xi) = \phi'(\xi)$, and

$$c_1(\xi) = c_1^a e^{\alpha(\phi^a-\phi(\xi))}, \quad c_2(\xi) = c_2^a e^{-\beta(\phi^a-\phi(\xi))}.$$

Similarly, potential boundary layers Γ_m^b at $x = b$ can be determined in the following way: the ϕ-component satisfies the Hamiltonian system

$$\phi'' + \alpha c_1^b e^{\alpha(\phi^b-\phi)} - \beta c_2^b e^{-\beta(\phi^b-\phi)} + Q = 0,$$

together with $\phi(0) = \phi^b$ and $\phi(\xi) \to \phi^{b,m}$ as $\xi \to -\infty$, $u(\xi) = \phi'(\xi)$, and

$$c_1(\xi) = c_1^b e^{\alpha(\phi^b-\phi(\xi))}, \quad c_2(\xi) = c_2^b e^{-\beta(\phi^b-\phi(\xi))}.$$

(iv) *Let $N_m^a = M_m^a \cap W^s(\mathcal{Z}_m)$ and $N_m^b = M_m^b \cap W^u(\mathcal{Z}_m)$, where M_m^a is the collection of orbits from B_a in forward time under the flow* (26) *and M_m^b is the collection of orbits from B_b in backward time under the flow* (26)*. Then,*

$$\omega(N_m^a) = \{(\phi^{a,m}, 0, c_1^{a,m}, c_2^{a,m}, J_1, J_2, a) : \text{ all } J_i\},$$
$$\alpha(N_m^b) = \{(\phi^{b,m}, 0, c_1^{b,m}, c_2^{b,m}, J_1, J_2, b) : \text{ all } J_i\}.$$

Remark 3.5. To show that the quantity under the square root in the display (28) is nonnegative, we assume $c_1^a > 0$ and $c_2^a > 0$ for the moment and let

$$f(x) = c_1^a + c_2^a - c_1^a e^{\alpha(\phi^a-x)} - c_2^a e^{-\beta(\phi^a-x)} - Q(\phi^a - x).$$

Then,

$$f'(x) = \alpha c_1^a e^{\alpha(\phi^a-x)} - \beta c_2^a e^{-\beta(\phi^a-x)} + Q$$

and

$$f''(x) = -\alpha^2 c_1^a e^{\alpha(\phi^a-x)} - \beta^2 c_2^a e^{-\beta(\phi^a-x)} < 0.$$

Therefore $f(x)$ is concave downward. Note that $f'(x) \to +\infty$ as $x \to -\infty$ and $f'(x) \to -\infty$ as $x \to +\infty$. Hence, $f(x)$ has a unique critical point and it must have a global maximum at this critical point. Since $x = \phi_m^a$ is the critical point, we have

$$f(\phi_m^a) \geq f(\phi^a) = 0.$$

By continuity, we have $f(\phi_m^a) \geq 0$ even if $c_1^a = 0$ and/or $c_2^a = 0$. Similarly, the quantity under the square root in the display (30) is nonnegative.

3.2.2. Outer dynamics on $[a, b]$: Regular layers or outer solutions. We now study the flow in the vicinity of the outer manifold $\mathcal{Z}_m$. Following the treatment of the outer flow on $\mathcal{Z}_l$ in section 3.1.2 (see also Remark 3.4), we make a scaling $u = \epsilon p$ and $\beta c_2 - \alpha c_1 - Q = \epsilon q$. System (7) becomes

$$
\begin{aligned}
&\dot{\phi} = p, \quad \epsilon \dot{p} = q - \epsilon \frac{h'(\tau)}{h(\tau)} p,\\
&\epsilon \dot{q} = ((\alpha+\beta)\alpha c_1 + \beta Q + \epsilon \beta q)p - h^{-1}(\tau)(\beta J_2 - \alpha J_1),\\
&\dot{c}_1 = -\alpha c_1 p - h^{-1}(\tau) J_1,\\
&\dot{J}_1 = \dot{J}_2 = 0, \quad \dot{\tau} = 1.
\end{aligned}
\tag{31}
$$

Its limit, as $\epsilon \to 0$, is

$$
\begin{aligned}
&\dot{\phi} = p, \quad 0 = q,\\
&0 = ((\alpha+\beta)\alpha c_1 + \beta Q)p - h^{-1}(\tau)(\beta J_2 - \alpha J_1),\\
&\dot{c}_1 = -\alpha c_1 p - h^{-1}(\tau) J_1,\\
&\dot{J}_i = 0, \quad \dot{\tau} = 1.
\end{aligned}
\tag{32}
$$

For this system, the outer manifold is

$$
\mathcal{S}_m = \left\{ p = \frac{\beta J_2 - \alpha J_1}{h(\tau)((\alpha+\beta)\alpha c_1 + \beta Q)},\ q = 0 \right\}.
$$

The outer limit dynamics on $\mathcal{S}_m$ is governed by system (32), which reads as follows:

$$
\begin{aligned}
\dot{\phi} &= \frac{\beta J_2 - \alpha J_1}{h(\tau)((\alpha+\beta)\alpha c_1 + \beta Q)},\\
\dot{c}_1 &= -\frac{(\beta J_2 - \alpha J_1)\alpha c_1}{h(\tau)((\alpha+\beta)\alpha c_1 + \beta Q)} - h^{-1}(\tau) J_1\\
&= -\frac{\alpha\beta(J_1 + J_2)c_1 + \beta Q J_1}{h(\tau)((\alpha+\beta)\alpha c_1 + \beta Q)},\\
\dot{J}_i &= 0, \quad \dot{\tau} = 1.
\end{aligned}
\tag{33}
$$

Since $h(\tau) > 0$ and $\beta c_2 = \alpha c_1 + Q > 0$, system (33) has the same phase portrait as that of the following system obtained by multiplying $h(\tau)((\alpha+\beta)\alpha c_1 + \beta Q)$ on the right-hand side of system (33) (here we see the reason why $\tau = x$ and $\dot{\tau} = 1$ were introduced into the analysis; see (7)):

$$
\begin{aligned}
&\frac{d}{dy}\phi = \beta J_2 - \alpha J_1,\\
&\frac{d}{dy} c_1 = -\alpha\beta(J_1 + J_2)c_1 - \beta Q J_1,\\
&\frac{d}{dy} J_i = 0, \quad \frac{d}{dy}\tau = h(\tau)((\alpha+\beta)\alpha c_1 + \beta Q).
\end{aligned}
\tag{34}
$$

The solution with the initial condition $(\phi^{a,m}, c_1^{a,m}, J_1, J_2, a)$ that corresponds to the point $(\phi^{a,m}, 0, c_1^{a,m}, c_2^{a,m}, J_1, J_2, a) \in \omega(N_m^a)$ is

$$
\begin{aligned}
&\phi(y) = \phi^{a,m} + (\beta J_2 - \alpha J_1)y,\\
&c_1(y) = e^{-\alpha\beta(J_1+J_2)y}c_1^{a,m} - \frac{QJ_1}{\alpha(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y}\right),\\
&\int_a^\tau h^{-1}(s)ds = (\alpha+\beta)\alpha\int_0^y c_1 ds + \beta Q y \qquad (35)\\
&\qquad = \frac{(\alpha+\beta)c_1^{a,m}}{\beta(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y}\right)\\
&\qquad\quad - \frac{(\alpha+\beta)QJ_1}{J_1+J_2}\left(y - \frac{1}{\alpha\beta(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y}\right)\right) + \beta Q y.
\end{aligned}
$$

We are looking for solutions to reach $\alpha(N_m^b)$; that is, whenever $\tau(y) = b$, we require $\phi(y) = \phi^{b,m}$ and $c_1(y) = c_1^{b,m}$. Assume $\tau(y_0) = b$ for some $y_0 > 0$. Then, $\phi(y_0) = \phi^{b,m}$ and $c_1(y_0) = c_1^{b,m}$, and hence,

$$
\begin{aligned}
&\phi^{b,m} = \phi^{a,m} + (\beta J_2 - \alpha J_1)y_0,\\
&c_1^{b,m} = e^{-\alpha\beta(J_1+J_2)y_0}c_1^{a,m} - \frac{QJ_1}{\alpha(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y_0}\right),\\
&\int_a^b h^{-1}(s)ds = \frac{(\alpha+\beta)c_1^{a,m}}{\beta(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y_0}\right) \qquad (36)\\
&\qquad - \frac{(\alpha+\beta)QJ_1}{J_1+J_2}\left(y_0 - \frac{1}{\alpha\beta(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y_0}\right)\right) + \beta Q y_0.
\end{aligned}
$$

System (36) is equivalent to

$$
\begin{aligned}
&\phi^{b,m} = \phi^{a,m} + (\beta J_2 - \alpha J_1)y_0,\\
&c_1^{b,m} = e^{-\alpha\beta(J_1+J_2)y_0}c_1^{a,m} - \frac{QJ_1}{\alpha(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y_0}\right), \qquad (37)\\
&J_1 + J_2 = \frac{\alpha(\alpha+\beta)(c_1^{a,m} - c_1^{b,m}) - \alpha\beta Q(\phi^{a,m} - \phi^{b,m})}{\alpha\beta\int_a^b h^{-1}(s)ds}.
\end{aligned}
$$

Therefore, the outer or regular layer solution Λ_m on $[a, b]$ is given by (35) with J_1 and J_2 determined by (37). Together with the boundary layers Γ_m^a and Γ_m^b in statement (iii) of Proposition 3.4, this gives the singular orbit on the interval $[a, b]$.

3.3. Singular orbits on $[b, 1]$ with $Q(x) = 0$. The construction of singular orbits on $[b, 1]$ is virtually identical to the construction of singular orbits on $[0, a]$ in section 3.1. We will state only the results for later use.

3.3.1. Inner dynamics on $[b, 1]$: Boundary layers or inner solutions. The inner limit system is

$$
\begin{aligned}
&\phi' = u, \quad u' = \beta c_2 - \alpha c_1,\\
&c_1' = -\alpha c_1 u,\\
&c_2' = \beta c_2 u, \qquad (38)\\
&J_1' = J_2' = 0, \quad \tau' = 0.
\end{aligned}
$$

The outer manifold is

$$\mathcal{Z}_r = \{u = 0, \alpha c_1 = \beta c_2\}.$$

It consists of equilibria of system (38) and is normally hyperbolic with a six-dimensional stable manifold $W^s(\mathcal{Z}_r)$ and a six-dimensional unstable manifold $W^u(\mathcal{Z}_r)$. Concerning the boundary layers, we have the following proposition.

PROPOSITION 3.5. (i) *System* (38) *has the following six integrals:*

$$H_1 = e^{\alpha\phi}c_1, \quad H_2 = e^{-\beta\phi}c_2, \quad H_3 = c_1 + c_2 - \frac{1}{2}u^2,$$
$$H_4 = J_1, \quad H_5 = J_2, \quad H_6 = \tau.$$

(ii) *Let* $\phi = \phi^{b,r}$ *be the unique solution of*

$$\alpha c_1^b e^{\alpha(\phi^b - \phi)} - \beta c_2^b e^{-\beta(\phi^b - \phi)} = 0, \quad \textit{that is, } \phi^{b,r} = \phi^b - \frac{1}{\alpha+\beta}\ln\frac{\beta c_2^b}{\alpha c_1^b},$$

and let

$$c_1^{b,r} = \frac{1}{\alpha}(\alpha c_1^b)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^b)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^{b,r} = \frac{1}{\beta}(\alpha c_1^b)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^b)^{\frac{\alpha}{\alpha+\beta}}.$$

The stable manifold $W^s(\mathcal{Z}_r)$ *intersects* B_b *transversally at points with*

$$u_r(b) = [sgn(\alpha c_1^b - \beta c_2^b)]\sqrt{2\left(c_1^b + c_2^b - \frac{\alpha+\beta}{\alpha\beta}(\alpha c_1^b)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^b)^{\frac{\alpha}{\alpha+\beta}}\right)} \tag{39}$$

and arbitrary J_i*'s.*

Let $\phi = \phi^R$ *be the unique solution of*

$$\alpha R_1 e^{-\alpha\phi} - \beta R_2 e^{\beta\phi} = 0, \quad \textit{that is, } \phi^R = -\frac{1}{\alpha+\beta}\ln\frac{\beta R_2}{\alpha R_1},$$

and let

$$c_1^R = \frac{1}{\alpha}(\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^R = \frac{1}{\beta}(\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}.$$

The unstable manifold $W^u(\mathcal{Z}_r)$ *intersects* B_R *transversally at points with*

$$u_1 = [sgn(\beta R_2 - \alpha R_1)]\sqrt{2\left(R_1 + R_2 - \frac{\alpha+\beta}{\alpha\beta}(\alpha R_1)^{\frac{\beta}{\alpha+\beta}}(\beta R_2)^{\frac{\alpha}{\alpha+\beta}}\right)} \tag{40}$$

and arbitrary J_i*'s.*

(iii) *Potential boundary layers* Γ_r^b *at* $x = b$ *can be determined in the following way: the* ϕ*-component satisfies the Hamiltonian system*

$$\phi'' + \alpha c_1^b e^{\alpha(\phi^b - \phi)} - \beta c_2^b e^{-\beta(\phi^b - \phi)} = 0,$$

together with $\phi(0) = \phi^b$ *and* $\phi(\xi) \to \phi^{b,r}$ *as* $\xi \to \infty$, $u(\xi) = \phi'(\xi)$, *and*

$$c_1(\xi) = c_1^b e^{\alpha(\phi^b - \phi(\xi))}, \quad c_2(\xi) = c_2^b e^{-\beta(\phi^b - \phi(\xi))}.$$

Similarly, potential boundary layers Γ_r^1 at $x = 1$ can be determined in the following way: the ϕ-component satisfies the Hamiltonian system

$$\phi'' + \alpha R_1 e^{-\alpha\phi} - \beta R_2 e^{\beta\phi} = 0,$$

together with $\phi(0) = 0$ and $\phi(\xi) \to \phi^R$ as $\xi \to -\infty$, $u(\xi) = \phi'(\xi)$, and

$$c_1(\xi) = R_1 e^{-\alpha\phi(\xi)}, \quad c_2(\xi) = R_2 e^{\beta\phi(\xi)}.$$

(iv) *Let $N_r^b = M_r^b \cap W^s(\mathcal{Z}_r)$ and $N_R = M_R \cap W^u(\mathcal{Z}_r)$, where M_r^b is the collection of orbits from B_b in forward time under the flow* (38) *and M_R is the collection of orbits from B_R in backward time under the flow* (38). *Then,*

$$\omega(N_r^b) = \{(\phi^{b,r}, 0, c_1^{b,r}, c_2^{b,r}, J_1, J_2, b) : \text{ all } J_i\},$$
$$\alpha(N_R) = \left\{\left(\phi^R, 0, c_1^R, c_2^R, J_1, J_2, 1\right) : \text{ all } J_i\right\}.$$

3.3.2. Outer dynamics on $[b, 1]$: Regular layers or outer solutions. We now examine the existence of regular layers or outer solutions that connect $\omega(N_r^b)$ to $\alpha(N_R)$. Following exactly the same analysis as in section 3.1.2, we find that the outer limit dynamics is

$$\begin{aligned} \dot{\phi} &= \frac{\beta J_2 - \alpha J_1}{(\alpha+\beta)\alpha h(\tau) c_1}, \\ \dot{c}_1 &= -\frac{\beta(J_1 + J_2)}{(\alpha+\beta)h(\tau)}, \\ \dot{J}_i &= 0, \quad \dot{\tau} = 1, \end{aligned} \tag{41}$$

and the outer solution Λ_r on $[b, 1]$ with the initial condition $(\phi^{b,r}, c_1^{b,r}, J_1, J_2, b)$ that corresponds to the point $(\phi^{b,r}, 0c_1^{b,r}, c_2^{b,r}, J_1, J_2, b) \in \omega(N_r^b)$ is given by

$$\begin{aligned} \phi(\xi) &= \phi^{b,r} - \frac{\beta J_2 - \alpha J_1}{\alpha\beta(J_1 + J_2)} \ln \frac{c_1(\xi)}{c_1^{b,r}}, \\ u(\xi) &= 0, \quad \alpha c_1(\xi) = \beta c_2(\xi), \\ c_1(\xi) &= c_1^{b,r} - \frac{\beta(J_1 + J_2)}{\alpha + \beta} \int_b^\xi h^{-1}(s) ds, \\ \tau(\xi) &= \xi. \end{aligned}$$

The outer solution Λ_r hits the point $(\phi^R, 0, c_1^R, c_2^R, J_1, J_2, 1) \in \alpha(N_R)$ if and only if

$$\begin{aligned} J_1 &= \frac{c_1^{b,r} - c_1^R}{\int_b^1 h^{-1}(s) ds} \left(1 + \frac{\alpha(\phi^{b,r} - \phi^R)}{\ln c_1^{b,r} - \ln c_1^R}\right), \\ J_2 &= \frac{c_2^{b,r} - c_2^R}{\int_b^1 h^{-1}(s) ds} \left(1 - \frac{\beta(\phi^{b,r} - \phi^R)}{\ln c_2^{b,r} - \ln c_2^R}\right). \end{aligned} \tag{42}$$

The outer solution Λ_r together with the inner solutions Γ_r^b and Γ_r^1 in statement (iii) of Proposition 3.5 gives the singular orbit on $[b, 1]$.

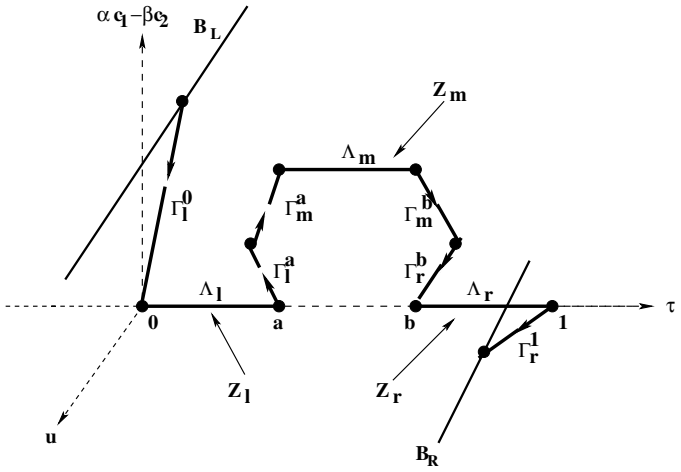


FIG. 3. *Schematic picture of the singular orbit (solid curves): left boundary layer Γ_l^0; right boundary layer Γ_r^1; four internal layers Γ_l^a, Γ_m^a, Γ_m^b, and Γ_r^b, and three regular layers Λ_l, Λ_m, and Λ_r.*

3.4. Matching and singular orbits on [0, 1]. A singular orbit on the whole interval $[0,1]$ will be the union of the singular orbits constructed on each of the subintervals (see Figure 3). The matching conditions are $u_l(a) = u_m(a)$, $u_m(b) = u_r(b)$, and J_1 and J_2 have to be the same on all subintervals; that is, from formulas (14), (23), (27), (28), (29), (30), (37), (39), and (42),

$$
\begin{aligned}
\alpha c_1^a e^{\alpha(\phi^a-\phi^{a,m})} - \beta c_2^a e^{-\beta(\phi^a-\phi^{a,m})} + Q &= 0,\\
\alpha c_1^b e^{\alpha(\phi^b-\phi^{b,m})} - \beta c_2^b e^{-\beta(\phi^b-\phi^{b,m})} + Q &= 0,\\
\frac{\alpha+\beta}{\beta} c_1^{a,l} &= c_1^a e^{\alpha(\phi^a-\phi^{a,m})} + c_2^a e^{-\beta(\phi^a-\phi^{a,m})} + Q(\phi^a-\phi^{a,m}),\\
\frac{\alpha+\beta}{\beta} c_1^{b,r} &= c_1^b e^{\alpha(\phi^b-\phi^{b,m})} + c_2^b e^{-\beta(\phi^b-\phi^{b,m})} + Q(\phi^b-\phi^{b,m}),\\
J_1 &= \frac{(c_1^L - c_1^{a,l})}{\int_0^a h^{-1}(s)ds}\left(1 + \frac{\alpha(\phi^L-\phi^{a,l})}{\ln c_1^L - \ln c_1^{a,l}}\right)\\
&= \frac{c_1^{b,r} - c_1^R}{\int_b^1 h^{-1}(s)ds}\left(1 + \frac{\alpha(\phi^{b,r}-\phi^R)}{\ln c_1^{b,r} - \ln c_1^R}\right),\\
J_2 &= \frac{(c_2^L - c_2^{a,l})}{\int_0^a h^{-1}(s)ds}\left(1 - \frac{\beta(\phi^L-\phi^{a,l})}{\ln c_2^L - \ln c_2^{a,l}}\right)\\
&= \frac{c_2^{b,r} - c_2^R}{\int_b^1 h^{-1}(s)ds}\left(1 - \frac{\beta(\phi^{b,r}-\phi^R)}{\ln c_2^{b,r} - \ln c_2^R}\right),\\
\phi^{b,m} &= \phi^{a,m} + (\beta J_2 - \alpha J_1)y_0,\\
c_1^{b,m} &= e^{-\alpha\beta(J_1+J_2)y_0} c_1^{a,m} - \frac{QJ_1}{\alpha(J_1+J_2)}\left(1 - e^{-\alpha\beta(J_1+J_2)y_0}\right),\\
J_1 + J_2 &= \frac{\alpha(\alpha+\beta)(c_1^{a,m} - c_1^{b,m}) - \alpha\beta Q(\phi^{a,m}-\phi^{b,m})}{\alpha\beta\int_a^b h^{-1}(s)ds},
\end{aligned}
\tag{43}
$$

where

$$
\begin{aligned}
c_1^L &= \frac{1}{\alpha}(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^L = \frac{1}{\beta}(\alpha L_1)^{\frac{\beta}{\alpha+\beta}}(\beta L_2)^{\frac{\alpha}{\alpha+\beta}},\\
c_1^{a,l} &= \frac{1}{\alpha}(\alpha c_1^a)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^a)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^{a,l} = \frac{1}{\beta}(\alpha c_1^a)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^a)^{\frac{\alpha}{\alpha+\beta}},\\
c_1^{b,r} &= \frac{1}{\alpha}(\alpha c_1^b)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^b)^{\frac{\alpha}{\alpha+\beta}}, \quad c_2^{b,r} = \frac{1}{\beta}(\alpha c_1^b)^{\frac{\beta}{\alpha+\beta}}(\beta c_2^b)^{\frac{\alpha}{\alpha+\beta}},\\
c_1^{a,m} &= e^{\alpha(\phi^a-\phi^{a,m})}c_1^a, \quad c_1^{b,m} = e^{\alpha(\phi^b-\phi^{b,m})}c_1^b.
\end{aligned}
$$

Recall that $h(x) = g_0^2(x)$, where $g_0(x)$ is the radius of the cross-section of the channel at x, Q is the concentration of the permanent charge over the interval $[a, b]$, (ϕ^a, c_1^a, c_2^a) and (ϕ^b, c_1^b, c_2^b) are the unknown values preassigned at $x = a$ and $x = b$, and J_1 and J_2 are the unknown values for the flux densities of the two types of ions.

There are also three auxiliary unknowns $\phi^{a,m}$, $\phi^{b,m}$, and y_0 in the set of equations (43). The total number of unknowns in (43) is eleven, which matches the total number of equations.

A qualitative important question is whether the set of nonlinear equations (43) has a unique solution. Next, we will consider a special case and demonstrate that (43) can have multiple solutions.

3.4.1. $\alpha = \beta = 1$, and $a = 1/3$, $b = 2/3$, and $h = 1$. We now consider a special case where $\alpha = \beta = 1$. It turns out that the nonlinear system of algebraic equations (43) in eleven unknowns can be reduced to a single algebraic equation with only one unknown. Further restrictions that $a = 1/3$, $b = 2/3$, and $h = 1$ will be posted later merely for simplicity.

Set $c_1^a c_2^a = A^2$, $c_1^b c_2^b = B^2$, $L_1 L_2 = L^2$, $R_1 R_2 = R^2$, and $Q = 2Q_0$. From the first two equations in (43), one has

$$
\begin{aligned}
\phi^a - \phi^{a,m} &= \ln \frac{\sqrt{Q_0^2 + A^2} - Q_0}{c_1^a},\\
\phi^b - \phi^{b,m} &= \ln \frac{\sqrt{Q_0^2 + B^2} - Q_0}{c_1^b}.
\end{aligned}
$$

System (43) becomes

$$
\begin{aligned}
A &= \sqrt{Q_0^2 + A^2} + Q_0 \ln \frac{\sqrt{Q_0^2 + A^2} - Q_0}{c_1^a},\\
B &= \sqrt{Q_0^2 + B^2} + Q_0 \ln \frac{\sqrt{Q_0^2 + B^2} - Q_0}{c_1^b},\\
J_1 &= \frac{L - A}{\int_0^a h^{-1}(s)ds} \cdot \frac{\nu_0 - \phi^a + \ln L_1 - \ln c_1^a}{\ln L - \ln A}\\
&= \frac{B - R}{\int_b^1 h^{-1}(s)ds} \cdot \frac{\phi^b + \ln c_1^b - \ln R_1}{\ln B - \ln R},
\end{aligned}
$$

$$
\begin{aligned}
J_2 &= \frac{L-A}{\int_0^a h^{-1}(s)ds}\left(2-\frac{\nu_0-\phi^a+\ln L_1-\ln c_1^a}{\ln L-\ln A}\right)\\
&= \frac{B-R}{\int_b^1 h^{-1}(s)ds}\left(2-\frac{\phi^b+\ln c_1^b-\ln R_1}{\ln B-\ln R}\right),\\
(J_2-J_1)y_0 &= \phi^b-\phi^a+\ln\frac{c_1^b(\sqrt{Q_0^2+A^2}-Q_0)}{c_1^a(\sqrt{Q_0^2+B^2}-Q_0)},\\
J_1+J_2 &= \frac{2(\sqrt{Q_0^2+A^2}-\sqrt{Q_0^2+B^2})-2Q_0(J_1-J_2)y_0}{\int_a^b h^{-1}(s)ds},\\
\sqrt{Q_0^2+B^2}-Q_0 &= e^{-(J_1+J_2)y_0}\left(\sqrt{Q_0^2+A^2}-Q_0\right)\\
&\quad -\frac{2Q_0J_1}{J_1+J_2}\left(1-e^{-(J_1+J_2)y_0}\right).
\end{aligned}
\tag{44}
$$

Add the J_1 and J_2 equations in (44) to get

$$
J_1+J_2 = 2\frac{L-A}{\int_0^a h^{-1}} = 2\frac{B-R}{\int_b^1 h^{-1}};\ \text{hence, } B=\frac{\int_b^1 h^{-1}}{\int_0^a h^{-1}}(L-A)+R.
$$

The first two equations in (44) give

$$
\begin{aligned}
c_1^a &= \left(\sqrt{Q_0^2+A^2}-Q_0\right)\exp\left\{\frac{\sqrt{Q_0^2+A^2}-A}{Q_0}\right\},\\
c_1^b &= \left(\sqrt{Q_0^2+B^2}-Q_0\right)\exp\left\{\frac{\sqrt{Q_0^2+B^2}-B}{Q_0}\right\}.
\end{aligned}
\tag{45}
$$

The first two equations together with $(J_2-J_1)y_0$ and the J_1+J_2 equations give

$$
J_1+J_2 = 2\frac{L-A}{\int_0^a h^{-1}} = 2\frac{B-R}{\int_b^1 h^{-1}} = 2\frac{A-B-Q_0(\phi^a-\phi^b)}{\int_a^b h^{-1}}.
$$

Hence,

$$
J_1+J_2 = 2\frac{L-R-Q_0(\phi^a-\phi^b)}{\int_0^1 h^{-1}},
$$

$$
\phi^b-\phi^a = \frac{(L-A)\int_0^1 h^{-1}-(L-R)\int_0^a h^{-1}}{Q_0\int_0^a h^{-1}},
\tag{46}
$$

and

$$
\begin{aligned}
(J_2-J_1)y_0 &= \phi^b-\phi^a-\ln\frac{\sqrt{Q_0^2+B^2}-Q_0}{c_1^b}+\ln\frac{\sqrt{Q_0^2+A^2}-Q_0}{c_1^a}\\
&= \frac{(L-A)\int_0^1 h^{-1}-(L-R)\int_0^a h^{-1}}{Q_0\int_0^a h^{-1}}+\ln\frac{(\sqrt{Q_0^2+A^2}-Q_0)c_1^b}{(\sqrt{Q_0^2+B^2}-Q_0)c_1^a}\\
&= \frac{(L-A)\int_a^b h^{-1}}{Q_0\int_0^a h^{-1}}+\frac{\sqrt{Q_0^2+B^2}-\sqrt{Q_0^2+A^2}}{Q_0}.
\end{aligned}
$$

Using

$$\frac{L-A}{\int_0^a h^{-1}} = \frac{B-R}{\int_b^1 h^{-1}}$$

and the second equality in the J_1 equation in (44), one has

$$\frac{\nu_0 - \phi^a + \ln L_1 - \ln c_1^a}{\ln L - \ln A} = \frac{\phi^b + \ln c_1^b - \ln R_1}{\ln B - \ln R}.$$

Hence,

$$\frac{\phi^b + \ln c_1^b - \ln R_1}{\ln B - \ln R} = \frac{\nu_0 + \phi^b - \phi^a + \ln(L_1 c_1^b) - \ln(R_1 c_1^a)}{\ln(BL) - \ln(AR)}.$$

The latter together with (46) and (45) gives

$$\begin{aligned}\phi^b &= \frac{\ln \frac{B}{R}}{\ln \frac{BL}{AR}} \left(\nu_0 + \ln \frac{L_1 c_1^b}{R_1 c_1^a} + \frac{(L-A)\int_0^1 h^{-1} - (L-R)\int_0^a h^{-1}}{Q_0 \int_0^a h^{-1}} \right) + \ln \frac{R_1}{c_1^b} \\ &= \frac{\ln \frac{B}{R}}{\ln \frac{BL}{AR}} \left(\nu_0 + \ln \frac{L_1(\sqrt{Q_0^2+B^2} - Q_0)}{R_1(\sqrt{Q_0^2+A^2} - Q_0)} + \frac{\sqrt{Q_0^2+B^2} - \sqrt{Q_0^2+A^2}}{Q_0} \right. \\ &\quad \left. + \frac{(L-A)\int_a^b h^{-1}}{Q_0 \int_0^a h^{-1}} \right) + \ln R_1 - \ln\left(\sqrt{Q_0^2+B^2} - Q_0\right) - \frac{\sqrt{Q_0^2+B^2} - B}{Q_0}.\end{aligned}$$

Note that all the variables in (44) can be expressed in terms of A. Substituting into the last equation in (44) we will get an equation $F(A) = 0$ in the variable A only. The expression of $F(A)$ is complicated but can be explicitly given.

We now suppose further that $a = 1/3$, $b = 2/3$, and $h = 1$. Then,

$$B = L + R - A, \quad J_1 + J_2 = 6(L - A), \tag{47}$$

$$c_1^a = \left(\sqrt{Q_0^2+A^2} - Q_0\right) \exp\left\{ \frac{\sqrt{Q_0^2+A^2} - A}{Q_0} \right\},$$

$$c_1^b = \left(\sqrt{Q_0^2+B^2} - Q_0\right) \exp\left\{ \frac{\sqrt{Q_0^2+B^2} - B}{Q_0} \right\},$$

$$\phi^b - \phi^a = \frac{2L + R - 3A}{Q_0},$$

$$\begin{aligned}\phi^b = {} & \frac{\ln \frac{B}{R}}{\ln \frac{BL}{AR}} \left(\nu_0 + \ln \frac{L_1(\sqrt{Q_0^2+B^2} - Q_0)}{R_1(\sqrt{Q_0^2+A^2} - Q_0)} + \frac{\sqrt{Q_0^2+B^2} - \sqrt{Q_0^2+A^2} + L - A}{Q_0} \right) \\ & + \ln R_1 - \ln\left(\sqrt{Q_0^2+B^2} - Q_0\right) - \frac{\sqrt{Q_0^2+B^2} - B}{Q_0},\end{aligned}$$

$$J_2 - J_1 = 6(L-A) - \frac{6(L-A)}{\ln \frac{BL}{AR}} \left(\nu_0 + \ln \frac{L_1(\sqrt{Q_0^2+B^2} - Q_0)}{R_1(\sqrt{Q_0^2+A^2} - Q_0)} \right) - \frac{6(L-A)(\sqrt{Q_0^2+B^2} - \sqrt{Q_0^2+A^2} + L - A)}{Q_0 \ln \frac{BL}{AR}},$$

$$(J_2 - J_1) y_0 = \frac{\sqrt{Q_0^2+B^2} - \sqrt{Q_0^2+A^2} + L - A}{Q_0}.$$

The final equation involving the only unknown A is $F(A) = 0$, where

$$F(A) = e^{K(A)} \left(\sqrt{Q_0^2+A^2} - \frac{Q_0(J_2-J_1)}{6(L-A)} \right) + \frac{Q_0(J_2-J_1)}{6(L-A)} - \sqrt{Q_0^2+B^2}, \tag{48}$$

where

$$K(A) = -6(L-A) \frac{\sqrt{Q_0^2+B^2} - \sqrt{Q_0^2+A^2} + L - A}{Q_0(J_2-J_1)},$$

$B = L + R - A$, and $J_2 - J_1$ is given above.

To summarize, for the special case where

$$\alpha = \beta = 1, \quad a = 1/3, \quad b = 2/3, \quad h = 1,$$

the set of nonlinear algebraic equations is equivalent to $F(A) = 0$, where $F(A)$ is given in (48). The formula $F(A)$, although terribly complicated, involves only one unknown $A = \sqrt{c_1^a c_2^a}$. Other parameters in $F(A)$ are L_1, $L = \sqrt{L_1 L_2}$, R_1, $R = \sqrt{R_1 R_2}$, ν_0, and Q_0.

For $L = L_1 = 2$, $R = R_1 = 3$, $Q = 2Q_0 = 2$, and $\nu_0 = -20$, we find, numerically, two solutions of $F(A) = 0$: $A_1 = 0.6858357$ and $A_2 = 2$ (the latter is a removable singularity of the functions $F(A)$, J_i's, ϕ^b, and ϕ^a).

Once a feasible value for A is determined, all the unknowns will be determined. We then get a singular orbit that consists of nine pieces $\Gamma_l^0 \cup \Lambda_l \cup \Gamma_l^a \cup \Gamma_m^a \cup \Lambda_m \cup \Gamma_m^b \cup \Gamma_r^b \cup \Lambda_r \cup \Gamma_r^1$ (see Figure 3).

4. Main results and numerical simulations. Any solution of the set of algebraic equations determines a singular orbit for the connecting problem. Once a singular orbit is constructed, we apply geometric singular perturbation theory to show that, for $\epsilon > 0$ small, there is a unique solution that is close to the singular orbit. Before giving the precise statement of our result and its proof, let us explain the ideas behind it.

Let $\Gamma_l^0 \cup \Lambda_l \cup \Gamma_l^a \cup \Gamma_m^a \cup \Lambda_m \cup \Gamma_m^b \cup \Gamma_r^b \cup \Lambda_r \cup \Gamma_r^1$ be a singular orbit to the connecting problem (7) associated to B_L and B_R. For $\epsilon > 0$ small, let $M_L(\epsilon)$ be the forward trace of B_L under the flow of system (7) or, equivalently, system (8). To establish the existence of a unique solution to the boundary value problem near the singular orbit, we will show that $M_L(\epsilon)$ intersects B_R transversally in a neighborhood of the singular orbit.

Roughly speaking, the evolution of $M_L(\epsilon)$ from $x = 0$ to $x = 1$ undergoes the following nine stages with each stage guided by one of the nine pieces of the singular orbit (see Figure 3):

(l1) Along Γ_l^0: Since B_L intersects $W^s(\mathcal{Z}_l)$ transversally, $M_L(\epsilon)$ will first follow the orbit Γ_l^0 towards the vicinity of $\mathcal{Z}_l$ under the inner limit flow (12) near $x = 0$.

(l2) Along Λ_l: Once $M_L(\epsilon)$ gets close to $\mathcal{Z}_l$, the outer limit flow (19) takes over, and $M_L(\epsilon)$ will then follow the outer flow on $\mathcal{Z}_l$ or $\mathcal{S}_l$ along the orbit Λ_l towards the hypersurface $\{x = a\}$.

(l3) Along Γ_l^a: Near but before $\{x = a\}$, $M_L(\epsilon)$ will leave the vicinity of $\mathcal{Z}_l$, follow the orbit Γ_l^a under the inner limit flow (12) near $x = a$, and hit the hypersurface $\{x = a\}$.

(m1) Along Γ_m^a: Upon hitting the hypersurface $\{x = a\}$, the flow switches to the inner limit flow (26) with $Q(x) = Q$. $M_L(\epsilon)$ then follows Γ_m^a towards the vicinity of $\mathcal{Z}_m$.

(m2) Along Λ_m: Once $M_L(\epsilon)$ gets close to $\mathcal{Z}_m$, the outer limit flow (33) takes over, and $M_L(\epsilon)$ will then follow the outer flow on $\mathcal{Z}_m$ or $\mathcal{S}_m$ along the orbit Λ_m towards the hypersurface $\{x = b\}$.

(m3) Along Γ_m^b: Near but before $\{x = b\}$, $M_L(\epsilon)$ will leave the vicinity of $\mathcal{Z}_m$, follow the orbit Γ_m^b under the inner limit flow (26) near $x = b$, and hit the hypersurface $\{x = b\}$.

(r1) Along Γ_r^b: Upon hitting the hypersurface $\{x = b\}$, the flow switches to the inner limit flow (38) with $Q(x) = 0$. $M_L(\epsilon)$ then follows Γ_r^b towards the vicinity of $\mathcal{Z}_r$.

(r2) Along Λ_r: Once $M_L(\epsilon)$ gets close to $\mathcal{Z}_r$, the outer limit flow (41) takes over, and $M_L(\epsilon)$ will then follow the outer flow on $\mathcal{Z}_r$ or $\mathcal{S}_r$ along the orbit Λ_r towards the hypersurface $\{x = 1\}$.

(r3) Along Γ_r^1: Near but before $\{x = 1\}$, $M_L(\epsilon)$ will leave the vicinity of $\mathcal{Z}_r$ and follow the orbit Γ_r^1 under the inner limit flow (38) near $x = 1$. If it hits B_R, then we get our solution.

The main task is to justify the above description of the stages that $M_L(\epsilon)$ undergoes. The exchange lemma—see, for example, [47, 45, 46, 50, 51, 73]—of geometric singular perturbation theory is a result that precisely characterizes the configuration of $M_L(\epsilon)$ during its evolution through the above stages. To apply this abstract theory, one need only verify certain transversality conditions of some limiting objects.

We now state our results and provide a proof using the geometric singular perturbation theory described above.

THEOREM 4.1. *Let* $\Gamma_l^0 \cup \Lambda_l \cup \Gamma_l^a \cup \Gamma_m^a \cup \Lambda_m \cup \Gamma_m^b \cup \Gamma_r^b \cup \Lambda_r \cup \Gamma_r^1$ *be a singular orbit to the connecting problem* (7) *associated to* B_L *and* B_R. *Then, for* $\epsilon > 0$ *small, the boundary value problem* (5) *and* (6) *has a unique continuous and piecewise smooth solution near the singular orbit.*

Proof. For $\epsilon > 0$ small, choose $\delta > 0$ small. Let

$$B_L(\delta) = \{(\nu_0, u, L_1, L_2, J_1, J_2, 0) : |u - u_0| < \delta, |J_i - J_i^0| < \delta\},$$

and let $M_L(\epsilon)$ be the forward trace of $B_L(\delta)$ under the flow of system (7) or, equivalently, system (8). To prove the theorem, we need to show that $M_L(\epsilon)$ intersects B_R transversally in a neighborhood of the singular orbit. Indeed, if we let $M_R(\epsilon)$ be the backward trace of B_R near the singular orbit, then $M_L(\epsilon)$ and $M_R(\epsilon)$ intersect transversally too. The transversality implies that $\dim(M_L(\epsilon) \cap M_R(\epsilon)) =$

$\dim M_L(\epsilon) + \dim M_R(\epsilon) - 7 = 1$. Therefore, the intersection $M_L(\epsilon) \cap M_R(\epsilon)$ consists of precisely one solution to the boundary value problem, and the solution is near the singular orbit.

To establish the transversal intersection of $M_L(\epsilon)$ with B_R near the singular orbit, we apply the exchange lemma successively along the stages described above. The first application of the exchange lemma verifies the descriptions for stages (l1), (l2), and (l3); the second one for stages (m1), (m2), and (m3); and the last application verifies the descriptions for stages (r1), (r2), and (r3).

Note that $\dim B_L(\delta) = 3$. Since the fast flow is not tangent to $B_L(\delta)$, one has $\dim M_L(\epsilon) = 4$. The transversality of the intersection $B_L \cap W^s(\mathcal{Z}_l)$ along Γ_l^0 implies the transversality of the intersection $M_L(0) \cap W^s(\mathcal{Z}_l)$. The exchange lemma implies that $M_L(\epsilon)$ will first follow Γ_l^0 towards $N_L \subset \mathcal{Z}_l$, then follow $N_L \cdot x$ in the vicinity of Λ_l towards $x = a$, and leave the vicinity of $\mathcal{Z}_l$. And upon exit, $M_L(\epsilon)$ is C^1 $O(\epsilon)$-close to $W^u(N_L \times (a - \delta, a))$ in the vicinity of Γ_l^a.

Denote the intersection of $W^u(N_L \times (a - \delta, a))$ with $\{x = a\}$ by $I(a)$. Then $I(a)$ intersects $W^s(\mathcal{Z}_m)$ transversally for the flow (26). Let $K(a)$ be the forward trace of $I(a)$ under (25). The exchange lemma implies that $M_L(\epsilon)$ will first follow $K(a)$ in the vicinity of Γ_m^a towards $N_m^a \subset \mathcal{Z}_m$, then follow $N_m^a \cdot x$ in the vicinity of Λ_m towards $x = b$, and leave the vicinity of $\mathcal{Z}_m$. And upon exit, $M_L(\epsilon)$ is C^1 $O(\epsilon)$-close to $W^u(N_m^a \times (b - \delta, b))$ in the vicinity of Γ_m^b.

Denote the intersection of $W^u(N_m^a \times (b - \delta, b))$ with $\{x = b\}$ by $I(b)$. Then $I(b)$ intersects $W^s(\mathcal{Z}_r)$ transversally for the flow (38). Let $K(b)$ be the forward trace of $I(b)$ under the full system. Then exchange lemma implies that $M_L(\epsilon)$ will first follow $K(b)$ in the vicinity of Γ_r^b towards $N_r^b \subset \mathcal{Z}_r$, then follow $N_r^b \cdot x$ in the vicinity of Λ_r towards $x = 1$, and leave the vicinity of $\mathcal{Z}_r$. And upon exit, $M_L(\epsilon)$ is C^1 $O(\epsilon)$-close to $W^u(N_R \times (1 - \delta, 1))$ in the vicinity of Γ_r^1.

In summary, after three applications of the exchange lemma, we determine that $M_L(\epsilon)$ is C^1 $O(\epsilon)$-close to $W^u(N_R \times (1 - \delta, 1))$ in the vicinity of Γ_r^1. Since $W^u(N_R \times (1 - \delta, 1))$ intersects B_R transversally along Γ_r^1, we have shown that $M_L(\epsilon)$ intersects B_R transversally. The proof is complete. □

Numerical simulations are performed for $A_1 = 0.6858357$ and $A_2 = 2$ (see Figures 4 and 5). The following properties of the two solutions are predicted from the analytical results and can be observed from the numerical simulations:

(i) For both A_1 and A_2, approximately $c_2(x) - c_1(x) = Q(x)$ for $x \in (0, 1)$ except around $x = 1/3$ and $x = 2/3$—the jumping points of Q.

(ii) For $A_2 = L$, $J_1 + J_2 = 0$ from (47). As a consequence of (19) and (41), $c_1(x) = c_2(x) = L = 2$ for $x \in (0, 1/3)$ and $c_1(x) = c_2(x) = R = 3$ for $x \in (2/3, 1)$. The decreasing behavior of $c_1(x) = c_2(x)$ for $x \in (0, 1/3) \cup (2/3, 1)$ can be also predicted from that of the singular orbit corresponding to A_1.

(iii) There is a significant difference between the two solutions for $A_1 \neq L$ and $A_2 = L$: the solution for A_1 has two internal layers with limit orbits Γ_l^a and Γ_m^a at $x = a = 1/3$ that match at a point on B_a (see Figure 3); the solution for A_2 has only one internal layer $\Gamma_l^a = \Gamma_m^a$ at $x = 1/3$. This analytical consequence is not clearly shown in the figures but is indicated by the different behaviors of the ϕ-component: for A_1, with the extra transition through B_a, the layers near $x = 1/3$ are smoother than the one layer for A_2. The same remarks are true for the two solutions near $x = b = 2/3$.

5. Remarks. The defining equation $F(A) = 0$ in (48) that determines multiplicity of steady-states of the PNP system should be investigated thoroughly.

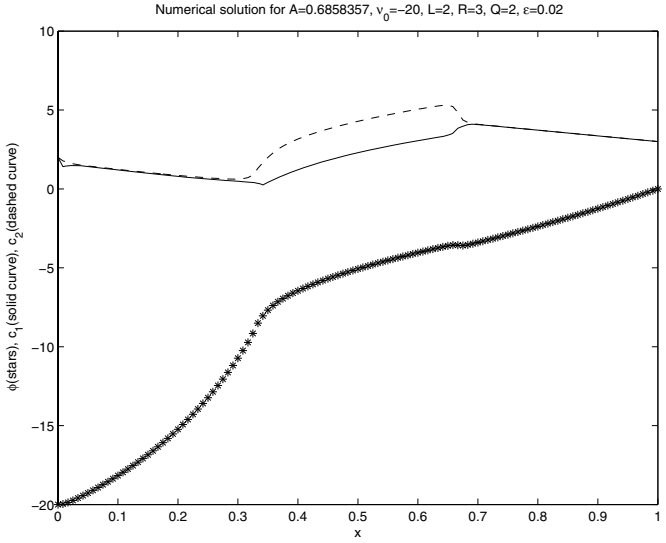


FIG. 4. *ϕ (stars), c_1 (solid curve), and c_2 (dashed curve) for $A_1 = 0.6858357$ with $L_1 = L_2 = 2$, $R_1 = R_2 = 3$, $Q = 2Q_0 = 2$, $\nu_0 = -20$, and $\epsilon = 0.02$.*

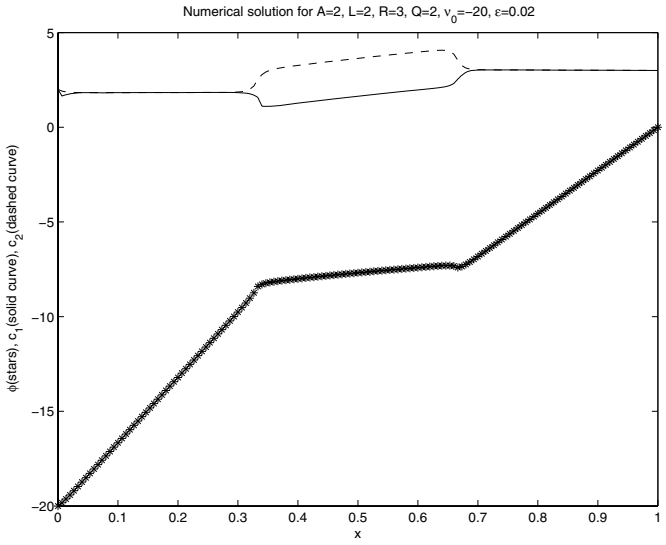


FIG. 5. *ϕ (stars), c_1 (solid curve), and c_2 (dashed curve) for $A_2 = 2$ with $L_1 = L_2 = 2$, $R_1 = R_2 = 3$, $Q = 2Q_0 = 2$, $\nu_0 = -20$, and $\epsilon = 0.02$.*

This could be studied using bifurcation theory of dynamical systems and numerical tools (e.g., AUTO) due to the presence of multiple parameters (L_i, R_i, ν_0, Q, etc. should be viewed as perturbation parameters). Another important problem is the stability of each solution in the full time evolution PNP system. Both multiplicity and stability have important biological consequences for ion channels. Single channels are in fact often defined in the laboratory by their characteristic current signal which switches from one nearly zero level ("the closed channel") to another nonzero level ("the open channel") in a random telegraph signal, with brief incomplete

spiky interruptions. Different types of channels perform their functions by controlling the open probability and/or mean duration of the stochastic signal. These gating phenomena are central to the biological function of channels and are almost always explained by saying the channel changes shape ("conformation") when it switches current level. Another explanation could be that the steady-state solutions of the PNP equations themselves have multiple solutions, and the different current levels correspond to those different solutions. Because the actual current data is stochastic, it is not clear whether the "open channel" state is stationary or not. Indeed, the open probability and/or duration of the open state might be stochastic representations of the instability of the PNP equations. Ion channels also act (in many cases) as if they have two spatially distinct gates, one of which is normally open and the other normally closed. The opening and closing processes of these gates do not overlap in ion channels so there is always a time when both gates are open and current flows through the channel. The stability properties of the equations may determine many of these gating properties. It is hard to see how the stability properties of the equations (and underlying physics) could not be involved to some significant extent, even if that gating is modulated by other processes and involves additional physics or conformational changes. Finally, there is a vitally important class of "channel" proteins in which the two gates open and close in ping pong fashion, so current can never flow right through the channel pore. These channels form mediated transporters of the greatest biological importance. It is hard to imagine that the stability of multiple solutions of the PNP equations (and the underlying physics) is not involved in the correlated gating properties of transporters, even if that gating is modulated by other processes and involves additional physics or even conformational changes.

Clearly our methods will be challenged when we try to extend them to other geometries of channels, multiple regions with nonzero permanent charges, and the even more important problems of three or more ions of different charge (e.g., Na^+, Ca^{2+}, Cl^-). The depletion layers that then occur allow the wide diversity of devices (from amplifier, to limiter, to multiplier, etc.) that can be built from a single PNP transistor, and that can be described by numerical solutions of the PNP equations [67, 35, 41, 28, 71, 34]. An alarming diversity of treatments must arise from any perturbation analysis of PNP systems because such a diversity of real devices actually exist and are built on that (physical and intellectual) substrate! Existing mathematical analysis of the PNP equations will need to be extended to show how those different devices can be built on one substrate. That is to say, analysis is needed to show how different devices arise from different values of the boundary potential but just one set of differential equations (and boundary equations), with one set of parameters (other than boundary potentials). Many useful applications in the design of channels and semiconductors depend on this analysis.

Acknowledgments. The authors are grateful to the anonymous referees whose suggestions have in our opinion, significantly improved the paper. We thank Feng He for the numerical simulations.

REFERENCES

[1] S. Aboud, D. Marreiro, M. Saraniti, and R. S. Eisenberg, *A Poisson P3M force field scheme for particle-based simulations of ionic liquids*, J. Comput. Electron., 3 (2004), pp. 117–133.

[2] S. Aboud, M. Saraniti, and R. Eisenberg, *Computational issues in modeling ion transport in biological channels: Self-consistent particle-based simulations*, J. Comput. Electron., 2 (2003), pp. 239–243.

[3] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 3rd ed., Garland, New York, 1994.
[4] R. Allen, J.-P. Hansen, and S. Melchionna, *Electrostatic potential inside ionic solutions confined by dielectrics: A variational approach*, Phys. Chem. Chem. Physics, 3 (2001), pp. 4177–4186.
[5] V. Barcilon, *Ion flow through narrow membrane channels: Part* I, SIAM J. Appl. Math., 52 (1992), pp. 1391–1404.
[6] V. Barcilon, D.-P. Chen, and R. S. Eisenberg, *Ion flow through narrow membrane channels: Part* II, SIAM J. Appl. Math., 52 (1992), pp. 1405–1425.
[7] V. Barcilon, D.-P. Chen, R. S. Eisenberg, and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Perturbation and simulation study*, SIAM J. Appl. Math. 57 (1997), pp. 631–648.
[8] J. Barthel, H. Krienke, and W. Kunz, *Physical Chemistry of Electrolyte Solutions: Modern Aspects*, Springer-Verlag, New York, 1998.
[9] M. Bazant, K. Thornton, and A. Ajdari, *Diffuse-charge dynamics in electrochemical systems*, Phys. Review E, 70 (2004), pp. 1–24.
[10] M. Z. Bazant, K. T. Chu, and B. J. Bayly, *Current-voltage relations for electrochemical thin films*, SIAM J. Appl. Math., 65 (2005), pp. 1463–1484.
[11] S. R. Berry, S. A. Rice, and J. Ross, *Physical Chemistry*, 2nd ed., Oxford University Press, New York, 2000.
[12] D. Boda, D. Busath, B. Eisenberg, D. Henderson, and W. Nonner, *Monte Carlo simulations of ion selectivity in a biological Na+ channel: Charge-space competition*, Phys. Chem. Chem. Phys., 4 (2002), pp. 5154–5160.
[13] D. Boda, D. Gillespie, W. Nonner, D. Henderson, and B. Eisenberg, *Computing induced charges in inhomogeneous dielectric media: Application in a Monte Carlo simulation of complex ionic systems*, Phys. Rev. E, 69 (2004), 046702.
[14] N. Brillantiv and T. Poschel, *Kinetic Theory of Granular Gases*, Oxford University Press, New York, 2004.
[15] J.-N. Chazalviel, *Coulomb Screening by Mobile Charges*, Birkhaüser, New York, 1999.
[16] D.-P. Chen and R. S. Eisenberg, *Charges, currents and potentials in ionic channels of one conformation*, Biophys. J., 64 (1993), pp. 1405–1421.
[17] S. Chung and S. Kuyucak, *Predicting channel function from channel structure using Brownian dynamics simulations*, Clin. Exp. Pharmacol Physiol., 28 (2001), pp. 89–94.
[18] K. E. Cooper, P. Y. Gates, and R. S. Eisenberg, *Surmounting barriers in ionic channels*, Q. Rev. Biophys., 21 (1988), pp. 331–364.
[19] S. Durand-Vidal, P. Turq, O. Bernard, C. Treiner, and L. Blum, *New perspectives in transport phenomena in electrolytes*, Phys. A, 231 (1996), pp. 123–143.
[20] B. Eisenberg, *Ion channels as devices*, J. Comput. Electron., 2 (2003), pp. 245–249.
[21] B. Eisenberg, *Proteins, channels, and crowded ions*, Biophys. Chem., 100 (2003), pp. 507–517.
[22] B. Eisenberg, *Living Transistors: A Physicist's View of Ion Channels*, http://arxiv.org/abs q-bio.BM/0506016 (June 14, 2005).
[23] R. S. Eisenberg, *Channels as enzymes*, J. Memb. Biol., 115 (1990), pp. 1–12.
[24] R. S. Eisenberg, *Atomic biology, electrostatics, and ionic channels*, in New Developments and Theoretical Studies of Proteins, R. Elber, ed., World Scientific, Philadelphia, 1996, pp. 269–357.
[25] R. S. Eisenberg, *From structure to function in open ionic channels*, J. Memb. Biol., 171 (1999), pp. 1–24.
[26] W. R. Fawcett, *Liquids, Solutions, and Interfaces: From Classical Macroscopic Descriptions to Modern Microscopic Details*, Oxford University Press, New York, 2004.
[27] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[28] D. K. Ferry, *Semiconductor Transport,* Taylor and Francis, New York, 2000.
[29] D. Gillespie and R. S. Eisenberg, *Modified Donnan potentials for ion transport through biological ion channels*, Phys. Rev. E, 63 (2001), 061902.
[30] D. Gillespie and R. S. Eisenberg, *Physical descriptions of experimental selectivity measurements in ion channels*, European Biophys. J., 31 (2002), pp. 454–466.
[31] D. Gillespie, W. Nonner, and R. S. Eisenberg, *Coupling Poisson–Nernst–Planck and density functional theory to calculate ion flux*, J. Phys. Condens. Matter, 14 (2002), pp. 12129–12145.
[32] D. Gillespie, W. Nonner, and R. S. Eisenberg, *Density functional theory of charged, hard-sphere fluids*, Phys. Rev. E, 68 (2003), 0313503.

[33] L. J. Henderson, *The Fitness of the Environment: An Inquiry into the Biological Significance of the Properties of Matter*, Macmillan, New York, 1927.
[34] H. K. Henisch, *Semiconductor Contacts: An Approach to Ideas and Models*, Oxford University Press, New York, 1989.
[35] K. Hess, J. P. Leburton, and U. Ravaioli, eds., *Computational Electronics: Semiconductor Transport and Device Simulation*, Kluwer, Boston, 1991.
[36] U. Hollerbach, D.-P. Chen, and R. S. Eisenberg, *Two- and three-dimensional Poisson–Nernst–Planck simulations of current flow through gramicidin A*, J. Comput. Sci., 16 (2001), pp. 373–409.
[37] U. Hollerbach and R. S. Eisenberg, *Concentration-dependent shielding of electrostatic potentials inside the gramicidin A channels*, Langmuir, 18 (2002), pp. 3626–3631.
[38] U. Hollerbach, D.-P. Chen, W. Nonner, and B. Eisenberg, *Three-dimensional Poisson-Nernst-Planck theory of open channels*, Biophys. J., 76 (1999), A205.
[39] M. H. Holmes, *Nonlinear ionic diffusion through charged polymeric gels*, SIAM J. Appl. Math., 50 (1990), pp. 839–852.
[40] W. Im and B. Roux, *Ion permeation and selectivity of OmpF porin: A theoretical study based on molecular dynamics, Brownian dynamics, and continuum electrodiffusion theory*, J. Mol. Biol., 322 (2002), pp. 851–869.
[41] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*, Springer-Verlag, New York, 1989.
[42] J. W. Jerome, *Consistency of semiconductor modeling: An existence/stability analysis for the stationary Van Roosbroeck system*, SIAM J. Appl. Math., 45 (1985), pp. 565–590.
[43] J. W. Jerome, *Mathematical Theory and Approximation of Semiconductor Models*, Springer-Verlag, New York, 1995.
[44] J. W. Jerome and T. Kerkhoven, *A finite element approximation theory for the drift diffusion semiconductor model*, SIAM J. Numer. Anal., 28 (1991), pp. 403–422.
[45] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme, 1994), Lect. Notes in Math. 1609, Springer-Verlag, Berlin, 1995, pp. 44–118.
[46] C. K. R. T. Jones, T. J. Kaper, and N. Kopell, *Tracking invariant manifolds up to exponentially small errors*, SIAM J. Math. Anal., 27 (1996), pp. 558–577.
[47] C. Jones and N. Kopell, *Tracking invariant manifolds with differential forms in singularly perturbed systems*, J. Differential Equations, 108 (1994), pp. 64–88.
[48] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics.* II. *Nonequilibrium Statistical Mechanics*, 2nd ed., Springer-Verlag, New York, 1995.
[49] M. G. Kurnikova, R. D. Coalson, P. Graf, and A. Nitzan, *A lattice relaxation algorithm for 3D Poisson–Nernst–Planck theory with application to ion transport through the gramicidin A channel*, Biophys. J., 76 (1999), pp. 642–656.
[50] W. Liu, *Exchange lemmas for singular perturbations with certain turning points*, J. Differential Equations, 167 (2000), pp. 134–180.
[51] W. Liu, *Geometric singular perturbation approach to steady-state Poisson–Nernst–Planck systems*, SIAM J. Appl. Math., 65 (2005), pp. 754–766.
[52] W. Liu and B. Wang, *Poisson–Nernst–Planck systems for narrow tubular-like membrane channels*, submitted.
[53] M. Lundstrom, *Fundamentals of Carrier Transport*, 2nd ed., Addison-Wesley, New York, 2000.
[54] E. Mason and E. McDaniel, *Transport Properties of Ions in Gases*, John Wiley & Sons, New York, 1988.
[55] B. Nadler, U. Hollerbach, and R. S. Eisenberg, *Dielectric boundary force and its crucial role in gramicidin*, Phys. Rev. E, 68 (2003), 021905.
[56] B. Nadler, Z. Schuss, A. Singer, and B. Eisenberg, *Diffusion through protein channels: From molecular description to continuum equations*, Nanotech., 3 (2003), 439–442.
[57] W. Nonner, L. Catacuzzeno, and B. Eisenberg, *Binding and selectivity in L-type Ca channels: A mean spherical approximation*, Biophys. J., 79 (2000), pp. 1976–1992.
[58] W. Nonner, D.-P. Chen, and B. Eisenberg, *Anomalous mole fraction effect, electrostatics, and binding in ionic channels*, Biophys. J., 74 (1998), pp. 2327–2334.
[59] W. Nonner, D. Gillespie, D. Henderson, and B. Eisenberg, *Ion accumulation in a biological calcium channel: Effects of solvent and confining pressure*, J. Phys. Chem. B, 105 (2001), pp. 6427–6436.
[60] W. Nonner, A. Peyser, D. Gillespie, and B. Eisenberg, *Relating microscopic charge movement to macroscopic currents: The Ramo–Shockley theorem applied to ion channels*, Biophys. J., 87 (2004), pp. 3716–3722.
[61] J.-H. Park and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Mathematical study*, SIAM J. Appl. Math., 57 (1997), pp. 609–630.

[62] D. J. Rouston, *Bipolar Semiconductor Devices*, McGraw-Hill, New York, 1990.
[63] I. Rubinstein, *Multiple steady states in one-dimensional electrodiffusion with local electroneutrality*, SIAM J. Appl. Math., 47 (1987), pp. 1076–1093.
[64] I. Rubinstein, *Electro-Diffusion of Ions,* SIAM Stud. Appl. Math. 11, SIAM, Philadelphia, 1990.
[65] M. Saraniti, S. Aboud, and R. S. Eisenberg, *The simulation of ionic charge transport in biological ion channels: An introduction to numerical methods*, in Reviews in Computational Chemistry, Vol. 22, John Wiley & Sons, New York, 2006, pp. 229–294.
[66] Z. Schuss, B. Nadler, and R. S. Eisenberg, *Derivation of Poisson and Nernst–Planck equations in a bath and channel from a molecular model*, Phys. Rev. E, 64 (2001), 036116.
[67] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, New York, 1984.
[68] Z. S. Siwy, M. R. Powell, E. Kalman, R. D. Asumian, and R. S. Eisenberg, *Negative incremental resistance induced by calcium in asymmetric nanopores*, Nano Lett., 6 (2006), pp. 473–477.
[69] Z. S. Siwy, M. R. Powell, A. Petrov, E. Kalman, C. Trautmann, and R. S. Eisenberg, *Calcium-induced voltage gating in single conical nanopores*, Nano Lett., 6 (2006), pp. 1729–1734.
[70] T. A. van der Straaten, G. Kathawala, R. S. Eisenberg, and U. Ravaioli, *BioMOCA—A Boltzmann transport Monte Carlo model for ion channel simulation*, Molecular Simulation, 31 (2005), pp. 151–171.
[71] B. G. Streetman, *Solid State Electronic Devices*, 4th ed., Prentice–Hall, Englewood Cliffs, NJ, 1972.
[72] C. Tanford and J. Reynolds, *Nature's Robots: A History of Proteins*, Oxford University Press, New York, 2001.
[73] S.-K. Tin, N. Kopell, and C. K. R. T. Jones, *Invariant manifolds and singularly perturbed boundary value problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1558–1576.
[74] M. Toda, R. Kubo, and N. Saito, *Statistical Physics* I, Springer-Verlag, New York, 1983.
[75] R. M. Warner, Jr., *Microelectronics: Its unusual origin and personality*, IEEE Trans. Electron. Devices, 48 (2001), pp. 2457–2467.
[76] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, New York, 2001.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1967–1980

COMPLEX SPHERICAL WAVES FOR THE ELASTICITY SYSTEM AND PROBING OF INCLUSIONS*

GUNTHER UHLMANN† AND JENN-NAN WANG‡

Abstract. We construct complex geometrical optics solutions for the isotropic elasticity system concentrated near spheres. We then use these special solutions, called complex spherical waves, to identify inclusions embedded in an isotropic, inhomogeneous, elastic background.

Key words. complex spherical waves, elasticity system, inverse problem

AMS subject classifications. 35R30, 74B05

DOI. 10.1137/060651434

1. Introduction.

Let $\Omega \subset \mathbb{R}^3$ be an open bounded domain with smooth boundary. The domain Ω is modeled as an inhomogeneous, isotropic, elastic medium characterized by the Lamé parameters $\lambda(x)$ and $\mu(x)$. Assume that $\lambda(x) \in C^2(\overline{\Omega})$, $\mu(x) \in C^4(\overline{\Omega})$ and the following inequalities hold:

$$\mu(x) > 0 \quad \text{and} \quad \lambda(x) + 2\mu(x) > 0 \quad \forall\, x \in \overline{\Omega} \quad \text{(strong ellipticity)}. \tag{1.1}$$

We consider the static isotropic elasticity system without sources:

$$\mathcal{L}u := \nabla \cdot (\lambda(\nabla \cdot u)I + 2\mu\ \mathrm{Sym}(\nabla u)) = 0 \quad \text{in} \quad \Omega, \tag{1.2}$$

where $\mathrm{Sym}(A) = (A + A^T)/2$ denotes the symmetric part of the matrix $A \in \mathbb{C}^{3\times 3}$. Equivalently, if we denote $\sigma(u) = \lambda(\nabla \cdot u)I + 2\mu\ \mathrm{Sym}(\nabla u)$ to be the stress tensor, then (1.2) becomes

$$\mathcal{L}u = \nabla \cdot \sigma = 0 \quad \text{and} \quad \Omega.$$

On the other hand, since the Lamé parameters are differentiable, we can also write (1.2) in the nondivergence form

$$\mathcal{L}u = \mu\Delta u + (\lambda + \mu)\nabla(\nabla \cdot u) + \nabla\lambda\nabla \cdot u + 2\mathrm{Sym}(\nabla u)\nabla\mu = 0 \quad \text{in} \quad \Omega. \tag{1.3}$$

Special types of solutions for elliptic equations or systems have played an important role in inverse problems since the pioneering work of Calderón [2]. In 1987, Sylvester and Uhlmann [21] introduced complex geometrical optics solutions to solve the inverse boundary value problem for the conductivity equation. For system (1.2), complex geometrical optics solutions were constructed in [5], using [4], and in [16], [17], and [18]. In [16], [17], and [18] the authors introduced an intertwining technique using pseudodifferential operators. In both [5] and [16], [17], [18], the phase functions of the complex geometrical optics solutions are linear. Other types of special solutions,

*Received by the editors February 2, 2006; accepted for publication (in revised form) October 2, 2006; published electronically March 30, 2007.
http://www.siam.org/journals/sima/38-6/65143.html
†Department of Mathematics, University of Washington, Seattle, WA 98195-4350 (gunther@math.washington.edu). The work of this author was partially supported by the NSF.
‡Department of Mathematics, Taida Institute for Mathematical Sciences, and NCTS (Taipei), National Taiwan University, Taipei 106, Taiwan (jnwang@math.ntu.edu.tw). The work of this author was supported in part by the National Science Council of Taiwan.

called oscillating-decaying solutions, were constructed for general elliptic systems in [19] and [20]. These oscillating-decaying solutions have been used in solving inverse problems, particularly in detecting inclusions and cavities [19].

In developing the theory for inverse boundary value problems with partial or local measurements, the authors of [7] and [14] gave, respectively, approximate complex geometrical optics solutions concentrated near hyperplanes and near hemispheres for the Schrödinger equation. In [14], the construction was based on hyperbolic geometry and was applied in [8] to construct complex geometrical optics solutions for the Schrödinger equation where the real part of the phase function is a radial function, i.e., its level surfaces are spheres. They call these solutions *complex spherical waves*. The hyperbolic geometry approach does not work for the Laplacian with first order perturbations such as the Schrödinger equation with magnetic potential and the isotropic elasticity equation (1.2) (see below). Recently, complex geometrical optics solutions with more general phase functions were constructed in [15] for the Schrödinger equation and in [3] for the Schrödinger equation with magnetic potential. The method used in [15] and [3] relies on Carleman-type estimates, which is a more flexible tool in treating lower order perturbations. Hence, we shall apply the method in [15] and [3] to construct complex geometrical optics solutions for (1.2) with the real part of the phase function being a radial function, i.e., complex spherical waves.

With these complex spherical waves at hand, we can study the inverse problem of detecting unknown inclusions inside an elastic body with known isotropic background medium. The investigation of this inverse problem is motivated by [8], in which the same problem was treated for the conductivity equation. There are several results, both theoretical and numerical, concerning the object identification problem by boundary measurements for the conductivity equation. We will not try to give a full account of these developments here. For detailed references, we refer to [8]. For the elasticity system, we will compare our result to some existing ones. In [10], Ikehata generalized his *probe method* to the isotropic elasticity system. Ikehata's probe method is based on singular solutions and Runge's approximation property (which is closely related to the unique continuation property). These ideas are due to Isakov [13]. On the other hand, for the general (anisotropic) elasticity system, a reconstruction method using oscillating-decaying solutions was given by the authors in [19]. The method in [19] shares the same spirit as Ikehata's *enclosure method* (see Ikehata's survey article [9]). Both methods enable us to reconstruct the support function of the inclusion by the Dirichlet-to-Neumann map. It should be noted that Runge's property was used in [19]. Ikehata's results on the enclosure method did not rely on Runge's property because he used the Laplacian as the background and explicit complex geometrical optics solutions are available for this case. Our approach here lies between the method in [19] and Ikehata's enclosure method in the sense that we treat the isotropic elasticity without using Runge's property. Furthermore, since we probe the region by complex spherical waves, it is possible to recover some concave parts of inclusions. Also, as in [8], we can localize the measurements with these complex spherical waves.

This paper is organized as follows. In section 2, (1.2) or (1.3) is transformed to a system of dimension four, and a Carleman estimate is derived for the new system. The construction of complex spherical waves for (1.2) is given in section 3. The study of the inverse problem is carried out in section 4.

2. Carleman estimate and its consequence. It suffices to work with system (1.3) here. Since the leading order of (1.3) is strongly coupled, we want to find

a reduced system whose leading part is decoupled (precisely, the Laplacian), and solutions of (1.3) can be constructed more easily. We will use the reduced system derived by Ikehata [11]. This reduction had already been mentioned in [22]. Let $W = \binom{w}{g}$ satisfy

$$PW := \Delta \begin{pmatrix} w \\ g \end{pmatrix} + \tilde{A}_1(x) \begin{pmatrix} \nabla g \\ \nabla \cdot w \end{pmatrix} + \tilde{A}_0(x) \begin{pmatrix} w \\ g \end{pmatrix} = 0, \tag{2.1}$$

where

$$\tilde{A}_1(x) = \begin{pmatrix} 2\mu^{-1/2}(-\nabla^2 + \Delta)\mu^{-1} & -\nabla \log \mu \\ 0 & \frac{\lambda+\mu}{\lambda+2\mu}\mu^{1/2} \end{pmatrix}$$

and

$$\tilde{A}_0(x) = \begin{pmatrix} -\mu^{-1/2}(2\nabla^2 + \Delta)\mu^{1/2} & 2\mu^{-5/2}(\nabla^2 - \Delta)\mu \ \nabla\mu \\ -\frac{\lambda-\mu}{\lambda+2\mu}(\nabla\mu^{1/2})^T & -\mu\Delta\mu^{-1} \end{pmatrix}.$$

Here $\nabla^2 f$ is the Hessian of the scalar function f. Then

$$u := \mu^{-1/2}w + \mu^{-1}\nabla g - g\nabla\mu^{-1}$$

satisfies (1.3). A similar form was also used in [5] for studying the inverse boundary value problem for the isotropic elasticity system.

With (2.1) at hand, we now consider the matrix operator $P_h = -h^2 P$. More precisely, we have

$$P_h = (hD)^2 + ihA_1(x) \begin{pmatrix} hD \\ hD\cdot \end{pmatrix} + h^2 A_0,$$

where $D = -i\nabla$, $A_1 = -\tilde{A}_1$, and $A_0 = -\tilde{A}_0$. Later on we shall denote the matrix operator

$$iA_1(x) \begin{pmatrix} hD \\ hD\cdot \end{pmatrix} = A_1(x, hD).$$

To construct complex geometrical optics solutions, we will follow closely the papers [3] and [15]. The construction here is simpler than the one given in [16], [17], where the technique of intertwining operators was first introduced. Furthermore, we do not need to work with C^∞ coefficients here. As in [3] and [15], we will use semiclassical Weyl calculus. Our goal here is to derive a Carleman estimate with semiclassical H^{-2} norm for P_h.

The conjugation of P_h with $e^{\varphi/h}$ is given by

$$e^{\varphi/h} \circ P_h \circ e^{-\varphi/h} = (hD + i\nabla\varphi)^2 + hA_1(x, hD + i\nabla\varphi) + h^2 A_0(x).$$

We first consider the leading operator $(hD + i\nabla\varphi)^2$ and denote

$$(hD + i\nabla\varphi)^2 = A + iB,$$

where $A = (hD)^2 - (\nabla\varphi)^2$ and $B = \nabla\varphi \circ hD + hD \circ \nabla\varphi$. The Weyl symbols of A and B are given as

$$a(x,\xi) = \xi^2 - (\nabla\varphi)^2 \quad \text{and} \quad b(x,\xi) = 2\nabla\varphi \cdot \xi,$$

respectively. Let Ω_0 be an open bounded domain such that $\bar{\Omega} \subset \Omega_0$. Accordingly, we extend λ and μ to Ω_0 by preserving their smoothness. We now let φ have nonvanishing gradient in Ω_0 and be a limit Carleman weight in Ω_0:

$$\{a, b\} = 0 \quad \text{when} \quad a = b = 0,$$

i.e.,

$$\langle \varphi'' | \nabla\varphi \otimes \nabla\varphi + \xi \otimes \xi \rangle = 0 \quad \text{when} \quad \xi^2 = (\nabla\varphi)^2 \text{ and } \nabla\varphi \cdot \xi = 0.$$

In order to get positivity in proving the Carleman estimate, we will modify the weight φ as in [3] and [15]. Let us denote $\varphi_\varepsilon = \varphi + h\varphi^2/(2\varepsilon)$, where $\varepsilon > 0$ will be chosen later. Also, we denote a_ε and b_ε to be the corresponding symbols as φ is replaced by φ_ε. Then one can easily check that

$$\{a_\varepsilon, b_\varepsilon\} = \frac{4h}{\varepsilon}\left(1 + \frac{h}{\varepsilon}\varphi\right)^2 (\nabla\varphi)^4 > 0 \quad \text{when} \quad a_\varepsilon = b_\varepsilon = 0.$$

Arguing as in [15], we get

$$\{a_\varepsilon, b_\varepsilon\} = \frac{4h}{\varepsilon}\left(1 + \frac{h}{\varepsilon}\varphi\right)^2 (\nabla\varphi)^4 + \alpha(x) a_\varepsilon + \beta(x,\xi) b_\varepsilon,$$

where $\beta(x,\xi)$ is linear in ξ. Therefore, at the operator level, we have

$$\begin{aligned} i[A_\varepsilon, B_\varepsilon] = {} & \frac{4h^2}{\varepsilon}\left(1 + \frac{h}{\varepsilon}\varphi\right)^2 (\nabla\varphi)^4 + \frac{h}{2}(\alpha \circ A_\varepsilon + A_\varepsilon \circ \alpha) \\ & + \frac{h}{2}(\beta^w \circ B_\varepsilon + B_\varepsilon \circ \beta^w) + h^3 c(x), \end{aligned} \tag{2.2}$$

where β^w denotes the Weyl quantization of β.

With the help (2.2), we now can estimate

$$\|(A_\varepsilon + iB_\varepsilon)V\|^2 = \|A_\varepsilon V\|^2 + \|B_\varepsilon V\|^2 + i\langle B_\varepsilon V | A_\varepsilon V\rangle - i\langle A_\varepsilon V | B_\varepsilon V\rangle$$

for $V \in C_0^\infty(\Omega)$. Here and below, we define the norm $\|\cdot\|$ and the inner $\langle \cdot \mid \cdot \rangle$ in terms of $L^2(\Omega)$. Integrating by parts, we conclude

$$\langle B_\varepsilon V | A_\varepsilon V\rangle = \langle A_\varepsilon B_\varepsilon V | V\rangle \quad \text{and} \quad \langle A_\varepsilon V | B_\varepsilon V\rangle = \langle B_\varepsilon A_\varepsilon V | V\rangle. \tag{2.3}$$

On the other hand, we observe that

$$\|h\nabla V\|^2 = \langle A_\varepsilon V | V\rangle + \|\sqrt{(\nabla\varphi)} V\|^2 \lesssim \|A_\varepsilon V\|^2 + \|V\|^2 \tag{2.4}$$

and the obvious estimate

$$\|(h\nabla)^2 V\|^2 \lesssim \|A_\varepsilon V\|^2 + \|V\|^2. \tag{2.5}$$

Using (2.2), (2.3), (2.4), and (2.5) gives

$$\begin{aligned} & \|(A_\varepsilon + iB_\varepsilon)V\|^2 \\ & \gtrsim \|A_\varepsilon V\|^2 + \|B_\varepsilon V\|^2 + \frac{h^2}{\varepsilon}\|V\|^2 - h(\|A_\varepsilon V\|\|V\| + \|B_\varepsilon V\|\|h\nabla V\|) \\ & \gtrsim \|A_\varepsilon V\|^2 + \|B_\varepsilon V\|^2 + \frac{h^2}{\varepsilon}\|V\|^2 - \frac{1}{2}\|A_\varepsilon V\|^2 - \frac{h^2}{2}\|V\|^2 - \frac{1}{2}\|B_\varepsilon V\|^2 \\ & \quad - \frac{h^2}{2}(\|A_\varepsilon V\|^2 + \|V\|^2) \\ & \gtrsim \left(1 - O\left(\frac{h^2}{\varepsilon}\right)\right)\|A_\varepsilon V\|^2 + \frac{h^2}{\varepsilon}(\|A_\varepsilon V\|^2 + \|V\|^2). \end{aligned}$$

Thus, taking h and ε ($h \ll \varepsilon$) sufficiently small, we arrive at

$$\|(A_\varepsilon + iB_\varepsilon)V\|^2 \gtrsim \frac{h^2}{\varepsilon}(\|V\|^2 + \|h\nabla V\|^2 + \|(h\nabla)^2 V\|^2),$$

namely,

$$\|(A_\varepsilon + iB_\varepsilon)V\|^2 \gtrsim \frac{h^2}{\varepsilon}\|V\|^2_{H^2_h(\Omega)}. \tag{2.6}$$

Here we define the semiclassical Sobolev norms

$$\|v\|^2_{H^m_h(\Omega)} = \sum_{|\alpha|\le m} \|(h\nabla)^\alpha v\|^2 \quad \forall\ m \in \mathbb{N}$$

and

$$\|v\|^2_{H^s_h(\mathbb{R}^3)} = \int (1 + |h\xi|^2)^s |\hat{v}(\xi)|^2 d\xi = \|\langle hD\rangle^s v\|^2 \quad \forall\ s \in \mathbb{R}.$$

Now let Ω_1 be open and $\bar{\Omega} \subset \Omega_1 \subset \Omega_0$. The estimate (2.6) also holds for $V \in C_0^\infty(\Omega_1)$. Then, as done in [3], we can obtain that

$$\frac{h^2}{\varepsilon}\|V\|^2_{H^2_h(\mathbb{R}^3)} \lesssim \|(A_\varepsilon + iB_\varepsilon)\langle hD\rangle^2 V\|^2_{H^{-2}_h(\mathbb{R}^3)}. \tag{2.7}$$

To add the first order perturbation $hA_{1,\varepsilon}V + h^2A_0V = hA_1(x, hD + i\nabla\varphi_\varepsilon)V + h^2A_0V$ into (2.7), we note that

$$\|(hA_{1,\varepsilon} + h^2A_0)\langle hD\rangle^2 V\|^2_{H^{-2}_h(\mathbb{R}^3)} \lesssim h^2\|V\|^2_{H^1_h(\mathbb{R}^3)}. \tag{2.8}$$

In view of (2.8), we get from (2.7) that

$$\|(A_\varepsilon + iB_\varepsilon + hA_{1,\varepsilon} + h^2A_0)\langle hD\rangle^2 V\|^2_{H^{-2}_h(\mathbb{R}^3)} \gtrsim h^2\|\langle hD\rangle^2 V\|^2, \tag{2.9}$$

provided $\varepsilon \ll 1$. Transforming back to the original operator, (2.9) is equivalent to

$$\|\langle hD\rangle^2 V\| \lesssim h\|e^{\phi_\varepsilon/h}Pe^{-\varphi_\varepsilon/h}\langle hD\rangle^2 V\|_{H^{-2}_h(\mathbb{R}^3)} \tag{2.10}$$

for $V \in C_0^\infty(\Omega_1)$.

Let $\chi \in C_0^\infty(\Omega_1)$ with $\chi = 1$ on Ω and $W \in C_0^\infty(\Omega)$. Substituting $V = \chi\langle hD\rangle^{-2}W$ into (2.10) and using the property that

$$\|(1-\chi)\langle hD\rangle^{-2}W\|_{H^s_h} = O(h^\infty)\|W\|$$

for any $s \in \mathbb{R}$, we get that

$$\|W\| \lesssim h\|e^{\phi_\varepsilon/h}Pe^{-\varphi_\varepsilon/h}W\|_{H^{-2}_h(\mathbb{R}^3)}. \tag{2.11}$$

Now since $e^{\varphi_\varepsilon/h} = e^{\varphi^2/\varepsilon}e^{\varphi/h}$ and $e^{\varphi^2/\varepsilon} = O(1)$, (2.11) becomes

$$\|W\| \lesssim h\|e^{\phi/h}Pe^{-\varphi/h}W\|_{H^{-2}_h(\mathbb{R}^3)}. \tag{2.12}$$

Note that (2.12) also holds when φ is replaced by $-\varphi$. Therefore, by the Hahn–Banach theorem, we have the following existence theorem.

THEOREM 2.1. *For h sufficiently small, for any $F \in L^2(\Omega)$, there exists $V \in H^2_h(\Omega)$ such that*

$$e^{\varphi/h}P_h(e^{-\varphi/h}V) = F$$

and $h\|V\|_{H^2_h(\Omega)} \lesssim \|F\|_{L^2(\Omega)}$.

3. Construction of complex spherical waves. In this section we will construct complex spherical waves for the elasticity system (1.3). We apply the method of [3] and [15] to our system here. We will work with the reduced system (2.1). Let ψ be a solution of the eikonal equation

$$a(x, \nabla\psi) = b(x, \nabla\psi) = 0 \quad \forall\ x \in \Omega,$$

i.e.,

$$\begin{cases} (\nabla\psi)^2 = (\nabla\varphi)^2 \\ \nabla\varphi \cdot \nabla\psi = 0 \end{cases} \quad \forall\ x \in \Omega. \tag{3.1}$$

Since $\{a, b\} = 0$ on $a = b = 0$, there exists a solution to (3.1). To construct complex spherical waves, we choose the limit Carleman weight

$$\varphi(x) = \log|x - x_0| \quad \text{for} \quad x_0 \notin \overline{\mathrm{ch}(\Omega)};$$

then a solution of (3.1) is

$$\psi(x) = \frac{\pi}{2} - \arctan \frac{\omega \cdot (x - x_0)}{\sqrt{(x - x_0)^2 - (\omega \cdot (x - x_0))^2}} = d_{\mathbb{S}^2}\left(\frac{x - x_0}{|x - x_0|}, \omega\right),$$

where $\mathrm{ch}(\Omega) :=$ convex hull of Ω and $\omega \in \mathbb{S}^2$ such that $\omega \neq (x - x_0)/|x - x_0|$ for all $x \in \overline{\Omega}$ [3]. We can be more explicit in the choices of φ and ψ. In fact, by suitable translation and rotation, we can take $x_0 = 0$, $\omega = (1, 0, 0)$ and set $z = x_1 + i|x'|$ with $x' = (x_2, x_3)$; then $\varphi + i\psi = \log z$ (see [3, Remark 3.1]). Having found ψ, we look for $U = e^{-(\varphi+i\psi)/h}(L + R)$ satisfying

$$(-h^2\Delta + h^2 A_1(x, D) + h^2 A_0(x))U = 0 \quad \text{in} \quad \Omega.$$

Equivalently, we need to solve

$$e^{(\varphi+i\psi)/h} P_h(e^{-(\varphi+i\psi)/h}(L + R)) = 0 \quad \text{in} \quad \Omega.$$

We can compute that

$$\begin{aligned} &e^{(\varphi+i\psi)/h} P_h e^{-(\varphi+i\psi)/h} \\ &= ((hD - \nabla\psi)^2 - (\nabla\varphi)^2) + i(\nabla\varphi \cdot (hD - \nabla\psi) + (hD - \nabla\psi) \cdot \nabla\varphi) \\ &\quad + h^2 A_1(x, D) + hA_1(x, i\nabla\varphi - \nabla\psi) + h^2 A_0 \\ &= h(-\nabla\psi \cdot D - D \cdot \nabla\psi + i\nabla\varphi \cdot D + iD \cdot \nabla\varphi + A_1(x, i\nabla\varphi - \nabla\psi)) + P_h \\ &= hQ + P_h, \end{aligned}$$

where $Q = -\nabla\psi \cdot D - D \cdot \nabla\psi + i\nabla\varphi \cdot D + iD \cdot \nabla\varphi + A_1(x, i\nabla\varphi - \nabla\psi)$. Hence we want to find L, independent of h, so that

$$QL = 0 \quad \text{in} \quad \Omega. \tag{3.2}$$

Equation (3.2) is a system of Cauchy–Riemann type. In fact, in view of the choices of φ and ψ above, (3.2) is equivalent to

$$\partial_{\bar{z}} L + \widetilde{A}(z, \theta)L = 0 \quad \text{in} \quad \Omega, \tag{3.3}$$

where $\widetilde{A}(z,\theta)$ is a C^2 matrix-valued function. Here we have used the cylindrical coordinates for $\mathbb{R}^3$, i.e., $x=(x_1,r,\theta)\in\mathbb{R}\times\mathbb{R}_+\times\mathbb{S}^1$, and $z=x_1+ir$. Using the results in [4], [6], or [18], one can find an invertible 4×4 matrix $G(x)\in C^2(\overline{\Omega})$ satisfying (3.2). For the sake of clarity, we outline the proof of the existence of G. We refer to, for example, [6, pp. 59–60] for more detailed arguments. It suffices to consider (3.3). Let $M>0$ satisfy $\overline{\Omega}\subset\{(x_1,r,\theta):|x_1|\le M,\ 0\le r\le M,\ \theta\in\mathbb{S}^1\}:=\mathcal{U}$. Without restriction, we can assume that (3.3) holds in $\mathcal{U}$ by suitably extending the matrix $\widetilde{A}$. By using cut-off functions with sufficiently small supports, one can show that G exists near $z_0=x_1^0+ir_0$, with $|x_1^0|<M, 0<r_0<M$, and depends C^2 smoothly on θ for all $\theta\in\mathbb{S}^1$. To construct a global invertible G in $\mathcal{U}$, we simply patch local solutions together with the help of Cartan's lemma.

So L can be chosen from columns of G. Then R is required to satisfy

$$e^{\varphi/h}P_h(e^{-(\varphi+i\psi)/h}R)=-e^{-i\psi/h}P_hL. \tag{3.4}$$

Note that $\|e^{-i\psi/h}P_hL\|\lesssim h^2$. Thus Theorem 2.1 implies that

$$\|e^{-i\psi/h}R\|_{H_h^2(\Omega)}\lesssim h, \tag{3.5}$$

which leads to

$$\|\partial^\alpha R\|_{L^2(\Omega)}\lesssim h^{1-|\alpha|}\quad\text{for}\quad|\alpha|\le2. \tag{3.6}$$

So if we write $L=\binom{\ell}{d}$ and $R=\binom{r}{s}$ with $\ell,r\in\mathbb{C}^3$, then

$$w=e^{-(\varphi+i\psi)/h}(\ell+r)\quad\text{and}\quad g=e^{-(\varphi+i\psi)/h}(d+s),$$

where r and s satisfy the estimate (3.6). Therefore, $u=\mu^{-1/2}w+\mu^{-1}\nabla g-g\nabla\mu^{-1}$ is the complex spherical wave for (1.3).

Remark 3.1. Even though the four-vector $\binom{\ell}{d}$ is nonzero in Ω, we cannot conclude that both ℓ and d never vanish in Ω. However, for any point $y\in\Omega$, it is easy to show that there exists a small ball $B_\delta(y)$ of y with $B_\delta(y)\subset\Omega$ such that one can find a pair of ℓ and d which does not vanish in $B_\delta(y)$. We will use this fact in studying our inverse problem in the next section.

4. Probing for inclusions. In this section we shall apply the complex spherical waves we constructed above to the problem of identifying the inclusion embedded inside an elastic body with isotropic medium. We now begin to set up the problem. Let D be an open subset of Ω with Lipschitz boundary satisfying the facts that $D\subset\subset\Omega$ and $\Omega\setminus\overline{D}$ is connected. Assume that $\lambda_0(x)\in C^2(\overline{\Omega})$ and $\mu_0(x)\in C^4(\overline{\Omega})$ satisfy the strong convexity condition, i.e.,

$$3\lambda_0(x)+2\mu_0(x)>0\quad\text{and}\quad\mu_0(x)>0\quad\forall\ x\in\overline{\Omega}. \tag{4.1}$$

It is obvious that (4.1) implies (1.1). On the other hand, we assume that $\widetilde{\lambda}(x)$, $\widetilde{\mu}(x)$ are two essentially bounded functions such that either

$$\widetilde{\mu}\ge0\quad\text{and}\quad3\widetilde{\lambda}+2\widetilde{\mu}\ge0\quad\text{a.e. in}\quad D$$

or

$$\widetilde{\mu}\le0\quad\text{and}\quad3\widetilde{\lambda}+2\widetilde{\mu}\le0\quad\text{a.e. in}\quad D.$$

For our inverse problem here, we shall also assume appropriate jump conditions across ∂D:

$$\text{For } y \in \partial D, \text{ there exists a ball } B_\epsilon(y) \text{ such that one of the following conditions holds:}$$
$$\begin{cases} (\mu+): & \widetilde{\mu} > \epsilon,\ 3\widetilde{\lambda} + 2\widetilde{\mu} \geq 0 \\ (\lambda+): & \widetilde{\mu} = 0,\ \widetilde{\lambda} > \epsilon \\ (\mu-): & \widetilde{\mu} < -\epsilon,\ 3\widetilde{\lambda} + 2\widetilde{\mu} \leq 0 \\ (\lambda-): & \widetilde{\mu} = 0,\ \widetilde{\lambda} < -\epsilon \end{cases} \quad \forall\ x \in B_\epsilon(y) \cap D. \tag{4.2}$$

To make sure that the forward problem is well-posed, we suppose that $\lambda = \lambda_0 + \chi_D\widetilde{\lambda}$ and $\mu = \mu_0 + \chi_D\widetilde{\mu}$ satisfy (4.1) a.e. in Ω, where χ_D is the characteristic function of D. Therefore, for any $f \in H^{1/2}(\partial\Omega)$, there exists a unique (weak) solution u to

$$\begin{cases} \mathcal{L}_D u = 0 & \text{in} \quad \Omega, \\ u = f & \text{on} \quad \partial\Omega. \end{cases}$$

Here the elastic operator $\mathcal{L}_D$ is defined in terms of λ and μ. The Dirichlet-to-Neumann map related to $\mathcal{L}_D$ is now defined as

$$\Lambda_D : f \to \sigma(u)\nu|_{\partial\Omega},$$

where ν is the unit outer normal of $\partial\Omega$ and for $x \in \partial\Omega$

$$\sigma(u) = \lambda(\nabla \cdot u)I + 2\mu\ \mathrm{Sym}(\nabla u) = \lambda_0(\nabla \cdot u)I + 2\mu_0\ \mathrm{Sym}(\nabla u).$$

Now assume that all parameters are known except $\widetilde{\lambda}$, $\widetilde{\mu}$, and D. The inverse problem is to determine D by Λ_D. This inverse problem was studied by Ikehata [10] with the so-called probe method. However, as we mentioned in the introduction, this method relies on Runge's approximation property, which is difficult to realize in practice. In this paper we approach this inverse problem from a different viewpoint. We would like to get partial information of D by local measurements. Our main tool is the use of complex spherical waves to probe for the inclusions. One of the advantages of our method is that we do not need Runge's property, and we can quickly determine roughly where the inclusion is located by only a few measurements that can be advantageous in practical applications.

We first derive some integral inequalities that we need. Let Λ_0 be the Dirichlet-to-Neumann map related to $\mathcal{L}_0$, where $\mathcal{L}_0$ is the elastic operator defined in terms of λ_0 and μ_0. Assume that u_0 is the solution of

$$\begin{cases} \mathcal{L}_0 u_0 = 0 & \text{in} \quad \Omega, \\ u_0 = f & \text{on} \quad \partial\Omega. \end{cases} \tag{4.3}$$

Then we have the following inequalities:

$$\begin{aligned} &\int_D \left\{ \frac{3\lambda_0 + 2\mu_0}{3(3\lambda + 2\mu)} \big(3\widetilde{\lambda} + 2\widetilde{\mu}\big)|\nabla \cdot u_0|^2 + 2\frac{\mu_0}{\mu}\widetilde{\mu} \left| \mathrm{Sym}(\nabla u_0) - \frac{\nabla \cdot u_0}{3} I \right|^2 \right\} dx \\ &\leq\ \langle (\Lambda_D - \Lambda_0) f, \overline{f} \rangle \\ &\leq \int_D \left\{ \frac{3\widetilde{\lambda} + 2\widetilde{\mu}}{3} |\nabla \cdot u_0|^2 + 2\widetilde{\mu} \left| \mathrm{Sym}(\nabla u_0) - \frac{\nabla \cdot u_0}{3} I \right|^2 \right\} dx \end{aligned} \tag{4.4}$$

(see [10, Proposition 5.1]). The plan now is to plug complex spherical waves u_0 given in Ω with parameters $h > 0$ and $t > 0$, denoted by $u_{0,h,t}$, into (4.4). For brevity, we will suppress the subscript 0 and denote $u_{0,h,t} = u_{h,t}$. We set $u_{h,t} = e^{\log t/h} v$ and $v_h = \mu_0^{-1/2} w + \mu_0^{-1} \nabla g - g \nabla \mu_0^{-1}$ with $w = e^{-(\varphi+i\psi)/h}(\ell + r)$ and $g = e^{-(\varphi+i\psi)/h}(d+s)$, where $W = \binom{w}{g}$ satisfies $PW = 0$ in Ω with λ, μ being replaced by λ_0, μ_0 (see (2.1)). Recall that r and s satisfy (3.6). Furthermore, for any $x \in \Omega$, we can choose a neighborhood of x such that $\ell(x)$ and $d(x)$ never vanish in such a neighborhood. In view of (4.4), we need to compute $\nabla \cdot u_{h,t}$ and $\mathrm{Sym}(\nabla u_{h,t})$ in detail. We note that

$$\Delta g = -\mu_0^{1/2} \frac{\lambda_0 + \mu_0}{\lambda_0 + 2\mu_0} \nabla \cdot w + b_0 \cdot w + c_0 g, \tag{4.5}$$

where (b_0, c_0) is the bottom row of A_0. From (4.5) we have

$$\begin{aligned}
&\nabla \cdot v_h \\
&= \nabla \mu_0^{-1/2} \cdot w + \mu_0^{-1/2} \nabla \cdot w + \nabla \mu_0^{-1} \cdot \nabla g + \mu_0^{-1} \Delta g - \nabla g \cdot \nabla \mu_0^{-1} - g \Delta \mu_0^{-1} \\
&= \nabla \mu_0^{-1/2} \cdot w + \mu_0^{-1/2} \nabla \cdot w + \mu_0^{-1} \Delta g - g \Delta \mu_0^{-1} \\
&= (\nabla \mu_0^{-1/2} + \mu_0^{-1} b_0) \cdot w + \mu_0^{-1/2} \left(1 - \frac{\lambda_0 + \mu_0}{\lambda_0 + 2\mu_0}\right) \nabla \cdot w + (\mu_0^{-1} c_0 - \Delta \mu_0^{-1}) g \\
&= e^{-(\varphi+i\psi)/h} \left\{ (\nabla \mu_0^{-1/2} + \mu_0^{-1} b_0) \cdot (\ell + r) - \mu_0^{-1/2} \left(1 - \frac{\lambda_0 + \mu_0}{\lambda_0 + 2\mu_0}\right) \frac{\nabla \varphi + i \nabla \psi}{h} \cdot (\ell + r) \right. \\
&\quad \left. + \mu_0^{-1/2} \left(1 - \frac{\lambda_0 + \mu_0}{\lambda_0 + 2\mu_0}\right) \nabla \cdot (\ell + r) + (\mu_0^{-1} c_0 - \Delta \mu_0^{-1})(d + s) \right\}.
\end{aligned} \tag{4.6}$$

Next we observe that

$$\mathrm{Sym}(\nabla v_h) = \mathrm{Sym}(\nabla \mu_0^{-1/2} \otimes w) + \mu_0^{-1/2} \mathrm{Sym}(\nabla w) + \mu_0^{-1} \nabla^2 g - g \nabla^2 \mu_0^{-1}$$

and hence

$$\begin{aligned}
&\mathrm{Sym}(\nabla v_h) \\
&= e^{-(\varphi+i\psi)/h} \left\{ \mathrm{Sym}(\nabla \mu_0^{-1/2} \otimes (\ell + r)) - \frac{1}{h} \mu_0^{-1/2} \mathrm{Sym}((\nabla \varphi + i \nabla \psi) \otimes (\ell + r)) \right. \\
&\quad + \mu_0^{-1/2} \mathrm{Sym}(\nabla(\ell + r)) + \mu_0^{-1} \nabla^2 (d + s) - \mu_0^{-1} \frac{1}{h} (d + s) \nabla^2 (\varphi + i\psi) \\
&\quad - \mu_0^{-1} \frac{2}{h} \mathrm{Sym}(\nabla(\varphi + i\psi) \otimes \nabla(d + s)) + \mu_0^{-1} \frac{1}{h^2} \nabla(\varphi + i\psi) \otimes \nabla(\varphi + i\psi)(d + s) \\
&\quad \left. - (d + s) \nabla^2 \mu_0^{-1} \right\},
\end{aligned} \tag{4.7}$$

where $(a \otimes b)_{jk} = (a_j b_k)$ for $1 \le j, k \le 3$.

We are now in a position to discuss the inverse problem. Recall that $\varphi = \log|x - x_0|$ with $x_0 \notin \overline{\mathrm{ch}(\Omega)}$. Let $f_{h,t}$ be the boundary value of $u_{h,t}$ on $\partial\Omega$ and denote

$$E(h, t) = |\langle (\Lambda_D - \Lambda_0) f_{h,t}, \overline{f_{h,t}} \rangle|.$$

Our main result for the inverse problem is the following.

THEOREM 4.1. *Assume that the jump condition* (4.2) *holds. For* $t > 0$ *and sufficiently small* h*, we have the following:*

(i) *If* $\operatorname{dist}(D, x_0) =: d_0 > t$*, then* $E(h,t) \le Ca^{1/h}$ *for some constants* $C > 0$ *and* $a < 1$*.*

(ii) *If* $d_0 < t$*, then* $E(h,t) \ge Cb^{1/h}$ *for some constants* $C > 0$ *and* $b > 1$ *with appropriate choices of* $f_{h,t}$*.*

(iii) *If* $\overline{D} \cap \overline{B_t(x_0)} = y$*, then*

$$\begin{cases} C'h^{-1} \le E(h,t) \le Ch^{-3} \text{ if } (\mu\pm) \text{ holds near } y, \\ C'h \le E(h,t) \le Ch^{-1} \text{ if } (\lambda\pm) \text{ holds near } y, \end{cases} \tag{4.8}$$

provided ℓ *and* d *of* $u_{h,t}$ *do not vanish near* y*.*

Proof. To prove the theorem, we simply substitute $u_{h,t}$ into (4.4). The key observation comes from (4.6) and (4.7). We consider only the cases $(\mu+)$ and $(\lambda+)$ of (4.2) here. The same arguments work for $(\mu-)$ and $(\lambda-)$ of (4.2). The only change is to use integral inequalities obtained by multiplying "$-$" on (4.4). If $(\mu+)$ holds, then the leading terms of two integrals in (4.4) come from $\operatorname{Sym}(\nabla u_{h,t})$ and are determined by

$$\frac{1}{h^4}\left(\frac{t}{|x-x_0|}\right)^{2/h}((\nabla\varphi)^2 + (\nabla\psi)^2)^2|d|^2 = \frac{4}{h^4}\left(\frac{t}{|x-x_0|}\right)^{2/h}(\nabla\varphi)^4|d|^2. \tag{4.9}$$

On the other hand, if $(\lambda+)$ holds, then the leading terms in those integrals in (4.4) come from $\nabla \cdot u_{h,t}$ and are governed by

$$\frac{2}{h^2}\left(\frac{t}{|x-x_0|}\right)^{2/h}(\nabla\varphi)^2|\ell|^2. \tag{4.10}$$

Using (4.4), (4.9), and (4.10), the proof of (i) follows easily from

$$E(h,t) \le C\frac{1}{h^4}\left(\frac{t}{d_0}\right)^{2/h} \quad \text{when } (\mu+) \text{ holds}$$

and

$$E(h,t) \le C\frac{1}{h^2}\left(\frac{t}{d_0}\right)^{2/h} \quad \text{when } (\lambda+) \text{ holds.}$$

For the proof of (ii), we pick a small ball $B_\delta \subset\subset B_t(x_0) \cap D$ such that the jump conditions $(\mu+)$ or $(\lambda+)$ hold in B_δ and $\ell(x)$, $d(x)$ of $u_{h,t}$ never vanish in $B_\delta(x)$. The latter property is guaranteed by Remark 3.1. For such choice of ℓ and d, the Dirichlet data is a priori given by

$$f_{h,t} = u_{h,t}|_{\partial\Omega} = e^{(\log t - \varphi - i\psi)/h}(\ell + d)|_{\partial\Omega}.$$

Thus, argued as above, we have that either

$$E(h,t) \ge C\frac{1}{h^4}\left(\frac{t}{d_0}\right)^{2/h} \quad \text{when } (\mu+) \text{ holds}$$

or

$$E(h,t) \geq C\frac{1}{h^2}\left(\frac{t}{d_0}\right)^{2/h} \quad \text{when } (\lambda+) \text{ holds,}$$

which implies (ii).

Now let $y \in \overline{D} \cap \overline{B_t(x_0)}$ and choose a ball $B_\epsilon(y)$ such that (4.2) holds and $\ell(x)$, $d(x)$ of $u_{h,t}$ never vanish in $B_\epsilon(y) \cap D$. Pick a small cone with vertex at y, say Γ, so that there exists an $\eta > 0$ satisfying

$$\Gamma_\eta := \Gamma \cap \{0 < |x-y| < \eta\} \subset B_\epsilon(y) \cap D.$$

We observe that if $x \in \Gamma_\eta$ with $|x-y| = \rho < \eta$, then $|x-x_0| \leq \rho + t$, i.e.,

$$\frac{1}{|x-x_0|} \geq \frac{1}{\rho+t}.$$

Thus, for the case $(\mu+)$, we get that from (4.4) and (4.9)

$$\begin{aligned} E(h,t) &\geq C\frac{1}{h^4}\int_D \widetilde{\mu}^2(\nabla\varphi)^4|d|^2\left(\frac{t}{|x-x_0|}\right)^{2/h} dx \\ &\geq C\epsilon\frac{1}{h^4}\int_0^\eta \left(\frac{t}{\rho+t}\right)^{2/h}\rho^2 d\rho \\ &\geq C\epsilon h^{-1}. \end{aligned} \tag{4.11}$$

On the other hand, we can choose a cone $\widetilde{\Gamma}$ with vertex at x_0 such that $\overline{D} \subset \widetilde{\Gamma} \cap \{|x-x_0| > t\}$. Hence, we can estimate

$$\begin{aligned} E(h,t) &\leq C\frac{1}{h^4}\int_{\widetilde{\Gamma}\cap\{t<|x-x_0|<t+\eta\}}\left(\frac{t}{|x-x_0|}\right)^{2/h} dx \\ &\quad + C\frac{1}{h^4}\int_{\widetilde{\Gamma}\cap\{t+\eta\leq|x-x_0|\}}\left(\frac{t}{|x-x_0|}\right)^{2/h} dx \\ &\leq C\frac{1}{h^4}\int_t^{t+\eta}\left(\frac{t}{r}\right)^{2/h} r^2 dr + O\left(\left(\frac{t}{t+\eta}\right)^{2/h}\right) \\ &\leq Ch^{-3}. \end{aligned} \tag{4.12}$$

Combining (4.11) and (4.12) yields the first estimate of (4.8). Using similar arguments, we can get the second estimate of (4.8) for the case $(\lambda+)$. □

Remark 4.1. (1) Using Theorem 4.1, we can clearly determine whether the probing front $\{|x-x_0| = t\}$ intersects the inclusion. In view of (iii) of the theorem, it is also possible to determine whether we have material jumps in μ or λ when the front touches the boundary of the inclusion.

(2) In the proofs of (ii) and (iii) we need to choose ℓ and d, which are nonvanishing in small subdomains of Ω. Since ℓ and d depend only on the known background medium, they can be chosen to be nonvanishing near any point in Ω at our disposal. In fact, it suffices to take ℓ and d, which are nonvanishing near the probe front $\{|x-x_0| = t\}$. Different choices of ℓ and d will give rise to different Dirichlet data $f_{h,t}$ and therefore different measurements.

(3) In real applications, we believe that the concerns in (ii) and (iii) can be ignored.

Taking advantage of the decay of the complex spherical waves in the region $\{|x - x_0| > t\}$, we can localize the measurements, which is of great practical value. Let $\phi_{\delta,t}(x) \in C_0^\infty(\mathbb{R}^3)$ satisfy

$$\phi_{\delta,t}(x) = \begin{cases} 1 & \text{on} \quad B_{t+\delta/2}(x_0), \\ 0 & \text{on} \quad \mathbb{R}^3 \setminus \overline{B_{t+\delta}(x_0)}, \end{cases}$$

where $\delta > 0$ is sufficiently small. Now we are going to use the measurements $f_{\delta,h,t} = \phi_{\delta,t} f_{h,t} = \phi_{\delta,t} u_{h,t}|_{\partial\Omega}$. Clearly, the measurements $f_{\delta,h,t}$ are localized on $B_{t+\delta}(x_0) \cap \partial\Omega$. Let us define

$$E_\delta(h,t) = |\langle (\Lambda_D - \Lambda_0) f_{\delta,h,t}, \overline{f_{\delta,h,t}} \rangle|.$$

THEOREM 4.2. *The statements of Theorem* 4.1 *are valid for* $E_\delta(h,t)$.

Proof. The main idea is to prove that the error caused by the remaining part of the measurement $(1 - \phi_{\delta,t}) f_{h,t} =: g_{\delta,h,t}$ is exponentially small. Let $w_{\delta,h,t}$ be the solution of (4.3) with boundary value $g_{\delta,h,t}$. We now want to compare $w_{\delta,h,t}$ with $(1 - \phi_{\delta,t}) u_{h,t}$. To this end, we first observe that

$$\begin{cases} \mathcal{L}_0((1 - \phi_{\delta,t}) u_{h,t} - w_{\delta,h,t}) = \mathcal{L}_0((1 - \phi_{\delta,t}) u_{h,t}), \\ (1 - \phi_{\delta,t}) u_{h,t} - w_{\delta,h,t} = 0 \quad \text{on} \quad \partial\Omega. \end{cases}$$

Since

$$\|\mathcal{L}_0((1 - \phi_{\delta,t}) u_{h,t})\|_{L^2(\Omega)} \le C\beta^{1/h}$$

for some $0 < \beta < 1$, we have that

$$\|(1 - \phi_{\delta,t}) u_{h,t} - w_{\delta,h,t}\|_{H^1(\Omega)} \le C\beta^{1/h}. \tag{4.13}$$

Using (4.4) for $\langle (\Lambda_D - \Lambda_0) g_{\delta,h,t}, \overline{g_{\delta,h,t}} \rangle$ with u_0 being replaced by $w_{\delta,h,t}$, we get from (4.13) and the decaying property of $u_{h,t}$ that

$$|\langle (\Lambda_D - \Lambda_0) g_{\delta,h,t}, \overline{g_{\delta,h,t}} \rangle| \le C\widetilde{\beta}^{1/h}$$

for some $0 < \widetilde{\beta} < 1$.

Now we first consider (i) of Theorem 4.1 for $E_\delta(h,t)$. We begin with the case $(\mu+)$ of (4.2). In view of the first inequality of (4.4), we see that

$$0 \le \langle (\Lambda_D - \Lambda_0)(\zeta f_{\delta,h,t} \pm \zeta^{-1} g_{\delta,h,t}), \overline{\zeta f_{\delta,h,t} \pm \zeta^{-1} g_{\delta,h,t}} \rangle$$

for any $\zeta > 0$, which leads to

$$\begin{aligned} &|\langle (\Lambda_D - \Lambda_0) f_{\delta,h,t}, \overline{g_{\delta,h,t}} \rangle + \langle (\Lambda_D - \Lambda_0) g_{\delta,h,t}, \overline{f_{\delta,h,t}} \rangle| \\ &\le \zeta^2 \langle (\Lambda_D - \Lambda_0) f_{\delta,h,t}, \overline{f_{\delta,h,t}} \rangle + \zeta^{-2} \langle (\Lambda_D - \Lambda_0) g_{\delta,h,t}, \overline{g_{\delta,h,t}} \rangle. \end{aligned} \tag{4.14}$$

It now follows from $f_{h,t} = f_{\delta,h,t} + g_{\delta,h,t}$ and (4.14) with $\zeta = 1/\sqrt{2}$ that

$$\begin{aligned} &\frac{1}{2} \langle (\Lambda_D - \Lambda_0) f_{\delta,h,t}, \overline{f_{\delta,h,t}} \rangle \\ &\le \langle (\Lambda_D - \Lambda_0) g_{\delta,h,t}, \overline{g_{\delta,h,t}} \rangle + \langle (\Lambda_D - \Lambda_0) f_{h,t}, \overline{f_{h,t}} \rangle \\ &\le C\widetilde{\beta}^{1/h} + \langle (\Lambda_D - \Lambda_0) f_{h,t}, \overline{f_{h,t}} \rangle. \end{aligned} \tag{4.15}$$

So from (i) of Theorem 4.1, the same statement holds for $E_\delta(h,t)$. Other cases of (4.2) are treated similarly.

Next we consider (ii) and (iii) of Theorem 4.1 for $E_{\delta,h,t}$. As before, we treat only $(\mu+)$ of (4.2). Choosing $\zeta = 1$ in (4.14), we get that

$$\begin{aligned} &\frac{1}{2}\langle(\Lambda_D - \Lambda_0)f_{h,t}, \overline{f_{h,t}}\rangle \\ &\leq \langle(\Lambda_D - \Lambda_0)g_{\delta,h,t}, \overline{g_{\delta,h,t}}\rangle + \langle(\Lambda_D - \Lambda_0)f_{\delta,h,t}, \overline{f_{\delta,h,t}}\rangle \\ &\leq C\widetilde{\beta}^{1/h} + \langle(\Lambda_D - \Lambda_0)f_{\delta,h,t}, \overline{f_{\delta,h,t}}\rangle. \end{aligned} \tag{4.16}$$

Therefore, (ii) of Theorem 4.1 and (4.16) imply that the same fact is true for $E_\delta(h,t)$. Finally, combining (4.15) and (4.16) yields statement (iii) for $E_\delta(h,t)$. The proof is now complete. □

Remark 4.2. With the help of Theorem 4.2, when parts of ∂D are near the boundary $\partial\Omega$, it is possible to detect some points of ∂D from only a few measurements taken from a very small region of $\partial\Omega$.

To end this section, we provide an algorithm of the method.

Step 1. Pick a point x_0 near $\mathrm{ch}(\Omega)$. Construct complex spherical waves $u_{h,t}$.

Step 2. Draw two balls $B_t(x_0)$ and $B_{t+\delta}(x_0)$. Set the Dirichlet data $f_{\delta,h,t} = \phi_{\delta,t}u_{h,t}|_{\partial\Omega}$. Measure the Neumann data $\Lambda_D f_{\delta,h,t}$ over the region $B_{t+\delta}(x_0) \cap \partial\Omega$.

Step 3. Calculate $E_\delta(h,t) = \langle(\Lambda_D - \Lambda_0)f_{\delta,h,t}, f_{\delta,h,t}\rangle$. If $E(h,t)$ tends to zero as $h \to 0$, then the probing front $\{|x - x_0| = t\}$ does not intersect the inclusion. Increase t and compute $E_\delta(h,t)$ again.

Step 4. If $E_{\delta,h,t}$ increases to ∞ as $h \to 0$, then the front $\{|x - x_0| = t\}$ intersects the inclusion. Decrease t to make a more accurate estimate of ∂D.

5. Conclusion. In this work we have constructed complex spherical waves or complex geometrical optics solutions for the elasticity system with isotropic inhomogeneous medium. We used these special solutions to investigate the inverse problem of identifying inclusions with localized measurements. Numerical realization of this method would be an interesting project. The same method should work for identifying cavities.

Acknowledgment. This work was done when the second author was visiting the Department of Mathematics at the University of Washington. He would like to thank this institute for its warm hospitality.

REFERENCES

[1] G. Alessandrini and A. Morassi, *Strong unique continuation for the Lamé system of elasticity*, Comm. Partial Differential Equations, 26 (2001), pp. 1787–1810.

[2] A. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and Its Applications to Continuum Physics, Soc. Brasileira de Matemática, Rio de Janeiro, Brazil, 1980, pp. 65–73.

[3] D. Dos Santos Ferreira, C. E. Kenig, J. Sjöstrand, and G. Uhlmann, *Determining the magnetic Schrödinger operator from partial Cauchy data*, Comm. Math. Phys., to appear.

[4] G. Eskin, *Global uniqueness in the inverse scattering problem for the Schrödinger operator with external Yang-Mills potentials*, Comm. Math. Phys., 222 (2001), pp. 503–531.

[5] G. Eskin and J. Ralston, *On the inverse boundary value problem for linear isotropic elasticity*, Inverse Problems, 18 (2002), pp. 907–921.

[6] G. Eskin and J. Ralston, *On the inverse boundary value problem for linear isotropic elasticity and Cauchy-Riemann system*, in Inverse Problems and Spectral Theory, Contemp. Math. 348, AMS, Providence, RI, 2004, pp. 53–69.

[7] A. Greenleaf and G. Uhlmann, *Local uniqueness for the Dirichlet-to-Neumann map via the two-plane transform*, Duke Math. J., 108 (2001), pp. 599–617.
[8] T. Ide, H. Isozaki, S. Nakata, S. Siltanen, and G. Uhlmann, *Probing for electrical inclusions with complex spherical waves*, Comm. Pure Appl. Math., to appear.
[9] M. Ikehata, *The enclosure method and its applications*, in Analytic Extension Formulas and Their Applications (Fukuoka, 1999/Kyoto, 2000), Int. Soc. Anal. Appl. Comput. 9, Kluwer Academic, Dordrecht, The Netherlands, 2001, pp. 87–103.
[10] M. Ikehata, *Reconstruction of inclusion from boundary measurements*, J. Inverse Ill-Posed Probl., 10 (2002), pp. 37–65.
[11] M. Ikehata, *A Remark on an Inverse Boundary Value Problem Arising in Elasticity*, preprint.
[12] M. Ikehata, G. Nakamura, and M. Yamamoto, *Uniqueness in inverse problems for the isotropic Lame system*, J. Math. Sci. Univ. Tokyo, 5 (1998), pp. 627–692.
[13] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.
[14] H. Isozaki and G. Uhlmann, *Hyperbolic geometry and the local Dirichlet-to-Neumann map*, Adv. Math., 188 (2004), pp. 294–314.
[15] C. E. Kenig, J. Sjöstrand, and G. Uhlmann, *The Calderón problem with partial data*, Ann. of Math. (2), to appear.
[16] G. Nakamura and G. Uhlmann, *Global uniqueness for an inverse boundary problem arising in elasticity*, Invent. Math., 118 (1994), pp. 457–474.
[17] G. Nakamura and G. Uhlmann, *Erratum: "Global uniqueness for an inverse boundary value problem arising in elasticity,"* Invent. Math., 152 (2003), pp. 205–207.
[18] G. Nakamura and G. Uhlmann, *Complex geometric optics solutions and pseudoanalytic matrices*, in Ill-Posed and Inverse Problems, VSP, Zeist, 2002, pp. 305–338.
[19] G. Nakamura, G. Uhlmann, and J.-N. Wang, *Oscillating-decaying solutions, Runge approximation property for the anisotropic elasticity system and their applications to inverse problems*, J. Math. Pures Appl., 84 (2005), pp. 21–54.
[20] G. Nakamura, G. Uhlmann, and J.-N. Wang, *Oscillating-decaying solutions for elliptic systems*, in Inverse Problems, Multi-Scale Analysis, and Effective Medium Theory, Contemp. Math. 408, AMS, Providence, RI, 2006, pp. 219–230.
[21] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.
[22] G. Uhlmann, *Developments in inverse problems since Calderón's foundational paper*, in Harmonic Analysis and Partial Differential Equations (Essays in Honor of Alberto P. Calderón), The University of Chicago Press, Chicago, 1999, pp. 295–345.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 1981–2004

A CLASS OF FREE BOUNDARY PROBLEMS WITH ONSET OF A NEW PHASE*

PATRICK GUIDOTTI†

Abstract. A class of diffusion-driven free boundary problems is considered which is characterized by the initial onset of a phase and by an explicit kinematic condition for the evolution of the free boundary. By a domain fixing change of variables it naturally leads to coupled systems comprising a singular parabolic initial boundary value problem and a Hamilton–Jacobi equation. Even though the one-dimensional case has been thoroughly investigated, results as basic as well-posedness and regularity have so far not been obtained for its higher-dimensional counterpart. In this paper a recently developed regularity theory for abstract singular parabolic Cauchy problems is utilized to obtain the first well-posedness results for the free boundary problems under consideration. The derivation of elliptic regularity results for the underlying static singular problems will play an important role.

Key words. free boundary problem, kinematic condition, singular parabolic and elliptic equations, well-posedness, existence, regularity

AMS subject classifications. 35A07, 35C15, 35M10, 35J70, 35K65

DOI. 10.1137/060660369

1. Introduction.

In this paper we consider a class of one phase free boundary problems (FBPs) characterized by the initial onset of a phase. Such FBPs arise in the description of diffusion in polymers, for instance. Under physically natural conditions, these problems lead to a formulation in which the phase is initially absent. This feature manifests itself mathematically in that some of the equations in the nonlinear system become singular (if they are written in a fixed reference domain). In spite of the fact that these types of problems have been intensively studied in the literature over an extended period of time, only one publication deals with the singular case considered here in more than one space dimension. The simplifying assumption that the phase be initially nonempty is typically added to avoid the mathematical complications stemming from the singularity. The one-dimensional case has, however, been thoroughly investigated [9, 10, 11, 8, 7, 14] in the specific context of diffusion in polymers and, more generally, for diffusion-driven FBPs (see [18, 12], for instance). The methods used for the one-dimensional case rely on the explicit use of the heat kernel to reduce the problem to the boundary. This approach cannot be used in higher dimensions because the singular behavior induced by the initial condition on the FBP cannot be decoupled from the diffusion operator. This is due to the fact that the free boundary has a nontrivial geometry in this case and to the fact that fundamental solutions (evolution operators) have not been studied for singular parabolic problems for which the singularity affects the underlying elliptic operator in an anisotropic way. The higher-dimensional problem has recently been studied in [16] but only in its simpler quasi-stationary form. [16] established well-posedness for the problem in a class of functions which needs to be carefully crafted and leads to an asymptotic expansion for the singularity which is valid in the corresponding topologies and has practical relevance (cf. [7, 14]). Regularity results, maximal regularity in particular, for singular

*Received by the editors May 19, 2006; accepted for publication (in revised form) October 10, 2006; published electronically March 30, 2007.
http://www.siam.org/journals/sima/38-6/66036.html
†Department of Mathematics, University of California, Irvine, CA 92697-3875 (gpatrick@math.uci.edu).

parabolic equations play a crucial role in this paper and have been recently obtained by the author in [13], where a construction of the evolution operator is given for a wide class of singular parabolic problems. Maximal regularity is needed because the full problem will be solved by reduction to a model problem and perturbation about it. Since the ensuing perturbation is of maximal order (both in the singularity and differentiation) optimal regularity results are necessary. Previous abstract results were obtained in [15, 21], but they do not apply to the situation considered here because their validity does not cover the case of spaces of classical pointwise regularity, nor the case in which the singularity affects the equations anisotropically. Classical pointwise regularity is needed in the analysis proposed here since the singular parabolic equation is coupled to a Hamilton–Jacobi-type equation modeling the front's dynamics. It has long been observed, as mentioned above, that one-dimensional techniques cannot be extended to the higher-dimensional case. Summarizing, the presence of singular coefficients and the coupling to a Hamilton–Jacobi equation are two of the characterizing features of the problem under consideration. They make its analysis more difficult and delicate than that of the related but different classical Stefan problem. This paper therefore offers a successful approach that fully overcomes these difficulties.

The unknowns of the problem are a function $u : \Omega_t \to \mathbb{R}$ defined on the open domain $\Omega_t \subset \mathbb{R}^{n+1} \ni (x, y)$ and its unknown boundary Γ_t. The function u measures the concentration of the penetrant in the case of diffusion in polymers. Many geometries can be chosen for the domain Ω_t. Here a strip-like setting is chosen where the domain is bounded by a fixed lower and an upper moving hypersurface denoted by Γ_0 and Γ_t, respectively. Other configurations, like annulus-type domains, are possible and interesting, and the results obtained here would apply to those since they all would lead to the same local model problems. The system of equations satisfied by (u, Γ_t) proposed in [17] and generalizing [1, 6] then reads

$$\varepsilon u_t - \triangle_{x,y} u = 0 \qquad \text{in } \bigcup_{t>0} \{t\} \times \Omega_t \,, \tag{1.1}$$

$$u = g \qquad \text{on } (0,\infty) \times \Gamma_0 \,, \tag{1.2}$$

$$-\partial_{\nu_t} u = (1 + \varepsilon u) V \qquad \text{on } \bigcup_{t>0} \{t\} \times \Gamma_t \,, \tag{1.3}$$

$$V = (1 + \delta \, \mathrm{H}_t) u \qquad \text{on } \bigcup_{t>0} \{t\} \times \Gamma_t \,, \tag{1.4}$$

$$\Gamma_t\big|_{t=0} = \Gamma_0 \qquad \text{at } t = 0 \,. \tag{1.5}$$

In the above equations V denotes the front speed in normal (outward) direction ν_t, and H_t denotes its mean curvature. The function $g > 0$ describes the concentration profile of a reservoir of the diffusing molecule, whereas the positive constants ε, δ arise in the nondimensionalization process. They measure the deviation from a threshold concentration of the penetrant and the strength of the curvature term, respectively. Condition (1.3) models conservation of mass across the free boundary, whereas condition (1.4) is a phenomenological law which is capable of capturing the behavior observed in experiments known as case II or anomalous diffusion. The curvature term naturally appears if one assumes the front speed to be proportional to the local average concentration in a reference ball of radius $\delta > 0$ rather than to its pointwise value. The system is clearly nonlinear, and even if the boundary conditions were linear, it still would be. Indeed, two different solutions have different domains of definitions and cannot be added. This typical feature of FBPs becomes apparent after the simple

change of variables

$$(\tau,\xi,\eta)=\left(t,x,\frac{y}{s(t,x)}\right),\qquad \hat{u}(\tau,\xi,\eta)=u\big(\tau,\xi,\eta s(\tau,\xi)\big),$$

which transforms the problem to a corresponding one on a fixed domain. To do so, it is assumed that the fixed boundary of the unknown domain is given by

$$\Gamma_0=\mathbb{R}^{n-1}\times\{0\}$$

and that the unknown moving boundary can be described as the graph of a function $s(t,\cdot)$. The latter assumption is motivated by (initial) condition (1.5). In the new variables, the domain becomes the strip

$$S=\mathbb{R}^{n-1}\times[0,1]$$

with the obvious boundaries denoted by Γ_j, $j=0,1$. Using the old notation for the new variables, the system now reads

(1.6)
$$\varepsilon u_t-\triangle_x u-\frac{1+y^2|\nabla s|^2}{s^2}\partial_y^2 u=\varepsilon y\frac{\dot{s}}{s}\partial_y u-2y\frac{1}{s}\big(\nabla s\big|\partial_y\nabla u\big)-y\frac{s\triangle s-2|\nabla s|^2}{s^2}\partial_y u,$$

(1.7)
$$u=g \qquad \text{on } \Gamma_0,$$

(1.8)
$$-\frac{1+|\nabla s|^2}{s}\partial_y u=(1+\varepsilon u)\dot{s}-\big(\nabla s\big|\nabla u\big) \qquad \text{on } \Gamma_1,$$

(1.9)
$$\dot{s}=\sqrt{1+|\nabla s|^2}(1+\delta\,\mathrm{H}_t)u \qquad \text{on } \Gamma_1,$$

(1.10)
$$s(0,\cdot)=0.$$

In [16] the quasi-stationary approximation ($\varepsilon=0$) in two space dimensions was considered in the presence or absence of the mean curvature term in two space dimensions. Here there is no restriction on the spatial dimension, and the full evolutionary problem is analyzed but the curvature effects are neglected ($\delta=0$).

The main result of this paper is establishing the well-posedness of system (1.6)–(1.10). To do so, appropriate function spaces have to be used which are able to capture the regularity of the solution as well as its asymptotic behavior at the origin. The choice of function spaces is limited by the simultaneous presence of a singular parabolic problem and a Hamilton–Jacobi equation. Spaces of classical regularity are better suited for the latter and thus reduce the freedom of choice for the parabolic problem. A compromise can be reached by using spaces of classical Hölder regularity in space and of singular Hölder behavior in time. A number of preparatory results are needed; they are formulated in the next section. In section 3 elliptic and parabolic results for singular equations will be derived which play a crucial role in the existence proof given in the last section.

2. Preliminaries and setting. In order to obtain existence results for system (1.6)–(1.10) a linearization procedure will be used in combination with maximal regularity results for the relevant linearized problems. The presence of the singularity complicates the analysis significantly since the relevant regularity results are not available and need to be derived first. The linearization procedure is meant to capture the leading order terms of the differential operators at the origin with respect to both their differentiation order and their singularity degree.

Of interest in this paper are classical solutions. If such a solution existed, one would be able to guess from (1.7) and (1.9)–(1.10) that

$$s(t,x) \approx t\, g(x) \quad \text{for } t \approx 0\,.$$

This indicates that the relevant model problem which captures the leading order behavior of (1.6)–(1.8) has the form

$$\varepsilon u_t - \triangle_x u - \frac{1}{t^2 c^2(x)}\partial_y^2 u = f(t,x,y) \text{ in } (0,\infty)\times S\,, \tag{2.1}$$

$$u = g(t,x) \quad \text{ on } (0,\infty)\times \Gamma_0\,, \tag{2.2}$$

$$\frac{1}{t}\partial_y u = h(t,x) \quad \text{ on } (0,\infty)\times \Gamma_1\,, \tag{2.3}$$

where in the situation at hand g is independent of the time variable and $c = g$. The more general case of c independent of g is, however, of interest and motivates the different notation. For reasons which can be guessed and will become apparent later, both parabolic ($\varepsilon = 1$) and elliptic ($\varepsilon = 0$) regularity results are needed in the analysis of (2.1)–(2.3). The main ingredients needed to derive such regularity results are vector-valued Fourier multiplier theorems and the use of spaces of singularly Hölder continuous functions in the time variable and the construction of an evolution operator for singular families of generators. The first are crucial in the analysis of the elliptic problem with time frozen, and the latter allow for the quantitative characterization of the singular behavior in the origin ($t = 0$) both in the elliptic and parabolic cases (where maximal regularity results are needed).

For the sake of completeness the formulation of the relevant Fourier multiplier theorem and the definition of the classes of singular Hölder continuous functions needed in the analysis are given here. The basic observation illuminating the reason for their combined use will also be presented in this section.

Assume that E is a given Banach space and that $T > 0$ and $\beta \in (0,1)\ [\cup\{1-\}]$. Then the standard Hölder (Lipschitz) space is given by

$$\mathrm{C}^{\beta}\big([0,T],E\big) := \left\{ f \in \mathrm{C}\big([0,T],E\big) \,\middle|\, [u]_{\beta,[0T]} := \sup_{t\neq s} \frac{|u(t)-u(s)|}{|t-s|^{\beta}} < \infty \right\}$$

with norm

$$\|\cdot\|_{\beta} = \|\cdot\|_{\infty,[0,T]} + [\cdot]_{\beta,[0,T]}\,.$$

If $\beta = 1$ is chosen, one obtains the space of Lipschitz continuous functions. To distinguish it from the space of continuously differentiable functions the notational device $\beta = 1-$ is used. This means that $\beta = 1$ for all practical purposes except in the notation for the space which becomes C^{1-}. Singular counterparts are given by

$$\mathrm{C}^{\beta}_{\beta}\big((0,T],E\big) := \big\{ f \in \mathrm{B}\big((0,T],E\big) \,\big|\, [t \mapsto t^{\beta} u(t)] \in \mathrm{C}^{\beta}\big((0,T],E\big)\big\} \tag{2.4}$$

with weighted norm defined through

$$\|\cdot\|_{\beta,\beta} := \|u\|_{\infty,(0,T]} + [(\cdot)^{\beta} u]_{\beta,(0,T]}\,.$$

The symbol denoting the time interval in the notation for the norm will be dropped in what follows with the understanding that the interval of definition does not contain

the origin for singular spaces, whereas it does for regular ones. The following closed subspace of regular Hölder functions will also be useful:

$$\mathrm{C}_0^\beta([0,T],E) := \big\{ f \in \mathrm{C}^\beta\big([0,T],E\big) \,\big|\, f(0)=0 \big\}\,.$$

The vector-valued Fourier multiplier theorem which is needed here can be found in [2, Theorem 6.2]. We reproduce it here for the ease of the reader. The symbol class

$$\mathcal{S}^m(E_0,E_1) := \Big\{ a \in \mathrm{C}^{n+1}\big(\mathbb{R}^n\setminus\{0\},\mathcal{L}(E_0,E_1)\big) \,\Big|\, \sup_{\xi\in\mathbb{R}^n} \big|[1+|\xi|^2]^{(m+|\alpha|)/2}\partial^\alpha a(\xi)\big|_{\mathcal{L}(E_0,E_1)} < \infty\,,\ |\alpha|\le n+1 \Big\}, \tag{2.5}$$

where E_0 and E_1 are Banach spaces and $m\in\mathbb{Z}$, is instrumental in the formulation of the result. Given a symbol $a\in\mathcal{S}^m(E_0,E_1)$, an operator can be associated to it by

$$a(D) := \mathcal{F}^{-1} a\,\mathcal{F}$$

through conjugation with the vector-valued Fourier transform

$$\mathcal{F} \in \mathcal{L}_{is}\Big(\mathcal{S}\big(\mathbb{R}^n,E_j\big)\Big) \cap \mathcal{L}_{is}\Big(\mathcal{S}'\big(\mathbb{R}^n,E_j\big)\Big)\,,\quad j=0,1\,,$$

where $\mathcal{S}$ and $\mathcal{S}'$ are the operator-valued Schwartz spaces of fast decaying test functions and of tempered distributions, respectively. They are endowed with their natural topology.

THEOREM 2.1 (see [2, Theorem 6.2]). *Suppose that $\mathcal{B}\in\{B,\mathring{B},b\}$ and that $m\in\mathbb{R}$. Then*

$$\big[a\mapsto a(D)\big] \in \mathcal{L}\Big(\mathrm{S}^m(E_0,E_1),\mathcal{L}\big(\mathcal{B}^s_{p,q}(\mathbb{R}^n,E_0),\mathcal{B}^{s+m}_{p,q}(\mathbb{R}^n,E_1)\big)\Big)$$

for $s\in\mathbb{R}$ and $p,q\in[1,\infty]$.

The symbol B denotes the regular Besov spaces, whereas $\mathring{B}^s_{p,q}$ and $b^s_{p,q}$ denote the closures of $\mathcal{S}$ and $B^{s+1}_{p,q}$ in $B^s_{p,q}$, respectively. Various equivalent definitions are given in [2]. Here only regular Besov spaces with $p=q=\infty$ of positive fractional order $s\in\mathbb{R}^+\setminus\mathbb{N}$ are used, in which case one has

$$B^s_{\infty,\infty} = \mathrm{BUC}^s(\mathbb{R}^n,E)\,,$$

where the space on the right is the standard space of bounded and uniformly Hölder continuous functions given by

$$\mathrm{BUC}^s(\mathbb{R}^n,E) := \big\{ f\in\mathrm{BUC}^{[s]}(\mathbb{R}^n,E) \,\big|\, \partial^\alpha f\in\mathrm{BUC}^{s-[s]}(\mathbb{R}^n,E)\ \forall\,|\alpha|\le[s]\,\big\}\,.$$

A crucial observation connecting the dilation of symbols of type (2.5) and singular Hölder spaces was already obtained in [16]. It allows one to deal with singular elliptic and, eventually, singular parabolic boundary value problems.

PROPOSITION 2.2 (see [16, Lemma 2.5]). *Assume that $a\in\mathrm{S}^m(E_0,E_1)$ for some $m\in\mathbb{Z}$. Then*

$$\big[t\mapsto t^m\sigma_t a\big] \in \mathrm{C}_1^{1-}\big((0,T],\mathrm{S}^m(E_0,E_1)\big)$$

for $\sigma_t(a) := a(t\cdot)$.

Remark 2.1. A straightforward adaptation of the proof given in [16] of Proposition 2.2 also shows that

$$\big[t \mapsto t^m \sigma_t a\big] \in \mathrm{C}^{1-}\big([0,T], \mathrm{S}^{m-1}(E_0, E_1)\big).$$

The importance of this proposition lies in the fact that multiplication of singularly Hölder continuous functions is well defined and continuous as a map defined in various combinations of spaces

$$\mathrm{C}^\alpha_\alpha \times \mathrm{C}^\beta_\beta \to \mathrm{C}^\beta_\beta, \tag{2.6}$$

$$\mathrm{C}^\alpha_\alpha \times \mathrm{C}^\beta_0 \to \mathrm{C}^\beta_0, \tag{2.7}$$

$$\mathrm{C}^\alpha_0 \times \mathrm{C}^\beta_0 \to \mathrm{C}^\beta_{-\alpha}, \tag{2.8}$$

$$\mathrm{C}^\alpha_0 \times \mathrm{C}^\beta_\beta \to \mathrm{C}^\beta_{\beta-\alpha} \tag{2.9}$$

for $\alpha, \beta \in (0,1) \cup \{1-\}$ and $\beta \le \alpha$. The proof is elementary and can be found in [16] along with the (natural) definition of the spaces for negative lower indices.

Finally, optimal regularity results are yet another essential ingredient in dealing with the full nonlinear problem. They provide maximal regularity results for the relevant class of singular parabolic problems in classes of singularly Hölder continuous functions which are, as they need to be, perfectly compatible with the corresponding results for singular elliptic problems in the same class of functions. The result needed here has been derived in [13] and is formulated in the next theorem. It gives conditions for the well-posedness of the singular abstract Cauchy problem

$$\dot u - A(t)u = f(t)\,, \quad t > 0, \tag{2.10}$$

for a Banach space–valued function $u : (0,T] \to E_0$ and a singular family of "elliptic operators" A (that is, of generators of analytic semigroups). The symbol $\mathcal{H}^-(E_0, \omega)$ denotes the class of generators of exponentially decaying semigroups which are not necessarily strongly continuous (as introduced by [20]).

THEOREM 2.3 (see [13, Corollary 3.3]). *Assume that A satisfies the assumptions*

$$\text{(i) } A(t) \in \mathcal{H}^-(E_0, \omega)\,, \quad t > 0\,, \tag{2.11}$$

$$\text{(ii) } \|\big[A(t) - A(s)\big]A^{-1}(\tau)\|_{\mathcal{L}(E_0)} \le c\frac{t-s}{t} \text{ and}$$

$$\|\big[A(t) - A(s)\big](-A)^{-\rho}(\tau)\|_{\mathcal{L}(E_0)} \le c(t-s)\,, \tag{2.12}$$

$$\text{(iii) } \lim_{t\to 0} A^{-1}(t) = 0\,, \tag{2.13}$$

for some $\rho \in (1,2)$ and $0 < \tau \le s \le t \le T$. Let $f \in \mathrm{C}^\beta_\gamma\big((0,T], E_0\big)$ for some $\beta \in (0,1)$ and $\gamma = 0, \beta$. Then (2.10) *has a unique solution $u \in \mathrm{C}^\beta_\gamma\big((0,T], E_0\big)$ satisfying*

$$\dot u, Au \in \mathrm{C}^\beta_\gamma \quad \text{and} \quad \|\dot u\|_{\beta,\gamma} + \|Au\|_{\beta,\gamma} \le c\|f\|_{\beta,\gamma}\,,$$

where $\|\cdot\|_{\beta,0} = \|\cdot\|_\beta$.

Using a combination of elliptic and parabolic estimates, it will be possible to obtain a satisfactory regularity theory for (2.1)–(2.3). The latter will be used in the analysis of the full evolution system (1.6)–(1.10) for $\delta = 0$.

3. Elliptic and parabolic estimates. In order to use the results presented in the previous section in the analysis of (1.6)–(1.10), appropriate function spaces have to be chosen in which to work. The choice can typically be justified by balancing the regularities in the nonlinear equations in such a way that a fixed point argument can be applied to the set of equations. Whereas a variety of function spaces of Sobolev type are available for dealing with elliptic and parabolic problems, the fact that the system of interest contains a Hamilton–Jacobi equation for the evolution of s, which, in turn, appears in differentiated form as a coefficient in the equations for the evolution of u, restricts the choice to spaces of classical regularity for s. It is indeed impossible to work in Sobolev–Slobodeckii spaces for u as the loss of regularity incurred in taking traces on the boundary cannot be made up for by (1.9), which does not possess any regularizing effect. It becomes clear that spaces of classical regularity for both u and s need to be chosen. In order to use the Fourier multiplier Theorem 2.1 the choice is therefore reduced to spaces of bounded uniformly continuous functions in the x-variable. As far as the y-variable is concerned a good choice is given by the space of continuous functions, as will become clear later. The "base" space E_0 for u is therefore chosen as

$$E_0 := \mathrm{BUC}^{1+\alpha}\big(\mathbb{R}^n, \mathrm{C}(0,1)\big) \tag{3.1}$$

for $\alpha \in (0,1)$ and where $\mathrm{C}(0,1)$ denotes the standard space of continuous functions on $[0,1]$. The space with one less regularity degree in x would seem like a more natural choice, but the choice made here makes it much easier to deal with inhomogeneous Neumann boundary conditions such as (1.8). This point will become clear later in the analysis in section 3.3. In view of the singular nonautonomous nature of the problem, the standard procedures [19] for dealing with inhomogeneous (nonlinear) boundary conditions cannot be utilized.

It follows that a proper choice of "base" space for the full evolutionary problem is given by

$$\mathbb{E}_0 := \mathrm{C}^{\beta}_{\beta}\Big((0,T], \mathrm{BUC}^{1+\alpha}\big(\mathbb{R}^n, \mathrm{C}(0,1)\big)\Big) \tag{3.2}$$

for u and by

$$\mathbb{S} := \mathrm{C}^{1+\beta}\big([0,T], \mathrm{BUC}^{2+\alpha}(\mathbb{R}^n)\big) \cap \mathrm{C}^{\beta}\big([0,T], \mathrm{BUC}^{3+\alpha}(\mathbb{R}^n)\big) \tag{3.3}$$

for s. A closer look at the underlying elliptic problem ($\varepsilon = 0$) is necessary in order to better deal with the inhomogeneous boundary terms in (2.1)–(2.3). The operator $-\triangle_x - \frac{1}{t^2 g^2(x)}\partial_y^2$ has clearly nonconstant coefficients, and multiplier Theorem 2.1 is formulated only in the translation invariant case. It therefore needs to be shown that localization arguments apply in the operator-valued setting considered here.

The first step, however, is to obtain regularity results in the constant coefficient case.

3.1. The elliptic case: Constant coefficients. Consider problem (2.1)–(2.3) and assume that $\varepsilon = 0$ and that c is a constant which can be assumed to be 1 without loss of generality. Then taking a Fourier transform in the x-variables, one obtains a parameter-dependent boundary value problem for $\hat{u}$ which can be solved explicitly. The solution has the structure

$$\hat{u}(t,\xi,\cdot) = t^2\sigma_t c(\xi,\cdot)\hat{f} + t\sigma_t b(\xi,\cdot)\hat{h} + \sigma_t a(\xi,\cdot)\hat{g}, \tag{3.4}$$

where σ_t denotes dilation by $t>0$ and the operator-valued symbols a,b,c are given by

$$a(\xi,\cdot)\;=\;\frac{\cosh\big(|\xi|(1-\cdot)\big)}{\cosh(|\xi|)}\,,\;b(\xi,\cdot)\;=\;\frac{\sinh(|\xi|\cdot)}{|\xi|\cosh(|\xi|)}\,,\;c(\xi,\cdot)\;=\;\big(|\xi|^2+C\big)^{-1}\,. \tag{3.5}$$

The first two symbols have to be considered as multiplication operator–valued with respect to the variable y, whereas the last contains the sectorial operator C which is the operator $-\partial_y^2$ on $\mathrm{C}(0,1)$ with domain

$$\mathrm{dom}(C)=\{u\in\mathrm{C}^2(0,1)\,|\,u(0)=0\,,\;\partial_y u(1)=0\}=:\mathrm{C}^2_{0,0}(0,1)\,. \tag{3.6}$$

The spaces

$$\mathrm{C}^k_0:=\{u\in\mathrm{C}^k(0,1)\,|\,u(0)=0\}\,,\quad k=0,1,$$

will also be needed. Proposition 2.2 can be used in combination with the multiplications (2.6)–(2.9) to obtain regularity results in weighted Hölder spaces, provided (3.5) are in the appropriate symbol class.

THEOREM 3.1. *The symbols* (3.5) *satisfy*

$$\begin{aligned}
&a\in\mathcal{S}^0\big(\mathrm{C}(0,1),\mathrm{C}(0,1)\big)\,,\\
&b\in\mathcal{S}^k\big(\mathrm{C}(0,1),\mathrm{C}^{1-k}_0(0,1)\big)\,,\\
&c\in\mathcal{S}^j\big(\mathrm{C}(0,1),\mathrm{C}^{2-j}_{0,0}(0,1)\big)
\end{aligned}$$

for $k=0,1$ and $j=0,1,2$. The notation for the spaces has to be interpreted so that boundary conditions are imposed only if enough regularity is available but otherwise disregarded.

Proof. Take the last symbol first. It is known that the operator C is sectorial and invertible (cf. [19]). It follows that

$$\|(|\xi|^2+C)^{-1}\|_{\mathcal{L}(\mathrm{C},\mathrm{C}^{2-j})}\le\,c\frac{1}{1+|\xi|^j}\,,\quad j=0,1,2. \tag{3.7}$$

As for the derivatives, it is seen by induction that, for $\xi\neq 0$,

$$\partial^\alpha(|\xi|^2+C)^{-1}=(|\xi|^2+C)^{-1}\sum_{|\beta|\le|\alpha|}p_{\alpha,\beta}(\xi)(|\xi|^2+C)^{-|\beta|}$$

for polynomials $p_{\alpha,\beta}$ of degree at most $|\beta|$. The estimate therefore follows from (3.7). Next consider the "boundary symbols" a and b. A simple series expansion reveals that both symbols are actually analytic functions of ξ^2 (and y of course). It follows that they are smooth for all $\xi\in\mathbb{R}^n$ in spite of appearances. It is therefore legitimate to concentrate on the decay properties of the symbols without necessarily using the nowhere vanishing weight $(1+|\xi|^2)^{1/2}$ needed in the definition of the symbol class. Observe that both are smooth functions of y regardless of ξ. Consequently they define multiplication operators for $\mathrm{C}^{2-k}(0,1)$ and $k=0,1,2$. Also observe that, for b, the smoothing effect in x is due to the decay in ξ. They clearly satisfy the required boundary conditions. They further satisfy the relations

$$\begin{aligned}
\partial_y b(\xi,y)&=a(\xi,1-y)\,,\qquad \partial_y^2 b(\xi,y)=|\xi|^2 b(\xi,y)\,,\\
\partial_y a(\xi,y)&=-|\xi|^2 b(\xi,1-y)\,,\qquad \partial_y^2 a(\xi,y)=|\xi|^2 a(\xi,y)\,,
\end{aligned}$$

which make it easier to derive their mapping and symbol properties from one another. For instance, if it were known that $a, b \in \mathcal{S}^0(\mathrm{C}, \mathrm{C})$, the first relation would imply that $\partial_y b \in \mathcal{S}^0(\mathrm{C}, \mathrm{C})$ and therefore that $b \in \mathcal{S}^0(\mathrm{C}, \mathrm{C}^1)$. The last one, on the same assumptions, would show that $a \in \mathcal{S}^{-2}(\mathrm{C}, \mathrm{C}^2)$. More details are now given for the analysis of symbol a. Corresponding claims for the symbol b can be obtained by similar calculations. The claim is that $a \in \mathcal{S}^0(C, C)$. It is plain that

$$\sup_{\xi \in \mathbb{R}^n} \sup_{y \in [0,1]} |a(\xi, y)| < \infty \,.$$

As for the first derivatives in ξ one has

$$\begin{aligned} \partial_j a(\xi, y) &= \frac{\xi_j}{|\xi|} \frac{1}{\cosh(|\xi|)} \big[(1-y) \sinh(|\xi|(1-y)) - \tanh(|\xi|) \cosh(|\xi|(1-y)) \big] \\ &= \frac{\xi_j}{|\xi|} a(\xi, y) \underbrace{\big[(1-y) \tanh(|\xi|(1-y)) - \tanh(|\xi|) \big]}_{=: d(\xi, y)} \,. \end{aligned} \tag{3.8}$$

The maximum in y is taken either on the boundary $y = 0, 1$ or in the interior. On the boundary it either vanishes or is exponentially decaying in $|\xi|$. It therefore only needs to be controlled in the interior to conclude its estimation. Only the term in brackets on the right-hand side in the first line of (3.8) depends on y. Setting its first y-derivative equal to zero, one arrives at the equation

$$\frac{\tanh(|\xi|(1-y))}{(1-y)|\xi|} = \frac{1}{|\xi| \tanh(|\xi|) - 1},$$

which, for $|\xi|$ large, is asymptotic to

$$\frac{1}{(1-y)|\xi|} \sim \frac{1}{|\xi|} \,.$$

Thus the maximum is located at $y \sim \frac{1}{|\xi|}$ for large $|\xi|$. As for the maximal value for large $|\xi|$ one has

$$\partial_j a \left(\xi, \frac{1}{|\xi|} \right) \sim \frac{\xi_j}{|\xi|} a \left(\xi, \frac{1}{|\xi|} \right) \frac{1}{|\xi|}, \tag{3.9}$$

which entails the desired estimate

$$\sup_{x \in \mathbb{R}^n} \sup_{y \in [0,1]} (1 + |\xi|^2)^{1/2} |\partial_j a(\xi, y)| < \infty \,.$$

Equation (3.9) follows from

$$\left(1 - \frac{1}{|\xi|} \right) \tanh(|\xi| - 1) - \tanh(|\xi|), \tag{3.10}$$

observing that $\tanh(s) = 1$ up to exponentially small terms in s for $s > 0$ large, that is,

$$|\tanh(s) - 1| \leq c e^{-3s} \; s > 0 \,.$$

The second derivative satisfies

$$\partial_i \partial_j a(\xi, y) = \frac{\xi_i \xi_j}{|\xi|^3} a(\xi, y) d(\xi, y) + \frac{\xi_j \xi_j}{|\xi|^2} a(\xi, y) d^2(\xi, y) + \frac{\xi_j}{|\xi|} a(\xi, y) \partial_j d(\xi, y).$$

The first term on the right-hand side can be seen to decay like $1/|\xi|^2$ in view of the explicit prefactor and the estimate obtained for the first derivative (3.8). The second and third terms add up to

$$\frac{\xi_i\xi_j}{|\xi|^2}a(\xi,y)\big[(1-y)^2-1-2(1-y)\tanh(|\xi|)\tanh(|\xi|(1-y))+2\tanh^2(|\xi|)\big].$$

Similar calculations to those performed for the first derivative show that the maximum is now attained at

$$y\sim\frac{2}{|\xi|},$$

and it amounts to

$$\partial_i\partial_j a\left(\xi,\frac{2}{|\xi|}\right)-\frac{\xi_i\xi_j}{|\xi|^3}a\left(\xi,\frac{2}{|\xi|}\right)d\left(\xi,\frac{2}{|\xi|}\right)\sim\frac{4}{|\xi|^2}\,.$$

The latter follows from

$$\left(1-\frac{2}{|\xi|}\right)^2-1-2\left(1-\frac{2}{|\xi|}\right)\tanh(|\xi|-2)\tanh(|\xi|)+2\tanh^2(|\xi|)\sim\frac{4}{|\xi|^2},$$

which uses (3.10) again. In conclusion, it is obtained that

$$\sup_{\xi\in\mathbb{R}^n}\sup_{y\in[0,1]}(1+|\xi|^2)|\partial_i\partial_j a(\xi,y)|<\infty\,.$$

Comparing this with the proof of [16, Lemma 2.2], we see that the arguments are almost identical. The only differences are the additional factors like $\frac{\xi_j}{|\xi|}$ for the first derivative or $\frac{\xi_i\xi_j}{|\xi|^3}$ and $\frac{\xi_i\xi_j}{|\xi|^2}$ for the second. Terms containing these can always be handled through their explicit dependence on ξ and the estimates from previous derivatives. The inductive argument used in [16] can therefore be adapted to the current situation and leads to the desired result. □

Combining these results with Theorem 2.1, Proposition 2.2, and (2.6)–(2.9), we obtain the next important theorem. To simplify its formulation we introduce the abbreviated notation

$$\mathrm{C}^?_?\,\mathrm{BUC}^?\,\mathrm{C}^?_{?,?}$$

for the function spaces

$$\mathrm{C}^?_?\Big((0,T],\mathrm{BUC}^?\big(\mathbb{R}^n,\mathrm{C}^?_{?,?}(0,1)\big)\Big),$$

where the question marks can be substituted by any of the relevant regularity and singularity parameters.

THEOREM 3.2. *Let $\alpha,\beta\in(0,1)$ and $\gamma=0,\beta$. Given*

$$(f,g,h)\in\mathrm{C}^\beta_\gamma\,\mathrm{BUC}^{1+\alpha}\,\mathrm{C}\times\mathrm{C}^\beta_\gamma\,\mathrm{BUC}^{3+\alpha}\times\mathrm{C}^\beta_\gamma\,\mathrm{BUC}^{2+\alpha},$$

there exists a unique solution u of (2.1)–(2.3) *for $c\equiv$ const which belongs to the space*

$$\mathbb{E}_1:=\left\{u\in\mathbb{E}_0\,\Big|\,\partial_x^\alpha\left(\frac{\partial_y}{t}\right)^j u\in\mathbb{E}_0\ \textit{for}\ |\alpha|,j=0,1,2,\ \textit{s.t.}\ |\alpha|+j\le 2\right\},$$

where

$$\mathbb{E}_0 := \mathrm{C}^{\beta}_{\gamma}\,\mathrm{BUC}^{1+\alpha}\,\mathrm{C}\,.$$

It is given by

$$u = \underbrace{\mathcal{F}^{-1}t^2\sigma_t c\mathcal{F}}_{=:A(t)^{-1}} f + \underbrace{\mathcal{F}^{-1}t\sigma_t b\mathcal{F}}_{=:R_N(t)} h + \underbrace{\mathcal{F}^{-1}\sigma_t a\mathcal{F}}_{=:R_D(t)} g\,.$$

Modulo the singularity, the terms in the above representation clearly point to their asymptotic behavior at the origin. This result needs to be extended to the nonconstant coefficient case. This is done in the next section.

Remark 3.1. To make it easier to refer back to the above result in the case $\gamma = 0$, the corresponding spaces will be denoted by $\mathbb{E}_{k,0}$, $k = 0, 1$.

3.2. The elliptic nonconstant coefficient case. In this section we adapt an abstract formulation of a localization argument proposed by Angenent in [5] (see also [4, 3]) for $\mathbb{R}^n$ to cover the case of boundary value problems in an operator-valued context. It is used in combination with the Fourier multiplier Theorem 2.1 to give the basic regularity results needed for the singular elliptic boundary value problem underlying (2.1)–(2.3) in the case of constant coefficients. It should be observed that localization arguments rely on perturbation results for differential operators by lower order terms. These are small compared to the leading order operator only in a qualitative sense. Perturbation results can therefore only be applied for the resolvents with large λ where the decay properties of the inverse to the leading order operator yield smallness of the lower order terms. The structure of the singular operators considered here can be exploited in order to avoid shifting the leading order operator and in order to obtain direct invertibility results. The standard argument of course shows that nonconstant coefficient operators are sectorial whenever their constant coefficient counterpart is (see [5]). This remains valid here and can be used if the time variable is kept fixed.

The goal of this section is to extend the validity of Theorem 3.2 to the nonconstant coefficient singular boundary value problem

$$\left(\mathcal{A}(t), \gamma_0, \frac{1}{t}\gamma_1\partial_y\right) : \mathbb{E}_1 \to \mathbb{E}_0 \times \partial\mathbb{E}_1, \tag{3.11}$$

where $\mathcal{A}(t)$ is the elliptic operator driving the evolution in (2.1) and the "boundary space" $\partial\mathbb{E}_1$ is defined by

$$\partial\mathbb{E}_1 = \mathrm{C}^{\beta}_{\gamma}\,\mathrm{BUC}^{3+\alpha} \times \mathrm{C}^{\beta}_{\gamma}\,\mathrm{BUC}^{2+\alpha} =: \partial_D\mathbb{E}_1 \times \partial_N\mathbb{E}_1\,, \quad \gamma = 0, \beta\,. \tag{3.12}$$

Observe that Theorem 3.2 simply states that the above singular boundary value problem has a bounded inverse in the given topologies and for constant coefficients. The localization procedure proposed in [5] is abstract and is based on the concept of resolution. A resolution of a Banach space E is simply a triple (F, ε, δ), where F is a Banach space and the maps $\delta : E \to F$ and $\varepsilon : F \to E$ satisfy $\varepsilon \circ \delta = \mathrm{id}_E$. The operator δ plays the role of the localizing operator, whereas ε resynthesizes the local contributions. If the space E has subspaces of interest, they should correspond to "similar subspaces" of F and should be left invariant by the resolution maps. The localization operator can be constructed as follows:

$$\delta : \mathbb{E}_j \to \oplus_{k\in\mathbb{Z}}\mathbb{F}^k_j\,, \quad u \mapsto (u_k)_{k\in\mathbb{Z}^n} := (u\varphi_{k,r})_{k\in\mathbb{Z}^n}\,, \quad j = 0, 1,$$

where $\mathbb{F}_j^k = \mathbb{E}_j$ for $k \in \mathbb{Z}^n$ and the sequence space $\oplus_{k\in\mathbb{Z}}\mathbb{F}_j^k$ is endowed with the supremum norm $\sup_{k\in\mathbb{Z}} \|u_k\|_{\mathbb{F}_j}$. The maps $(\varphi_{k,r}^2)_{k\in\mathbb{Z}^n}$ are chosen as to form a smooth resolution of the identity in $\mathbb{R}^n$,

$$\sum_{k\in\mathbb{Z}^n} \varphi_{k,r}^2 \equiv 1,$$

subordinated to a cover by cubes $Q_{k\in\mathbb{Z}^n,r}$ of fixed side length $r > 0$ obtained by translation and dilation of the standard cube of side size $r = 1$ centered at the origin. The support of $\varphi_{k,r}$ should contain $Q_{k,r}$ and be contained in the union of at most finitely many adjacent cubes. The synthesis operator is then given by

$$\varepsilon : \oplus_{k\in\mathbb{Z}}\mathbb{F}_j^k \to \mathbb{E}_j\,, \quad (u_k)_{k\in\mathbb{Z}^n} \mapsto \sum_{k\in\mathbb{Z}^n} u_k\varphi_{k,r},$$

in which case the relation defining a resolution is clearly satisfied. Observe that the mappings δ and ε are continuous. The idea is to define an operator on $\oplus_{k\in\mathbb{Z}}\mathbb{F}_1$ which parallels the boundary value problem (3.11) and to use its inverse to approximate the desired inverse of (3.11) via the use of the maps ε and δ. Define the operator $(\mathcal{A}', \mathcal{B}')$ on $\oplus_{k\in\mathbb{Z}}\mathbb{F}_1$ through

$$(\mathcal{A}^k, \mathcal{B}^k)u_k = \left(-\triangle_x u_k - \frac{1}{t^2 c_k(x)}\partial_y^2 u_k, \gamma_0 u_k, \frac{1}{t}\gamma_1\partial_y u_k\right),$$

where $c_k|_{\mathrm{supp}(\varphi_{k,r})} = c|_{\mathrm{supp}(\varphi_{k,r})}$ and c_k is otherwise smoothly extended without increasing its norm to the whole space (the details of the extension procedure can be found in [4, 3]). The boundary value problem $(\mathcal{A}^k, \mathcal{B}^k)$ can be made arbitrarily close to a constant coefficient one, say, by substituting c_k by $c_k(x_k)$ and making r small. It is therefore invertible and enjoys the properties claimed in Theorem 3.2. The diagonal operator $(\mathcal{A}', \mathcal{B}')$ is then invertible itself and maps $\oplus_{k\in\mathbb{Z}}\big[\mathbb{F}_1 \times \partial\mathbb{F}_1\big]$ to $\oplus_{k\in\mathbb{Z}}\mathbb{F}_0$, where $\partial\mathbb{F}_1$ is defined in the obvious way. Denote by $\begin{bmatrix} A'(t)^{-1} & R'_D(t) & R'_N(t)\end{bmatrix}$ its inverse. Then

$$\varepsilon \begin{bmatrix} A'(t)^{-1} & R'_D(t) & R'_N(t)\end{bmatrix} \delta \tag{3.13}$$

should represent an approximation to the solution operator for the singular boundary value problem (2.1)–(2.3). In order to show this, it needs to be shown that (3.13) is an approximate left and right inverse. Starting with the latter, compute

$$\begin{bmatrix} \mathcal{A}(t) \\ \gamma_0 \\ \frac{1}{t}\gamma_1\partial_y \end{bmatrix} \varepsilon \begin{bmatrix} A'(t)^{-1} & R'_D(t) & R'_N(t)\end{bmatrix} \delta \tag{3.14}$$

$$= \begin{bmatrix} \mathrm{id}_{\mathbb{E}_0} + [\mathcal{A}(t), \varepsilon]A'(t)^{-1}\delta & [\mathcal{A}(t)\,, \varepsilon]\, R'_D(t)\delta & [\mathcal{A}(t)\,, \varepsilon]\, R'_N(t)\delta \\ 0 & \mathrm{id}_{\gamma_0\mathbb{E}_1} & 0 \\ 0 & 0 & \mathrm{id}_{\gamma_1\partial_y\mathbb{E}_1} \end{bmatrix},$$

where $[\mathcal{A}(t), \varepsilon] = \mathcal{A}(t)\varepsilon - \varepsilon\mathcal{A}(t)'$. Invertibility would follow if

$$\mathrm{id}_{\mathbb{E}_0} + [\mathcal{A}(t), \varepsilon]A'(t)^{-1}\delta \tag{3.15}$$

were invertible. Here the specific structure of the operator $\mathcal{A}(t)$ needs to be exploited since standard perturbation arguments would only allow us to show that the resolvent

$\mathcal{R}(\lambda, \mathcal{A})$ exists for large enough λ, in which case the lower order commutator term is small. The commutator term (3.15) can be computed to give

$$\sum_{k\in\mathbb{Z}^n} \left[2(\nabla\varphi_{k,r}|\nabla\cdot) + \triangle\varphi_{k,r}\cdot\right] \left[-\triangle_x - \frac{1}{t^2c_k^2(x)}\partial_y^2\right]^{-1} \varphi_k f\,. \tag{3.16}$$

To see that the inverse appearing in (3.16) does indeed exist on the desired spaces, observe that $r > 0$ can be chosen such that

$$-\triangle_x - \frac{1}{t^2c_k^2(x)}\partial_y^2 + \triangle_x + \frac{1}{t^2c_k^2(x_k)}\partial_y^2 = \left[\frac{1}{c_k^2(x)} - \frac{1}{c_k^2(x_k)}\right]\frac{1}{t^2}\partial_y^2$$

is small in $\mathcal{L}(\mathbb{E}_1, \mathbb{E}_0)$. The realization A^k of $\mathcal{A}^k$ with homogeneous boundary conditions is therefore invertible, and the constant coefficient estimates of Theorem 3.2 carry over to it (with larger constants, of course). Now, the estimates

$$\left\|2\nabla\varphi_{k,r}\nabla\left[-\triangle_x - \frac{1}{t^2c_k^2(x)}\partial_y^2\right]^{-1}\right\|_{\mathcal{L}(\mathbb{E}_0)} \le cT \quad \text{and}$$

$$\left\|\triangle\varphi_{k,r}\left[-\triangle_x - \frac{1}{t^2c_k^2(x)}\partial_y^2\right]^{-1}\right\|_{\mathcal{L}(\mathbb{E}_0)} \le cT^2$$

show that the commutator is indeed small, at least for T small enough. Here the decay properties of the resolvent as $t \to 0$ are used as a substitute for making λ large in the standard argument. Heuristically this makes sense, since making t small, just as making λ large, makes the operator "more and more elliptic." This is due to the specific nature of the singular operator. Since the difficulty of the problem stems from the origin, making T small does not in any way weaken the result.

Next it needs to be shown that (3.13) is also a good approximation for a left inverse. To that end, observe that

$$u - \varepsilon\big(\mathcal{A}'(t), \mathcal{B}'(t)\big)^{-1}\delta\big(\mathcal{A}(t), \mathcal{B}(t)\big)u$$

is the same as

$$u - \left(\varepsilon A'(t)^{-1}\delta\mathcal{A}(t)u + \varepsilon R'_D(t)\delta\gamma_0 u + \varepsilon R'_N(t)\delta\gamma_1\frac{\partial_y}{t}u\right) = \varepsilon A'(t)^{-1}[\mathcal{A}(t), \delta]u \tag{3.17}$$

because

$$u = \varepsilon A'(t)^{-1}\mathcal{A}'(t)\delta u + \varepsilon R'_D(t)\delta\gamma_0 u + \varepsilon R'_N(t)\gamma_1\frac{\partial_y}{t}u$$

by definition. The notation

$$[\mathcal{A}(t), \delta] = \mathcal{A}'(t)\delta - \delta\mathcal{A}(t)$$

was used in (3.17). It follows that

$$\varepsilon\big(\mathcal{A}'(t), \mathcal{B}'(t)\big)^{-1}\delta\big(\mathcal{A}(t), \mathcal{B}(t)\big) = \mathrm{id}_{\mathbb{E}_1} - \varepsilon A'(t)^{-1}[\mathcal{A}(t), \delta].$$

The term containing the commutator is lower order and can be estimated just as before, exploiting the structure of $\mathcal{A}(t)$ in order to yield the invertibility of

$$\mathrm{id}_{\mathbb{E}_1} - \varepsilon A'(t)^{-1}[\mathcal{A}(t), \delta]$$

and, thus, a left inverse for the singular boundary value problem. The result is summarized in the next theorem.

THEOREM 3.3. *Let $\alpha,\beta\in(0,1)$, $\gamma=0,\beta$, and $0<c_0\le c\in \mathrm{BUC}^{1+\alpha}$. Given*

$$(f,g,h)\in \mathrm{C}^{\beta}_{\gamma}\,\mathrm{BUC}^{1+\alpha}\,\mathrm{C}\times \mathrm{C}^{\beta}_{\gamma}\,\mathrm{BUC}^{3+\alpha}\times \mathrm{C}^{\beta}_{\gamma}\,\mathrm{BUC}^{2+\alpha},$$

there exists a unique solution u of (2.1)–(2.3) which belongs to the space

$$\mathbb{E}_1:=\left\{u\in\mathbb{E}_0\,\Big|\,\partial_x^{\alpha}\left(\frac{\partial_y}{t}\right)^{j}u\in\mathbb{E}_0 \text{ for } |\alpha|,j=0,1,2 \text{ s.t. } |\alpha|+j\le 2\right\},$$

where

$$\mathbb{E}_0:=\mathrm{C}^{\beta}_{\gamma}\,\mathrm{BUC}^{1+\alpha}\,\mathrm{C}\,.$$

It naturally splits into three components,

$$u=\tilde A(t)^{-1}f+\tilde R_N(t)h+\tilde R_D(t)g,$$

with the same asymptotic behavior at the origin as in the constant coefficient case and where

$$\tilde A(t)^{-1}f=\big(\mathcal{A}(t),\mathcal{B}(t)\big)^{-1}(f,0,0)\,,\quad \tilde R_D(t)g=\big(\mathcal{A}(t),\mathcal{B}(t)\big)^{-1}(0,g,0),$$
$$\text{and } \tilde R_N(t)h=\big(\mathcal{A}(t),\mathcal{B}(t)\big)^{-1}(0,0,h),$$

respectively.

Proof. The only part of the proof missing is for the claim about the asymptotic behavior of the solution. It is obtained using the representation

$$\begin{bmatrix} T & -T[\mathcal{A}(t),\varepsilon]\,R_D'(t)\delta & -T[\mathcal{A}(t),\varepsilon]\,R_N'(t)\delta\\ 0 & \mathrm{id}_{\gamma_0\mathbb{E}_1} & 0\\ 0 & 0 & \mathrm{id}_{\gamma_1\partial_y\mathbb{E}_1}\end{bmatrix}$$

for the inverse of

$$\begin{bmatrix} \mathrm{id}_{\mathbb{E}_0}-[\mathcal{A}(t),\varepsilon]A'(t)^{-1}\delta & [\mathcal{A},\varepsilon]R_D'(t)\delta & [\mathcal{A}(t),\varepsilon]R_N'(t)\delta\\ 0 & \mathrm{id}_{\gamma_0\mathbb{E}_1} & 0\\ 0 & 0 & \mathrm{id}_{\gamma_1\partial_y\mathbb{E}_1}\end{bmatrix},$$

where $T=\big[\mathrm{id}_{\mathbb{E}_0}-[\mathcal{A}(t),\varepsilon]A'(t)^{-1}\delta\big]^{-1}$, the factorization of the resolvent implied by (3.14) and the mapping properties for the operators involved which follow from the symbol analysis combined with Theorem 2.1, Proposition 2.2, and (2.6)–(2.9). Take, for instance, the entry $T\mathcal{A}(t)\,\varepsilon\,R_D'(t)\delta$. The claim follows from

$$T\in\mathcal{L}(\mathbb{E}_0),\quad \mathcal{A}\in\mathcal{L}(\mathbb{E}_1,\mathbb{E}_0),\quad \text{and}\quad \varepsilon R_D'\delta\in\mathcal{L}(\partial_D\mathbb{E}_1,\mathbb{E}_1)\,.\qquad \square$$

From now on the tildes on the solution operators of the nonconstant coefficient case will be omitted. Since the corresponding operators in the constant coefficient case will no longer be used in the analysis, no confusion seems likely.

3.3. The parabolic problem. It is now possible to return to the analysis of the model problem (2.1)–(2.3) for $\varepsilon = 1$. If the boundary conditions are homogeneous, then Theorem 2.3 gives existence and regularity of a solution $u \in \mathbb{E}_1$ whenever $f \in \mathbb{E}_0$. The notation $\mathbb{E}_1$ (see Theorem 3.2) has been used so far to describe the regularity space of the singular elliptic boundary value problem. In the parabolic case the same notation indicates the space

$$\mathbb{E}_1 := \left\{ u \in \mathbb{E}_0 \,\middle|\, \partial_t u \in \mathbb{E}_0\,,\ \partial_x^\alpha \left(\frac{\partial_y}{t}\right)^j u \in \mathbb{E}_0 \text{ for } |\alpha|, j = 0,1,2,\ \text{s.t. } |\alpha| + j \le 2 \right\}$$

in accordance with Theorem 2.3. To obtain optimal regularity results for

$$A(t) = \triangle_x + \frac{1}{t^2 c^2(x)} C$$

it is therefore sufficient to check that conditions (2.11)–(2.13) are satisfied. The operator C was defined in (3.6). Freezing coefficient arguments such as those in [4, 3] or [5] show that $A(t)$ does indeed generate an analytic semigroup on E_0 for each fixed $t > 0$, which is exponentially decaying since C is invertible (see also the beginning paragraph of subsection 3.2). Estimates (2.12)–(2.13) follow from the regularity theory developed in the previous section and the fact that $A(t)$ generates an exponentially decaying analytic semigroup. The latter makes it possible, in particular, to define the fractional power appearing in (2.12) (see [19] for more details). Condition (2.13) follows from

$$\|A(t)^{-1}\|_{\mathcal{L}(\mathbb{E}_0)} \le cT^2,$$

which, in turn, follows from the observation that $(\frac{1}{c^2(x)} C)^{-1}$ is bounded in $\mathbb{E}_0$ and from

$$A(t)^{-1} = t^2 \left(t^2 \triangle_x + \frac{1}{c^2(x)} C \right)^{-1}.$$

The first condition in (2.12) follows from

$$\big[A(t) - A(s)\big] A(\tau)^{-1} = \frac{(t^2 - s^2)\tau^2}{t^2 s^2} \frac{C}{c^2(x)} \left[-\tau^2 \triangle_x + \frac{C}{c^2(x)} \right]^{-1}$$

and the fact that

$$\left\| \frac{C}{c^2(x)} \left[-\tau^2 \triangle_x + \frac{C}{c^2(x)} \right]^{-1} \right\|_{\mathcal{L}(\mathbb{E}_0)}$$

is uniformly bounded for $\tau \in (0, T]$. The second condition in (2.12) follows similarly by using abstract mapping properties of fractional powers and, in particular, that

$$\left\| \left[-\tau^2 \triangle_x + \frac{C}{c^2(x)} \right]^{-\rho+1} \right\|_{\mathcal{L}(\mathbb{E}_0)} \le c\,.$$

It remains to be shown that problem (2.1)–(2.3) can be solved for inhomogeneous boundary conditions as well. The result is formulated in the next theorem.

THEOREM 3.4. *Let* $c \in \mathrm{BUC}^{1+\alpha}$ *and*

(3.18)
$$f \in \mathbb{E}_0\,, \quad g \in \mathrm{C}^{\beta}_{\beta}\,\mathrm{BUC}^{3+\alpha} \cap \mathrm{C}^{1+\beta}_{\beta}\,\mathrm{BUC}^{1+\alpha}\,, \quad h \in \mathrm{C}^{\beta}_{\beta}\,\mathrm{BUC}^{2+\alpha} \cap \mathrm{C}^{1+\beta}_{\beta}\,\mathrm{BUC}^{\alpha}\,.$$

Then there exists a unique solution $u \in \mathbb{E}_1$ *of* (2.1)–(2.3).

Proof. It can be assumed without loss of generality that $f \equiv 0$. Looking for a solution u in the form

$$u = v + R_D(t)g + R_N(t)h,$$

one obtains that v satisfies

$$\dot v - A(t)v = \frac{d}{dt}\big[R_D(t)g + R_N(t)h\big]\,.$$

Using (2.1)–(2.3), it can be checked that

$$\frac{d}{dt}\big[R_D(t)g\big] = \frac{2}{t}A(t)^{-1}\triangle_x R_D(t)g + R_D(t)\dot g,$$
$$\frac{d}{dt}\big[R_N(t)h\big] = \frac{2}{t}A(t)^{-1}\triangle_x R_N(t)h + \frac{1}{t}R_N(t)h + R_N(t)\dot h\,.$$

The regularity results obtained in the previous section combined with the assumptions then imply that

$$\left\{t \mapsto \frac{d}{dt}\left[R_D(t)g + R_N(t)h\right]\right\} \in \mathbb{E}_0\,.$$

Take, for instance, $\big[t \mapsto \frac{2}{t}A(t)^{-1}\triangle_x R_D(t)g\big]$ and observe that

$$R_D \in \mathcal{L}(\partial_D\mathbb{E}_1, \mathbb{E}_1)\,, \quad \triangle_x \in \mathcal{L}(\mathbb{E}_1, \mathbb{E}_0), \quad \text{and} \quad \left[t \mapsto \frac{2}{t}A(t)^{-1}\right] \in \mathcal{L}(\mathbb{E}_0)$$

or $R_D(t)\dot g$, in which case the stated regularity follows from

$$\dot g \in \gamma_0\mathbb{E}_0 \quad \text{and} \quad R_D \in \mathcal{L}(\gamma_0\mathbb{E}_0, \mathbb{E}_0).$$

The fact that $R_D \in \mathcal{L}(\gamma_0\mathbb{E}_0, \mathbb{E}_0)$ has not been explicitly proven but can be obtained by symbol analysis and freezing coefficients along the lines of sections 3.1 and 3.2. The claim then follows from Theorem 2.3. □

4. The Hamilton–Jacobi equation. Consider the Hamilton–Jacobi equation

$$s_t - \sqrt{1 + |\nabla s|^2}\, v = 0 \quad \text{in } (0,\infty)\times\mathbb{R}^n\,, \tag{4.1}$$
$$s(0,\cdot) \equiv 0 \quad \text{on } \mathbb{R}^n \tag{4.2}$$

for a given $v \in \mathrm{C}^{\beta}\,\mathrm{BUC}^{3+\alpha}$ satisfying $v(0,\cdot) = g$. The method of characteristics allows one to recast this Hamilton–Jacobi equation as a system of ODEs in the following manner:

$$\dot t = 1\,, \qquad t(0) = 0\,, \tag{4.3}$$
$$\dot x = -\frac{p}{\sqrt{1+|p|^2}}\, v(t,x)\,, \qquad x(0) = \rho \in \mathbb{R}^n\,, \tag{4.4}$$
$$\dot z = r - \frac{|p|^2}{\sqrt{1+|p|^2}}\, v(t,x)\,, \quad z(0) = 0\,, \tag{4.5}$$
$$\dot r = \sqrt{1+|p|^2}\,\partial_t v(t,x)\,, \qquad r(0) = g(\rho), \tag{4.6}$$
$$\dot p = \sqrt{1+|p|^2}\,\nabla_x v(t,x)\,, \qquad p(0) = 0\,. \tag{4.7}$$

This system is easily seen to reduce to

$$\dot{x} = -\frac{p}{\sqrt{1+|p|^2}}\, v(t,x)\,, \quad x(0) = \rho \in \mathbb{R}^n\,, \tag{4.8}$$

$$\dot{p} = \sqrt{1+|p|^2}\, \nabla_x v(t,x)\,, \quad p(0) = 0, \tag{4.9}$$

as all other unknowns can be obtained after solving this reduced system. The assumption on v makes it possible to solve this system on a possibly small time interval $[0,T]$ which is independent of $\rho \in \mathbb{R}^n$. Exploiting the regularity assumption on v, it can easily be seen that the flow mapping

$$(X_t, P_t) : \mathbb{R}^{2n} \to \mathbb{R}^{2n}\,, \quad (\rho,\eta) \mapsto \big(x(t,\rho,\eta), p(t,\rho,\eta)\big)$$

satisfies

$$[t \mapsto (X_t, P_t)] \in \mathrm{C}^\beta\, \mathrm{BUC}^{2+\alpha}(\mathbb{R}^{2n}, \mathbb{R}^{2n}),$$

where the map is obtained by solving system (4.8)–(4.9) with the second initial condition substituted by $p(0) = \eta$. It is clearly a flow of diffeomorphisms. Furthermore, since an equation satisfied by $(D_\rho x, D_\rho p)$ is easily derived from (4.8)–(4.9) and since

$$\big(D_\rho x(0,\rho,\eta), D_\rho p(0,\rho,\eta)\big) = (\mathrm{id}_{\mathbb{R}^n}, 0)\,,$$

it follows that $D_\rho X_t(\rho, 0)$ satisfies

$$\|\,\mathrm{id}_{\mathbb{R}^n} - D_\rho X_t(\rho,0)\| \le \frac{1}{2}$$

uniformly in $\rho \in \mathbb{R}^n$. This implies that $X_t(\cdot, 0)$ is a diffeomorphism and yields uniform estimates for the inverse of X_t. It follows that

$$[t \mapsto (X_t)^*]\,,\ [t \mapsto (X_t)_*] \in \mathrm{C}^\beta\, \mathcal{L}(\mathrm{BUC}^{2+\alpha}), \tag{4.10}$$

where we denoted the pull-back and push-forward with X_t by $(X_t)^*$ and $(X_t)_*$, respectively. Moreover, these operators are uniformly bounded in the norm in a small time interval.

THEOREM 4.1. *Assume that $v \in \mathrm{C}^\beta\, \mathrm{BUC}^{3+\alpha}$ is given with $v(0) = g \in \mathrm{BUC}^{3+\alpha}$. Then, for $T > 0$ small enough, there exists a unique solution s of (4.1)–(4.2) on $[0,T]$ with*

$$s \in \mathrm{C}^{1+\beta}\, \mathrm{BUC}^{2+\alpha} \cap \mathrm{C}^\beta\, \mathrm{BUC}^{3+\alpha}\ .$$

The existence interval is independent of v in a neighborhood of g. Furthermore, if $v_1, v_2 \in \mathrm{C}^\beta\, \mathrm{BUC}^{3+\alpha}$ are such that $v_1(0) = v_2(0) = g$, then

$$\begin{aligned}\|s_1 - s_2\|_{\mathrm{C}^{1+\beta}\,\mathrm{BUC}^{2+\alpha} \cap \mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha}} \\ \le\ cT\, \|v_1 - v_2\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha}} + c\, \|v_1 - v_2\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{2+\alpha}}\,.\end{aligned} \tag{4.11}$$

Proof. The solution can be computed by solving the reduced system of characteristic equations (4.8)–(4.9), which produces in particular the diffeomorphisms X_t. A minimal interval of existence can be chosen independently of the initial conditions thanks to the assumptions on v. Next

$$\nabla_x s(t,x) = p\big(t, X_t^{-1}(x), 0\big),$$

together with the discussion preceding the formulation of the theorem, in particular (4.10), shows that

$$s \in \mathrm{C}^{\beta}\,\mathrm{BUC}^{3+\alpha}\ .$$

The desired time regularity has to be backed out from (4.5)–(4.6). The time derivative appearing on the right-hand side of (4.6) is obviously not welcome in view of the assumptions made on v. It is, however, possible to rewrite the term containing said time derivative while integrating the equation to read

$$\begin{aligned}(4.12)\quad r(t,x) &= \sqrt{1+|p(t,x)|^2}v\big(t,x\big)\\ &\quad -\int_0^t [v\nabla_x v\cdot p]\big(\tau, X_\tau(X_t^{-1}(x))\big)\,d\tau - \int_0^t \left[\frac{|p|^2}{\sqrt{1+|p|^2}}v\right]\big(\tau, X_\tau(X_t^{-1}(x))\big)\,d\tau\,.\end{aligned}$$

The solution s can then be obtained by one further integration from (4.5). By (4.10) and the regularity assumption on v it then follows that

$$s \in \mathrm{C}^{1+\beta}\,\mathrm{BUC}^{2+\alpha}\ .$$

The use of this argument needs to be justified. This can be done by substituting v by a regularized version of it which is differentiable in time and for which the partial integration used to obtain (4.12) can safely be performed. Letting it converge back to the original function v produces a function satisfying the modified equation. Since the solution of the original problem can be constructed by solving only the reduced problem (4.8)–(4.9), no trouble is encountered in taking the limit.

The additional estimate follows from the characteristic system (4.3)–(4.7), the regularity assumptions on v_1, v_2, and the fact that $v_1(0) - v_2(0) = 0$. □

5. Existence result. The local existence result for regular solutions of (1.1)–(1.5) is based on the analysis of (1.6)–(1.10). The regularity theory for linear singular elliptic and parabolic problems described in the previous sections and the analysis for the Hamilton–Jacobi equation performed in the last section will be the main tools. It should be kept in mind that $\varepsilon = 1$ without loss of generality. To better highlight the structure of the problem and the use of the linear regularity theory, it is convenient to rewrite (1.6)–(1.8) using (1.9) in the form

$$\dot u - \widetilde{\mathcal{A}}(t)u = \overline{\mathcal{A}}(s)u \quad \text{in } S, \tag{5.1}$$

$$u = g \qquad \text{on } \Gamma_0, \tag{5.2}$$

$$\frac{1}{t}\partial_y u = H(s,u) \text{ on } \Gamma_1, \tag{5.3}$$

where $\widetilde{\mathcal{A}}(t) = \triangle_x + \frac{1}{t^2 g^2(x)}\partial_y^2 + \frac{y}{t}\partial_y$ and the operator $\overline{\mathcal{A}}(s)$ is defined by

$$\begin{aligned}\overline{\mathcal{A}}(s)u = \left[\frac{1+y^2|\nabla s|^2}{s^2} - \frac{1}{t^2 g^2(x)}\right]\partial_y^2 u + y\left[\frac{\dot s}{s} - \frac{1}{t}\right]\partial_y u - 2y\frac{1}{s}\big(\nabla s\big|\partial_y\nabla u\big)\\ - y\frac{s\triangle s - 2|\nabla s|^2}{s^2}\partial_y u\end{aligned}$$

and $H(s,u)$ is given by

$$H(s,u) = \frac{s}{t}\frac{1}{1+|\nabla s|^2}\big(\nabla s\big|\nabla u\big) - \frac{s}{t}\frac{1}{\sqrt{1+|\nabla s|^2}}u(1+u)\,,$$

respectively.

Remark 5.1. It should be observed that the operator family $\widetilde{\mathcal{A}}(t)$ does not coincide with the one considered in the sections devoted to elliptic and parabolic regularity theory and denoted by $A(t)$. It is, however, easy to check that it enjoys the same "singular" elliptic regularity theory, as a perturbation argument shows:

$$\widetilde{A}(t) = A(t)\left[\mathrm{id}_{\mathbb{E}_1} + A^{-1}(t)\frac{y}{t}\partial_y\right] = \left[\mathrm{id}_{\mathbb{E}_0} + \frac{y}{t}\partial_y A^{-1}(t)\right] A(t).$$

Associated with the new family, there will be boundary solution operators $\widetilde{R}_D(t)$ and $\widetilde{R}_N(t)$ with corresponding mapping properties. These new boundary operators are actually given by

$$\widetilde{R}_D(t) = R_D(t) + \widetilde{A}(t)^{-1}\left[\frac{y}{t}\partial_y R_D(t)\right] \quad \text{and} \quad \widetilde{R}_N(t) = R_N(t) + \widetilde{A}(t)^{-1}\left[\frac{y}{t}\partial_y R_N(t)\right],$$

where R_D and R_N are the original boundary operators. From now on the old notation will apply to the new operators.

Assume now that $s \in \mathbb{S}$, that is, that $s \in \mathrm{C}^{1+\beta}\,\mathrm{BUC}^{2+\alpha} \cap \mathrm{C}^{\beta}\,\mathrm{BUC}^{3+\alpha}$ is chosen in a small ball about the function $[t \mapsto tg(x)]$ (in the natural norm of $\mathbb{S}$) and satisfying

$$s(0) = 0\,, \qquad \dot{s}(0) = g\,. \tag{5.4}$$

It can be verified that the operator $\overline{\mathcal{A}}(s) \in \mathcal{L}(\mathbb{E}_1, \mathbb{E}_0)$ is small in the operator norm if the time interval length $T > 0$ is chosen small (this is the reason the modified operator $\widetilde{A}(t)$ is introduced; see Remark 5.1). This structure will be useful in the proof of the following existence result.

THEOREM 5.1. *Let $s \in \mathbb{S}$ be given with the above properties* (5.4)*, $g \in \mathrm{BUC}^{4+\alpha}$, and let $T > 0$ be small. Then there exists a unique solution $u \in \mathbb{E}_1$ of* (5.1)–(5.3)*, which is the fixed point of the operator*

$$\Phi_1(u) = v + R_D(t)g + R_N(t)H(s,u),$$

where v solves

$$\dot{v} - A(t)v = \overline{\mathcal{A}}(s)u - \frac{2}{t}A^{-1}(t)\left[\triangle_x + \frac{y}{t}\partial_y\right]\left[R_D(t)g + R_N(t)H(s,u)\right]. \tag{5.5}$$

Moreover, for $s_1, s_2 \in \mathbb{S}$, the estimates

$$\begin{aligned} \|\gamma_1 u_1 - \gamma_1 u_2\|_{\mathrm{C}^{\beta}\,\mathrm{BUC}^{3+\beta}} &\le c\|s_1 - s_2\|_{\mathbb{S}^{\beta}}\,, \\ \|\gamma_1 u_1 - \gamma_1 u_2\|_{\mathrm{C}^{\beta}\,\mathrm{BUC}^{2+\beta}} &\le cT^{\beta}\|s_1 - s_2\|_{\mathbb{S}^{\beta}} \end{aligned} \tag{5.6}$$

hold.

Proof. As already pointed out, the existence proof relies on maximal regularity results obtained in the previous sections, which are at the heart of the matter and now make it possible to use a simple Banach fixed point argument. The fixed point u will be looked for in the set $\{u \in \overline{\mathbb{B}}_{\mathbb{E}_1}(g,r) \mid u(0) = g\}$ for positive r to be fixed later. This set is endowed with the $\mathbb{E}_1$ topology and is therefore complete. It is easy but tedious to show that $v, R_N(t)H(s,u)$ are in $\mathbb{E}_1 \cap \mathbb{E}_{0,0}$. The first regularity claim follows directly from Theorems 3.3 and 3.4 both for v and $R_N(t)H(s,u)$. The vanishing property for v follows from its regularity using the equations it satisfies. Its regularity readily implies that $\partial_y^2 v(t=0) \equiv 0$. Since v satisfies homogeneous boundary conditions

$$v(y=0) = 0 = \partial_y v(y=1)$$

it then is $v(t=0)\equiv 0$. As for the boundary term $R_N(t)H(s,u)$ it follows from

$$\frac{1}{t}R_N(t)\in\mathcal{L}\big(\mathrm{C}_0^\beta\,\mathrm{BUC}^{1+\alpha},\mathbb{E}_{0,0}\big)$$

combined with

$$tH(s,u)\in\mathrm{C}_0^\beta\,\mathrm{BUC}^{1+\alpha}$$

or, alternatively, also from the equations. Finally, the equations also imply that

$$\partial_y^2R_D(0)g\equiv 0\,,\quad R_D(0)g(y=0)=g\,,\quad \partial_yR_D(0)g(y=1)=0\,.$$

It clearly follows that $R_D(0)g\equiv g$. Summarizing, it is obtained that $\Phi_1(u)(0)=g$. Next a norm estimate for $\Phi_1(u)-g$ is needed. The term $R_D(t)g-g$ can be estimated by some fixed constant r_1. As for the other boundary term, one has that

$$\left\|\frac{R_N(t)}{t}[tH(s,u)]\right\|_{\mathbb{E}_1}\ \le\ cr(1+r)T^\beta\|s\|_{\mathbb{S}},$$

and it can therefore be made arbitrarily small by reducing the interval length. The above estimate is possible since the map $[t\mapsto tH(s,u)]$ vanishes at the origin in view of the properties of s. Next v has to be estimated. Most terms can simply be estimated in a way that they produce a factor T^β and can be made arbitrarily small by interval length reduction. This is due to the fact that

$$\|v\|_{\mathbb{E}_1}\le\|\big(\partial_t-A(t)\big)^{-1}\|_{\mathcal{L}(\mathbb{E}_0,\mathbb{E}_1)}\|F(s,u)\|_{\mathbb{E}_0},$$

where $F(s,u)$ summarizes the terms on the right-hand side of (5.5), and the specifics of F. Consider, for instance, the first term $\overline{\mathcal{A}}(s)u$. Since $\overline{\mathcal{A}}(s)$ vanishes at the origin, it can be seen that

$$\|\overline{\mathcal{A}}(s)\|_{\mathcal{L}(\mathbb{E}_1,\mathbb{E}_0)}\le cT^\beta\|\dot s\|_{\mathcal{S}},$$

which implies the claimed estimate. The estimate for the second term follows from the fact that

$$\left\|\frac{1}{t}A(t)^{-1}\right\|_{\mathcal{L}(\mathbb{E}_0)}\le cT$$

and the general mapping properties of the operators involved. All remaining terms but one can be handled similarly. All terms in $\frac{d}{dt}[tH(s,u)]$ containing either s or ∇s are going to lead to more terms vanishing at the origin. The term

$$\frac{R_N(t)}{t}\left[\frac{\dot s}{\sqrt{1+|\nabla s|^2}}u(1+u)\right]$$

behaves slightly differently and needs separate consideration since $\dot s u(1+u)/\sqrt{1+|\nabla s|^2}$ does not vanish at the origin. Rewriting this term as

$$\frac{R_N(t)}{t}\left[\frac{\dot s}{\sqrt{1+|\nabla s|^2}}u(1+u)-g^2(1+g)\right]+\frac{R_N(t)}{t}[g^2(1+g)]$$

and using the fact that $u(0) = g$ up to vanishing terms of type T^β, it follows that

$$\left\| \frac{R_N(t)}{t} \left[\frac{\dot{s}}{\sqrt{1+|\nabla s|^2}} u(1+u) \right] \right\|_{\mathbb{E}_0} \leq r_2 + cr^2 T^\beta \|s\|_{\mathbb{S}},$$

where the constant r_2 is given by

$$r_2 = \left\| \big(\partial_t - A(t)\big)^{-1} \frac{R_N(t)}{t} [g^2(1+g)] \right\|_{\mathbb{E}_1}.$$

It follows that Φ_1 is a self-map if r is chosen to satisfy $r > r_1 + r_2$ and the time interval length is sufficiently small. It is important to observe that all estimates have constants which are independent of s in the chosen neighborhood.

Next it needs to be shown that Φ_1 is a contraction. The calculations are very similar to the above. The terms

$$\Phi(u_1) - \Phi_1(u_2) = v_1 - v_2 + R_N(t)\big[H(s,u_1) - H(s,u_2)\big]$$

need to be estimated. The difference $v_1 - v_2$ satisfies

$$v_1 - v_2 = \big(\partial_t - A(t)\big)^{-1}\big[F(s,u_1) - F(s,u_2)\big].$$

Maximal regularity yields that

$$\|v_1 - v_2\|_{\mathbb{E}_1} = \|\big(\partial_t - A(t)\big)^{-1}\big[F(s,u_1) - F(s,u_2)\big]\|_{\mathbb{E}_1} \leq c\,\|F(s,u_1) - F(s,u_2)\|_{\mathbb{E}_0}$$

so that only the last term needs to be estimated. As for the other summand, an estimate for

$$\|R_N(t)\big[H(s,u_1) - H(s,u_2)\big]\|_{\mathbb{E}_1}$$

needs to be established. Using similar estimates to those for the self-map property and observing that all the terms not vanishing at the origin drop out since they do not depend on the unknowns, one arrives at

$$\|\Phi_1(s,u_1) - \Phi_1(s,u_2)\|_{\mathbb{E}_1} \leq cT^\beta \|u_1 - u_2\|_{\mathbb{E}_1}$$

with a constant which is independent of s in the chosen neighborhood. Existence of a unique fixed point follows by making the interval length short enough.

Next observe that any solution actually satisfies

$$\gamma_1 u = \gamma_1 R_D(t) g + \gamma_1 [v + R_N(t) H(s,u)] \in \mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha} \tag{5.7}$$

by Remark 2.1 since $g \in \mathrm{BUC}^{4+\alpha}$. This follows from

$$v = \big(\partial_t - A(t)\big)^{-1}\left[F(s,u) + \frac{1}{t} R_N(t) g^2(1+g)\right] - \big(\partial_t - A(t)\big)^{-1} \frac{1}{t} R_N(t) g^2(1+g)$$

and

$$R_N(t)H(s,u) = R_N(t)[H(s,u) - g^2(1+g)] + R_N(t)g^2(1+g),$$

where the first terms on the right-hand sides above belong to $\mathbb{E}_{1,0}$ and the second ones are not singular due to the regularity of g, and therefore of $g^2(1+g)$, combined

with Remark 2.1. The first continuous dependence result follows along the lines of the above estimates from the fact that the solution to the initial boundary value problem depends smoothly (linearly) on the interior and boundary data $\overline{\mathcal{A}}(s)u$ and $H(s,u)$. The only difference stems from the fact that a stronger norm is now estimated which leads to the absence of the factor T^β in front of $\|s_1 - s_2\|_{\mathbb{S}}$. This factor can be regained if a weaker norm is estimated in view of the "desingularizing" properties of $[\partial_t - A(t)]^{-1}$, which leads to the stated continuous dependence estimate. The estimates are based on

$$\begin{aligned}\|\gamma_1 u_1 - \gamma_1 u_2\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{3+\beta}} &\le \|\gamma_1\Phi_1(s_1,u_1) - \gamma_1\Phi_1(s_2,u_2)\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{3+\beta}}\\ &\le \|\gamma_1\Phi_1(s_1,u_1) - \gamma_1\Phi_1(s_1,u_2)\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{3+\beta}}\\ &\quad + \|\gamma_1\Phi_1(s_1,u_2) - \gamma_1\Phi_1(s_2,u_2)\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{3+\beta}}.\end{aligned}$$

The estimate for the first term of $\overline{\mathcal{A}}(s_1) - \overline{\mathcal{A}}(s_2)$ is typical. It gives

$$\begin{aligned}&\left\|\left[\frac{t^2}{s_1^2}(1+y^2|\nabla s_1|^2) - \frac{t^2}{s_2^2}(1+y^2|\nabla s_2|^2)\right]\frac{\partial_y^2}{t^2}\right\|_{\mathcal{L}(\mathbb{E}_1,\mathbb{E}_{0,0})}\\ &\qquad\le \left\|\frac{t^2}{s_1^2}(1+y^2|\nabla s_1|^2) - \frac{t^2}{s_2^2}(1+y^2|\nabla s_2|^2)\right\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{1+\alpha}\,\mathrm{C}}\left\|\frac{\partial_y^2}{t^2}\right\|_{\mathcal{L}(\mathbb{E}_1,\mathbb{E}_0)},\end{aligned}$$

which easily entails the claim since

$$\begin{aligned}&\left\|\frac{t^2}{s_1^2}(1+y^2|\nabla s_1|^2) - \frac{t^2}{s_2^2}(1+y^2|\nabla s_2|^2\right\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{1+\alpha}\,\mathrm{C}}\\ &\qquad\le c\big(\|\dot s_1 - \dot s_2\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{1+\alpha}} + \|s_1 - s_2\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{2+\alpha}}\big) \le c\|s_1 - s_2\|_{\mathcal{S}}\,.\end{aligned}$$

One is eventually led to

$$\|\gamma_1 u_1 - \gamma_1 u_2\|_{\mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha}} \le\ cT^\beta\|u_1 - u_2\|_{\mathbb{E}_1} + \|s_1 - s_2\|_{\mathcal{S}^\beta},$$

from which the claim follows since the first term on the right can be absorbed on the left-hand side. □

All pieces are now in place in order to show existence for the original system of equations (1.6)–(1.10). By denoting with $u = \Phi_1(s)$ the solution of (1.6)–(1.8) for a given $s \in \mathcal{S}$, it follows that

$$u \in \mathbb{E}_1 \quad\text{and}\quad \gamma_1 u \in \mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha}\,.$$

Decomposition (5.7) shows that u is in a given neighborhood of g for any choice of $s \in \mathcal{S}$ in a neighborhood of $[t \mapsto tg(x)]$. Similarly let $\Phi_2(u)$ be the solution of (1.9)–(1.10) constructed in section 4. In this case, if $u \in \mathbb{E}_1$ with $\gamma_1 u \in \mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha}$, $\Phi_2(u) \in \mathcal{S}$ will be in a neighborhood of $tg(x)$ for u in a neighborhood of g. It follows that a solution to the full problem can be found by producing a fixed point s for the map $\Phi := \Phi_2 \circ \Phi_1$ and defining $u := \Phi_1(s)$. We summarize in the following theorem.

THEOREM 5.2. *For any given $g \in \mathrm{BUC}^{4+\alpha}$ such that $g(x) \ge g_0 > 0$, $x \in \mathbb{R}^n$, there exists a unique local solution (u,s) of the free boundary problem* (1.6)–(1.10) *such that*

$$\begin{aligned}u, \dot u, \partial_x^\alpha\frac{1}{t^k}\partial_y^k u \in \mathrm{C}^\beta_\beta\,\mathrm{BUC}^{1+\alpha}\,\mathrm{C}\ \textit{ for } 0 \le |\alpha| + k \le 2,\\ \gamma_1 u \in \mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha},\ \textit{and}\\ s \in \mathrm{C}^{1+\beta}\,\mathrm{BUC}^{2+\alpha} \cap \mathrm{C}^\beta\,\mathrm{BUC}^{3+\alpha}\,.\end{aligned}$$

Furthermore

$$u = R_D(t)g + R_N(t)H(s,u) + v,$$

where v is a solution of (5.5) *and the different terms have different asymptotic behavior in the origin. They behave, respectively, like t^k for $k = 0, 1, 2$ modulo the singular behavior built into the space chosen.*

Proof. Making the time interval as small as needed and choosing $r > 0$ large enough, it can be seen that (Φ_1, Φ_2) is a self-map on the complete set

$$\{u \in \overline{\mathbb{B}}_{\mathbb{E}_1}(0,r) \,|\, u(0) = g\} \times \{s \in \overline{\mathbb{B}}_{\mathbb{S}}(gt, g_0/2) \,|\, s(0) = 0\,,\ \dot{s}(0) = g\}\,,$$

where the additional requirements $u(0) \equiv g\,,\ s(0) = 0\,,\ \dot{s}(0) \equiv g$ are included in the definition of the balls. Combining the estimates of Theorems 4.1 and 5.1, it follows that

$$\|\Phi(s_1) - \Phi(s_2)\|_{\mathbb{S}^\beta} \le c\,T^\beta\,\|s_1 - s_2\|_{\mathbb{S}^\beta}$$

and the contraction principle can be applied to obtain existence. ∎

Remark 5.2. It should be pointed out that this is the first well-posedness result for this class of singular FBPs in more than one space dimension for the full evolutionary problem. A companion result for the quasi-stationary approximation ($\varepsilon = 0$) has previously been obtained in [16].

REFERENCES

[1] T. Alfrey, E. F. Gurnee, and W. G. Lloyd, *Diffusion in glassy polymers*, J. Polymer Sci. Part C, 12 (1966), pp. 249–261.

[2] H. Amann, *Operator-valued Fourier multipliers, vector-valued Besov-spaces, and applications*, Math. Nachr., 186 (1997), pp. 5–56.

[3] H. Amann, *Elliptic operators with infinite-dimensional state space*, J. Evol. Equ., 1 (2001), pp. 143–188.

[4] H. Amann, M. Hieber, and G. Simonett, *Bounded H_∞-calculus for elliptic operators*, Differential Integral Equations, 7 (1994), pp. 613–653.

[5] S. Angenent, *Constructions with analytic semigroups and abstract exponential decay results for eigenfunctions*, in Topics in Nonlinear Analysis, Progr. Nonlinear Differential Equations Appl. 35, Birkhäuser, Basel, 1999, pp. 11–27.

[6] G. Astarita and G. C. Sarti, *A class of mathematical models for sorption of swelling solvents in glassy polymers*, Polymer Engrg. Sci., 18 (1978), pp. 388–395.

[7] D. S. Cohen and T. Erneux, *Free boundary problems in controlled release pharmaceuticals. I: Diffusion in glassy polymers*, SIAM J. Appl. Math., 48 (1988), pp. 1451–1465.

[8] A. Fasano, G. H. Meyer, and M. Primicerio, *On a problem in the polymer industry: Theoretical and numerical investigation on swelling*, SIAM J. Math. Anal., 17 (1986), pp. 945–960.

[9] A. Fasano and M. Primicerio, *General free-boundary problems for the heat equation.* I, J. Math. Anal. Appl., 57 (1977), pp. 694–723.

[10] A. Fasano and M. Primicerio, *General free-boundary problems for the heat equation.* II, J. Math. Anal. Appl., 58 (1977), pp. 202–231.

[11] A. Fasano and M. Primicerio, *General free-boundary problems for the heat equation.* III, J. Math. Anal. Appl., 59 (1977), pp. 1–14.

[12] A. Friedman, *Partial Differential Equations of Parabolic Type*, Kruger, Malabar, FL, 1983.

[13] P. Guidotti, *Optimal regularity for a class of singular Cauchy problems*, J. Differential Equations, to appear.

[14] P. Guidotti, *Diffusion in glassy polymers: A free boundary problem*, Adv. Math. Sci. Appl., 7 (1997), pp. 675–693.

[15] P. Guidotti, *Singular quasilinear abstract Cauchy problems*, Nonlinear Anal., 32 (1998), pp. 667–695.

[16] P. Guidotti, *A 2d free boundary value problem and singular elliptic boundary value problems*, J. Evol. Equ., 2 (2002), pp. 395–424.

[17] P. Guidotti and J. A. Pelesko, *Transient instability in case* II *diffusion*, J. Polymer Sci. Part B., 36 (1998), pp. 2941–2947.
[18] W. T. Kyner, *An existence and uniqueness theorem for a nonlinear Stefan problem*, J. Math. Mech., 8 (1959), pp. 483–498.
[19] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, 1995.
[20] E. Sinestrari, *On the abstract Cauchy problem in spaces of continuous functions*, J. Math. Anal. Appl., 107 (1985), pp. 16–66.
[21] F. Weber, *On products on noncommuting sectorial operators*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 27 (1998), pp. 499–531.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 2005–2041

ON MULTIPLE RADIAL SOLUTIONS OF A SINGULARLY PERTURBED NONLINEAR ELLIPTIC SYSTEM*

EDWARD NORMAN DANCER†, XIAOFENG REN‡, AND SHUSEN YAN§

Abstract. We study radial solutions of a singularly perturbed nonlinear elliptic system of the FitzHugh–Nagumo type. In a particular parameter range, we find a large number of layered solutions. First we show the existence of solutions whose layers are well separated from each other and also separated from the origin and the boundary of the domain. Some of these solutions are local minimizers of a related functional while the others are critical points of saddle type. Although the local minimizers may be studied by the Γ-convergence method, the reduction procedure presented in this paper gives a more unified approach that shows the existence of both local minimizers and saddle points. Critical points of both types are all found in the reduced finite dimensional problem. The reduced finite dimensional problem is solved by a topological degree argument. Next we construct solutions with odd numbers of layers that cluster near the boundary, again using the reduction method. In this case the reduced finite dimensional problem is solved by a maximization argument.

Key words. radial solutions, singularly perturbed elliptic system, Lyapunov–Schmidt reduction, Γ-convergence

AMS subject classifications. 35J50, 35J55

DOI. 10.1137/050643507

1. Introduction.

We consider the singularly perturbed elliptic system

$$\begin{cases} -\epsilon^2 \Delta u + f(u) + \epsilon\gamma v = 0 \ \text{ in } \Omega, \\ -\Delta v + v - u = 0 \ \text{ in } \Omega, \\ \partial_\nu u = \partial_\nu v = 0 \ \text{ on } \partial\Omega \end{cases} \tag{1.1}$$

on a smooth bounded domain Ω. The perturbation parameter ϵ is positive and small. The outward normal derivatives of u and v on the boundary of Ω are denoted by $\partial_\nu u$ and $\partial_\nu v$, respectively.

The nonlinear function f in (1.1) is the cubic polynomial

$$f(u) = (u-a)\left(u - \frac{a+b}{2}\right)(u-b). \tag{1.2}$$

It has three zeros a, $\frac{a+b}{2}$, and b, in the increasing order. The function is balanced in the sense that

$$\int_a^b f(q)\, dq = 0. \tag{1.3}$$

The nonlinearity in the system (1.1) is of the FitzHugh–Nagumo type. It was originally proposed to study nerve impulses [10, 18]. The phenomenon that is modeled

*Received by the editors October 25, 2005; accepted for publication (in revised form) October 30, 2006; published electronically April 6, 2007.
http://www.siam.org/journals/sima/38-6/64350.html
†School of Mathematics and Statistics, University of Sydney, NSW 2005 Australia (E.Dancer@maths.usyd.edu.au). Supported in part by ARC.
‡Department of Mathematics and Statistics, Utah State University, Logan, UT (ren@math.usu.edu). Supported in part by NSF grant DMS-0509725.
§School of Mathematics, Statistics and Computer Science, The University of New England, Armidale, NSW 2351, Australia (syan@mcs.une.edu.au).

is the control of the electrical potential across a cell membrane. This control is done by the change of flow of the ionic channels of the cell membrane. This results in the change in potential which is used to send electrical signals between cells. This is readily observed in muscle and other excitable cells. The two variables in the system are the excitable variable u and the recovery variable v. The dynamics of the two variables are described by the reaction-diffusion system

$$\begin{cases} u_t = \epsilon^2 \Delta u - f(u) - \epsilon\gamma v, \\ \kappa v_t = \Delta v - v + u, \\ \partial_\nu u = \partial_\nu v = 0 \ \text{ on } \partial\Omega. \end{cases} \tag{1.4}$$

Steady state solutions of (1.1) often have layered structures. In most parts of the domain Ω, a solution u is close to a or b. However, there exist small regions in Ω where the value of u changes abruptly from a to b. These regions are called transition layers or interfaces.

The parameter range in this paper differs from the more extensively studied one where ϵ does not appear in the $\epsilon\gamma v$ term in the first equation of (1.1) (see, for example, [4, 3, 5, 7, 6, 9, 14, 11, 12, 20, 19, 21, 24]). We will show that the parameter range considered in this paper typically gives solutions with a finite number of interior layers. In the parameter range without ϵ in the $\epsilon\gamma v$ term, the number of interior layers of a solution typically approaches infinity as $\epsilon \to 0$ (see [4, 17, 23, 1] for this type of phenomenon). The reason for this difference is that with ϵ in $\epsilon\gamma v$ there is less impact from the coupling effect with v, and hence there are fewer layers in a solution.

If we solve the second equation in (1.1) for v in terms of u with the boundary condition $\partial_\nu v = 0$ on $\partial\Omega$, i.e., $v = (1-\Delta)^{-1}u$, and substitute the solution into the first equation, we obtain the equation for u:

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

This integro-differential equation can be viewed as the Euler–Lagrange equation of the functional

$$I_\epsilon(u) = \int_\Omega \left\{ \frac{\epsilon^2}{2}|\nabla u|^2 + W(u) + \frac{\epsilon\gamma}{2}|(1-\Delta)^{-1/2}u|^2 \right\} dx. \tag{1.6}$$

Here W is an antiderivative of f; i.e.,

$$W(u) = \int_a^u f(q)\, dq. \tag{1.7}$$

Note that $W(u) \geq 0$ for all $u \in (-\infty, \infty)$ and $W(u) = 0$ if and only if $u = a$ or $u = b$. Note that $(1-\Delta)^{-1/2}$ is a nonlocal linear operator. One first defines $(1-\Delta)^{-1}$ so that $v = (1-\Delta)^{-1}$ is the solution of

$$-\Delta v + v = u \text{ in } \Omega, \ \partial_\nu v = 0 \text{ on } \partial\Omega.$$

Since $(1-\Delta)^{-1}$ is a positive operator from $L^2(\Omega)$ to itself, we define $(1-\Delta)^{-1/2}$ to be the positive square root of $(1-\Delta)^{-1}$.

The fast inhibitor limit of (1.4) is the parabolic-elliptic system

$$\begin{cases} u_t = \epsilon^2 \Delta u - f(u) - \epsilon\gamma v, \\ 0 = \Delta v - v + u, \\ \partial_\nu u = \partial_\nu v = 0 \ \text{ on } \partial\Omega \end{cases} \tag{1.8}$$

obtained by setting $\kappa = 0$ in (1.4). This is the gradient flow of I_ϵ in the $L^2(\Omega)$ space. Finding and classifying the critical points of I_ϵ (solutions of (1.1)) help us understand the behavior of (1.8).

In this paper we look for radial solutions of (1.1) on a unit ball in R^n:

$$\Omega = \{x \in R^n : |x| < 1\}. \tag{1.9}$$

The functional I_ϵ is therefore defined in the admissible set of radial $W^{1,2}$ functions:

$$\{u \in W^{1,2}(\Omega) : u = u(|x|)\}. \tag{1.10}$$

In the first part of the paper, we study layered solutions whose interfaces are well separated and away from the origin and the boundary. We prove two theorems.

THEOREM 1.1. *Let $a < 0 < b$ and K be a positive integer. There exists $\gamma_0 > 0$ such that for each $\gamma > \gamma_0$, there exists $\epsilon_0 > 0$ so that when $\epsilon \in (0, \epsilon_0)$ there are four solutions, each of which has K interfaces. Two of the four solutions, if denoted by u_ϵ^a, satisfy $\lim_{\epsilon\to 0} u_\epsilon^a(0) = a$, and the other two, if denoted by u_ϵ^b, satisfy $\lim_{\epsilon\to 0} u_\epsilon^b(0) = b$.*

THEOREM 1.2.

1. *If $a < b < 0$ and $\gamma > \frac{(n-1)\tau}{(b-a)a}$, there exists $\epsilon_0 > 0$ such that for each $\epsilon < \epsilon_0$, there is a one-interface solution u_ϵ with the property $\lim_{\epsilon\to 0} u_\epsilon(0) = b$.*
2. *If $0 < a < b$ and $\gamma > \frac{(n-1)\tau}{(b-a)a}$, there exists $\epsilon_0 > 0$ such that for each $\epsilon < \epsilon_0$, there is a one-interface solution u_ϵ with the property $\lim_{\epsilon\to 0} u_\epsilon(0) = a$.*

The constant τ here is a positive number, often called the surface tension. It is given in (2.5). The proof of Theorem 1.1 uses a type of Lyapunov–Schmidt reduction procedure tailored for singular perturbation problems. It consists of two steps. First we reduce I_ϵ to a functional Q_ϵ that is defined on a finite dimensional set. This set is really the coordinates of interfaces. In this step we construct a family of approximate solutions with K interfaces whose coordinates serve as parameters. The family is a finite dimensional submanifold of the admissible set of I_ϵ. Near each approximate solution we find a function that "solves" (1.5) in a direction that is more or less perpendicular to the submanifold. These functions are again parameterized by their interfaces and they form an improved finite dimensional submanifold. The restriction of I_ϵ on this new submanifold is Q_ϵ, which is viewed as a function of the interfaces. As a consequence of this construction, we show that a critical point of Q_ϵ is a solution of (1.5).

In the second step of the proof, we look for critical points of Q_ϵ. We show that $\epsilon^{-1}Q_\epsilon$ converges in C^1_{loc} to a function J as $\epsilon \to 0$. When γ is sufficiently large, J has a minimum. Near this local minimum Q_ϵ also has a minimum for small ϵ. The topological degree of J and hence that of Q_ϵ are shown to be 0. We then conclude that when γ is large, there are at least two critical points of Q_ϵ.

The proof of Theorem 1.2 is similar. After the same reduction procedure, we show that when γ is large, the reduced problem J has one maximum in $\mathcal{A}_1^b$ if $a < b < 0$ and one maximum in $\mathcal{A}_1^a$ if $0 < a < b$.

Another purpose of this paper is to illustrate the power and limitation of the Γ-convergence theory [8, 16, 15, 13] applied to this problem. Consider the case covered in Theorem 1.1. The limit J of the reduced problem Q_ϵ can be easily identified in the Γ-convergence theory. If one can show that the minimum of J is isolated, then a local minimizer of I_ϵ exists according to the theory. For small values of K ($K = 1$ or $K = 2$), we are able to show that the minimum is indeed isolated. For general K this also appears to be true, but we do not have a proof.

Regarding the second critical point of J, which is found by the topological degree argument on Q_ϵ, one in general cannot derive from the Γ-convergence theory that there exists another critical point of I_ϵ corresponding to the second critical point of J. This is because the Γ-convergence theory addresses only isolated local minima of J but not other types of critical points of J.

Similarly, in the case covered by Theorem 1.2 the maximum of J in $\mathcal{A}_1^a$ and the maximum in $\mathcal{A}_1^b$ do not yield solutions of (1.1) by the Γ-convergence theory. This paper shows that the nonminimum critical points all correspond to solutions of (1.5). They are saddle points of (1.6).

In the second part of the paper we further demonstrate the effectiveness of the Lyapunov–Schmidt reduction method. We construct solutions with a number of interfaces that cluster near the boundary $r = 1$. Namely, we prove the following result.

THEOREM 1.3. *Suppose that $\gamma \geq 0$ and $a < b$. For any nonnegative integer k, there exists an $\epsilon_0 > 0$ such that for each $\epsilon \in (0, \epsilon_0]$, (1.1) has a solution u_ϵ, which has $2k + 1$ interfaces near the boundary $r = 1$. Moreover, $u_\epsilon \to b$ uniformly in $B_{1-\delta}(0)$ for any $\delta > 0$ small if $b > 0$, and $u_\epsilon \to a$ uniformly in $B_{1-\delta}(0)$ for any $\delta > 0$ small if $a < b \leq 0$.*

Here the $2k + 1$ layers are all close to the boundary $r = 1$. The distance between two successive interfaces is of order $\epsilon \log \frac{1}{\epsilon}$, and the distances between these interfaces and the boundary $r = 1$ are also of order $\epsilon \log \frac{1}{\epsilon}$. After reduction the problem becomes a finite dimensional maximization problem with respect to the interfaces. The solution constructed from this maximization procedure is again of saddle type.

In Theorem 1.3 the nonlocal term $\epsilon\gamma(1-\Delta)^{-1}u$ does not play a central role. The existence of solutions with layers clustering near the boundary is valid with ($\gamma > 0$) or without ($\gamma = 0$) the nonlocal term. See also the result in [2] for the unbalanced case. The existence of interior layer solutions (Theorems 1.1 and 1.2) is different. Our results show that to have solutions of multiple interior layers we must have sufficient nonlocality, i.e., γ must be large enough.

The organization of our paper serves these two purposes. In section 2 we recall how J is derived from the Γ-convergence theory. We show that the topological degree of J is always 0 in the case $a < 0 < b$. For large γ we show that J has a minimum and consequently there is another critical point of J. The main work starts in section 3, where we reduce the study of I_ϵ to that of the finite dimensional problem Q_ϵ. Then in section 4 we show that $\epsilon^{-1}Q_\epsilon$ converges to J in C^1_{loc} and prove Theorems 1.1 and 1.2.

In section 5 another reduction is used to prove Theorem 1.3. We again derive a reduced functional of the interfaces. This time the interfaces are close to each other and to the boundary (all the distances are of order $\epsilon \log \frac{1}{\epsilon}$). As the interfaces vary in this range the functional varies by a quantity that is much smaller than ϵ. This compares differently from the situation discussed in Theorems 1.1 and 1.2. We show that the reduced problem is maximized at an interior point.

The conditions on a, b, and γ in the three theorems are used when we solve the reduced problems. In the case of Theorem 1.1 J has many critical points, and in the case of Theorem 1.2 J has only one critical point. In the case of Theorem 1.3 we will solve the reduced problem by showing that it has an interior maximum point. To achieve this goal, the assumption on the sign of b is essential.

We use C to denote constants independent of ϵ. Their values may vary from line to line. The $L^p(\Omega)$ norm, $p \in [1, \infty]$, of a function is denoted by $\|\cdot\|_p$.

2. The Γ-limit. The limiting problem J is easily identified in the Γ-convergence theory. Other than the expression of J and its properties, given in Lemmas 2.3 and 2.4,

the details of the Γ-convergence and its consequences are not needed in this paper. Therefore, we omit the proofs of the statements in this section, with the exception of Lemmas 2.3 and 2.4. The interested reader may reconstruct them with the help of the references we provide.

The Γ-limit J of $\epsilon^{-1}I_\epsilon$ is defined on the admissible set

$$\{u \in BV(\Omega, \{a, b\}) : u = u(|x|)\}, \tag{2.1}$$

where $BV(\Omega, \{a, b\})$ is the set of functions of bounded variation which take values only in $\{a, b\}$. The set (2.1) consists of such functions that are radial.

A function in (2.1) has either a finite number of interfaces or infinitely many interfaces. If it has finite, say, K, interfaces, there exist $r_1, r_2, \dots, r_K$, with $0 < r_1 < r_2 < \cdots < r_K < 1$, that divide the interval $(0, 1)$ into $(0, r_1)$, (r_1, r_2), $\dots$,(r_{K-1}, r_K), $(r_K, 1)$, and

$$u(r) = a \text{ on } (0, r_1), \ = b \text{ on } (r_1, r_2), \ = a \text{ on } (r_2, r_3), \dots \tag{2.2}$$

or

$$u(r) = b \text{ on } (0, r_1), \ = a \text{ on } (r_1, r_2), \ = b \text{ on } (r_2, r_3), \dots . \tag{2.3}$$

In the case of (2.2) we say that $u \in \mathcal{A}_K^a$ and in the case of (2.3) we say that $u \in \mathcal{A}_K^b$. On $\mathcal{A}_K^a$ and $\mathcal{A}_K^b$ the Γ-limit J is given by

$$J(u) = \omega_{n-1}\tau \sum_{j=1}^{K} r_j^{n-1} + \frac{\omega_{n-1}\gamma}{2} \int_0^1 |(1-\Delta)^{-1/2}u|^2 \, r^{n-1} dr. \tag{2.4}$$

Here we denote the area of the $n-1$ dimensional unit sphere by ω_{n-1}. The constant τ in (2.4) is given by

$$\tau = \int_a^b \sqrt{2W(q)}\, dq. \tag{2.5}$$

A function in (2.1) may also have infinite interfaces. Then the interfaces must accumulate at the origin. Otherwise, if there were a cluster point not at the origin, the total length of the interfaces would be infinite and u could not be in (2.1). Hence there exists a *decreasing* sequence $r_1, r_2, \dots$, such that $1 > r_1 > r_2 > \dots$ and $\lim_{j\to\infty} r_j = 0$, and either

$$u = \begin{cases} a & \text{on} \quad (r_1, 1), \\ b & \text{on} \quad (r_2, r_1), \\ a & \text{on} \quad (r_3, r_2) \\ \dots \end{cases} \tag{2.6}$$

or

$$u = \begin{cases} b & \text{on} \quad (r_1, 1), \\ a & \text{on} \quad (r_2, r_1), \\ b & \text{on} \quad (r_3, r_2) \\ \dots . \end{cases} \tag{2.7}$$

In this case J is defined by (2.4) with $\sum_{j=1}^{K}$ replaced by $\sum_{j=1}^{\infty}$. Because u is assumed to have bounded variation, this infinite sum converges.

Now we have J which is defined in (2.1). But I_ϵ is defined in a different set (1.10). We trivially extend both to

$$\{u \in L^2(\Omega) : u = u(|x|)\}, \tag{2.8}$$

the radial L^2-functions, by setting $I_\epsilon(u) = \infty$ if u is in (2.8) but not in (1.10) and similarly $J(u) = \infty$ if u is in (2.8) but not in (2.1). In (2.8) distance is measured by the L^2 norm $\|\cdot\|_2$.

The Γ-convergence of $\epsilon^{-1}I_\epsilon$ to J is characterized by the two properties of the following lemma.

LEMMA 2.1. *As $\epsilon \to 0$, $\epsilon^{-1}I_\epsilon$ Γ-converges to J in the following sense:*

1. *For every family of functions ϕ_ϵ in (2.8) with $\lim_{\epsilon\to 0}\|\phi_\epsilon - \phi\|_2 = 0$, $\liminf_{\epsilon\to 0}\epsilon^{-1}I_\epsilon(\phi_\epsilon) \geq J(\phi)$.*
2. *For every ϕ in (2.8), there is a family of functions ϕ_ϵ in (2.8) such that $\lim_{\epsilon\to 0}\|\phi_\epsilon - \phi\|_2 = 0$ and $\limsup_{\epsilon\to 0}\epsilon^{-1}I_\epsilon(\phi_\epsilon) \leq J(\phi)$.*

One important consequence of the Γ-convergence of $\epsilon^{-1}I_\epsilon$ is the following existence result.

LEMMA 2.2. *If $u_0 \in \mathcal{A}_K^a$ (or $\mathcal{A}_K^b$, respectively) is an isolated local minimum of J in $\mathcal{A}_K^a$ (or $\mathcal{A}_K^b$, respectively), for sufficiently small ϵ, there exists a local minimizer u_ϵ of I_ϵ, and $\lim_{\epsilon\to 0}\|u_\epsilon - u_0\|_2 = 0$.*

These two lemmas may be proved by mimicking the argument in Ren and Wei [22]. Lemma 2.2 suggests that we look for minima of J in $\mathcal{A}_K^a$ and $\mathcal{A}_K^b$. We do this in the rest of this section. Moreover, we will also find critical points of J that are not local minima.

Let $G = G(r, s)$ be the Green's function

$$-G_{rr} - \frac{n-1}{r}G_r + G = \delta(r - s). \tag{2.9}$$

Note that $G(r, s)$ is not symmetric in r and s, but $r^{n-1}G(r, s)$ is. We define

$$v(r) = \int_0^1 G(r, s)u(s)\,ds \tag{2.10}$$

to be the solution of

$$-v_{rr} - \frac{n-1}{r}v_r + v = u. \tag{2.11}$$

When it is in $\mathcal{A}_K^a$ or $\mathcal{A}_K^b$, u is determined by its jump points $r_1, r_2, \ldots, r_K$, which we term interfaces. Collectively we set $\mathbf{r} = (r_1, r_2, \ldots, r_K)$. Because u depends on $\mathbf{r}$, we often write $u = u(r; \mathbf{r})$ and correspondingly $v = v(r; \mathbf{r})$.

The nonlocal part of J may be rewritten as

$$\int_0^1 |(1-\Delta)^{-1/2}u(\cdot;\mathbf{r})|^2\, r^{n-1}dr = \int_0^1 v(r;\mathbf{r})u(r;\mathbf{r})\, r^{n-1}dr. \tag{2.12}$$

We view J as a function of $\mathbf{r}$: $J = J(\mathbf{r})$. Now we compute the derivative of J. Note that

$$\begin{aligned}\frac{\partial}{\partial r_1}\int_0^1 v(r)u(r)\, r^{n-1}dr &= \frac{\partial}{\partial r_1}\left[\int_0^{r_1} v(r)a\, r^{n-1}dr + \int_{r_1}^{r_2} v(r)b\, r^{n-1}dr + \ldots\right] \\ &= (a-b)v(r_1)r_1^{n-1} + \int_0^1 \frac{\partial v(r;\mathbf{r})}{\partial r_1}u(r;\mathbf{r})\, r^{n-1}dr. \end{aligned} \tag{2.13}$$

Because

$$\frac{\partial v(r;\mathbf{r})}{\partial r_1} = \frac{\partial}{\partial r_1}\left[\int_0^{r_1} G(r,s)a\,ds + \int_{r_1}^{r_2} G(r,s)b\,ds + \dots\right]$$

$$= (a-b)G(r,r_1), \tag{2.14}$$

we find

$$\begin{aligned}
&\frac{\partial}{\partial r_1}\int_0^1 v(r)u(r)\,r^{n-1}dr \\
&\quad = (a-b)v(r_1)r_1^{n-1} + (a-b)\int_0^1 G(r,r_1)u(r;\mathbf{r})\,r^{n-1}dr \\
&\quad = (a-b)v(r_1)r_1^{n-1} + (a-b)\int_0^1 G(r_1,r)u(r;\mathbf{r})\,r_1^{n-1}dr \\
&\quad = (a-b)v(r_1)r_1^{n-1} + (a-b)v(r_1)r_1^{n-1} \;=\; 2(a-b)v(r_1)r_1^{n-1}. \qquad (2.15)
\end{aligned}$$

We have used the symmetry of $r^{n-1}G(r,r_1)$. For a general r_j we have

$$\frac{\partial}{\partial r_j}\int_0^1 v(r)u(r)\,r^{n-1}dr = 2(-1)^j(b-a)v(r_j;\mathbf{r})r_j^{n-1}. \tag{2.16}$$

Therefore

$$\frac{\partial J(\mathbf{r})}{\partial r_j} = \omega_{n-1}[(n-1)\tau r_j^{n-2} + \gamma(b-a)(-1)^j v(r_j;\mathbf{r})r_j^{n-1}], \quad \mathbf{r}\in\mathcal{A}_K^a. \tag{2.17}$$

The gradient of J in $\mathcal{A}_K^b$ is a bit different:

$$\frac{\partial J(\mathbf{r})}{\partial r_j} = \omega_{n-1}[(n-1)\tau r_j^{n-2} + \gamma(a-b)(-1)^j v(r_j;\mathbf{r})r_j^{n-1}], \quad \mathbf{r}\in\mathcal{A}_K^b. \tag{2.18}$$

The existence of critical points of J depends on a, b, and γ. We consider the following three cases:

- *Case* I. $a < b < 0$.
- *Case* II. $0 < a < b$.
- *Case* III. $a < 0 < b$.

The first two cases are relatively simple. We have the following result.

LEMMA 2.3.

1. *If $a < b < 0$ and $\gamma > \gamma_I$ where*

$$\gamma_I = \frac{(n-1)\tau}{(a-b)b}, \tag{2.19}$$

there is a local maximum in $\mathcal{A}_1^b$. There are no critical points in other classes.

2. *If $0 < a < b$ and $\gamma > \gamma_{II}$ where*

$$\gamma_{II} = \frac{(n-1)\tau}{(b-a)a}, \tag{2.20}$$

there is a local maximum in $\mathcal{A}_1^a$. There are no critical points in other classes.

Proof. We consider only case I, for case II may be similarly handled. Note that $u \le 0$ implies that $v < 0$. In each $\mathcal{A}_K^a$, $K \ge 1$, $\frac{\partial J}{\partial r_1} > 0$. There is no critical point in $\mathcal{A}_K^a$. In $\mathcal{A}_K^b$ for $K \ge 2$, $\frac{\partial J}{\partial r_2} > 0$. So there is no critical point in $\mathcal{A}_K^b$, $K \ge 2$. The only class left is $\mathcal{A}_1^b$. In this class

$$\frac{\partial J}{\partial r_1} = \omega_{n-1} r_1^{n-2}[(n-1)\tau - \gamma(a-b)v(r_1; r_1)r_1].$$

The quantity inside the brackets is $(n-1)\tau$ when $r_1 = 0$ and $(n-1)\tau - \gamma(a-b)v(1) = (n-1)\tau - \gamma(a-b)b$ when $r_1 = 1$. Hence when $\gamma > \gamma_I$ there is a local maximum of J in $\mathcal{A}_1^b$, where γ_I is given in (2.19). Here since when $b = 0$, $\gamma_I = \infty$, the condition $\gamma > \gamma_I$ can be satisfied only when $b < 0$. □

Case III is the most interesting. We have the following lemma.

LEMMA 2.4. *Suppose $a < 0 < b$.*

1. *When γ is sufficiently large, J attains a global minimum in $\mathcal{A}_K^a$ (or $\mathcal{A}_K^b$)—not on the boundary of $\mathcal{A}_K^a$ (or $\mathcal{A}_K^b$).*
2. *Given $\gamma' > 0$ and any compact subset $\mathcal{K}'$ of $\mathcal{A}_K^a$ (or $\mathcal{A}_K^b$) one can find a compact subset $\mathcal{K}$, such that $\mathcal{K}' \subset \mathcal{K}$ and for all $\gamma \in [0, \gamma']$ the topological degree of* grad J on $\mathcal{K}$ about $\vec{0}$ is zero.
3. When γ is large, there exist at least two critical points of J in each $\mathcal{A}_K^a$ (or $\mathcal{A}_K^b$).

Proof. To prove part 1, we note that as far as the minimum is concerned the condition that γ is large is equivalent to the condition that τ is small. Or J can be considered as a perturbation of the function

$$J_0(\mathbf{r}) = \frac{\omega_{n-1}\gamma}{2} \int_0^1 |((1-\Delta)^{-1/2} u(\cdot; \mathbf{r}))^2|\, r^{n-1} dr, \quad \mathbf{r} \in \mathcal{A}_K^a \text{ (or } \mathcal{A}_K^b). \tag{2.21}$$

We recall that $\mathcal{A}_K^a$ (and, similarly, $\mathcal{A}_K^b$) is identified with

$$\{\mathbf{r} = (r_1, \dots, r_K) : 0 < r_1 < \cdots < r_K\}$$

so that the boundary of $\mathcal{A}_K^a$ is not included in $\mathcal{A}_K^a$. We study J_0 on the boundary of $\mathcal{A}_K^a$ (the case $\mathcal{A}_K^b$ is left to the reader), which consists of three pieces: (1) $r_1 = 0$, (2) $r_K = 1$, and (3) $r_j = r_{j+1}$ for some $j = 1, 2, \dots, K-1$.

In $\mathcal{A}_K^a$,

$$\frac{\partial J_0}{\partial r_j} = \omega_{n-1}\gamma(b-a)(-1)^j v(r_j; \mathbf{r}) r_j^{n-1}. \tag{2.22}$$

If the minimum of J_0 is achieved on $r_1 = 0$, say, at $\mathbf{r} = (0, r_2, r_3, \dots)$, then

$$v(r_2; \mathbf{r}) = v(r_3; \mathbf{r}) = \cdots = v(r_K; \mathbf{r}) = 0.$$

However, $v(r_2; \mathbf{r}) = 0$ and $u(r; \mathbf{r}) = b > 0$ for $r \in (0, r_2)$ imply that $v(r; \mathbf{r}) > 0$ for $r \in [0, r_2)$ by the maximum principal. Then at this $\mathbf{r}$

$$\frac{\partial J_0(\mathbf{r})}{\partial r_1} = -\omega_{n-1}\gamma(b-a)v(0; \mathbf{r}) < 0. \tag{2.23}$$

This means that the gradient of J_0 points outward at this $\mathbf{r}$. Then $\mathbf{r}$ cannot be a minimum point.

If the minimum of J is achieved at $\mathbf{r} = (r_1, r_2, \dots, r_{K-1}, 1)$ on the boundary piece $r_N = 1$, then

$$v(r_1; \mathbf{r}) = v(r_2; \mathbf{r}) = \cdots = v(r_{K-1}; \mathbf{r}) = 0.$$

Since for $r \in (r_{K-1}, 1)$,

$$u(r; \mathbf{r}) = \begin{cases} a < 0 & \text{if } K \text{ is odd,} \\ b > 0 & \text{if } K \text{ is even,} \end{cases}$$

$v(r_{K-1}; \mathbf{r}) = 0$ implies that $v(r)$ is negative if K is odd and positive if K is even on $(r_{K-1}, 1]$. Then

$$\frac{\partial J_0}{\partial r_K} = \gamma(b-a)(-1)^K v(1; \mathbf{r}) > 0. \tag{2.24}$$

Hence the gradient of J points outward at this $\mathbf{r}$ and it cannot be the minimum.

If the minimum of J is achieved at a boundary point $\mathbf{r}$ on $r_j = r_{j+1}$, we have two possibilities. First we may have $\mathbf{r} = (r_1, r_2, \dots, r_{j-1}, r_j, r_{j+1}, r_{j+2}, \dots, r_K)$ with $r_1 < r_2 < \cdots < r_{j-1} < r_j = r_{j+1} < r_{j+2} < \cdots < r_K$. This means two interfaces coincide but other interfaces stay separate. Then

$$v(r_1; \mathbf{r}) = \cdots = v(r_{j-1}; \mathbf{r}) = v(r_{j+2}; \mathbf{r}) = \cdots = v(r_K; \mathbf{r}) = 0.$$

When $r \in (r_{j-1}, r_{j+2})$,

$$u(r; \mathbf{r}) = \begin{cases} a < 0 & \text{if } j \text{ is odd,} \\ b > 0 & \text{if } j \text{ is even.} \end{cases}$$

Then, since $v(r_{j-1}; \mathbf{r}) = v(r_{j+2}; \mathbf{r}) = 0$ for $r \in (r_{j-1}, r_{j+2})$, $v(r; \mathbf{r})$ is negative if j is odd and positive if j is even by the maximal principal. Note that at the minimum $\mathbf{r}$ the outward normal direction is $\nu = (0, 0, \dots 0, 1, -1, 0, \dots, 0)$, where 1 is the jth entry and -1 the $(j+1)$th entry. The directional derivative along ν is

$$\begin{aligned} \frac{\partial J}{\partial \nu} &= \gamma(b-a)[(-1)^j v(r_j; \mathbf{r}) r_j^{n-1} - (-1)^{j+1} v(r_{j+1}; \mathbf{r}) r_{j+1}^{n-1}] \\ &= 2\gamma(b-a)(-1)^j v(r_j; \mathbf{r}) r_j^{n-1} > 0. \end{aligned}$$

Hence $\mathbf{r}$ cannot be the minimum.

In this case there is also the possibility that more than two interfaces collapse at one point, where the minimum is attained—for example, at $\mathbf{r}$ where $r_1 < r_2 < \cdots < r_{j-2} < r_{j-1} = r_j = r_{j+1} < r_{j+2} < \cdots < r_K$. However, this point can be viewed as a point on the boundary of $\mathcal{A}^a_{K-1}$. We can make an induction assumption that in every $\mathcal{A}^a_N$ or $\mathcal{A}^b_N$ with $N \le K-1$, the minimum of J_0 is not achieved on the boundary. Therefore this possibility needs no consideration.

Therefore the minimum of J_0 is achieved on a compact subset of $\mathcal{A}^a_K$. Hence for large γ, the minimum of J is also achieved inside $\mathcal{A}^a_K$.

To prove part 2, we treat γ in J as a parameter for the homotopy argument. We consider the topological degree of $\operatorname{grad} J$. We are given a compact subset $\mathcal{K}'$ of $\mathcal{A}^a_K$ and γ in $[0, \gamma']$.

First we show that grad J is not $\vec{0}$ on the boundary of some compact $\mathcal{K} \supset \mathcal{K}'$. When $\gamma = 0$,

$$\frac{\partial J}{\partial r_j} = (n-1)\omega_{n-1}\tau r_j^{n-2}, \tag{2.25}$$

which is not 0 anywhere in $\mathcal{A}_K^a$. When $\gamma > 0$, we consider the three pieces of the boundary of $\mathcal{A}_K^a$ again.

Although on the piece $r_1 = 0$ of the boundary $\partial J/\partial r_1 = 0$ if $n > 2$, we move slightly away from $r_1 = 0$ and consider small and positive r_1. Then

$$\frac{\partial J}{\partial r_1} = (n-1)\omega_{n-1}\tau r_1^{n-2} + \gamma O(r_1^{n-1}) > 0. \tag{2.26}$$

Hence grad J is not $\vec{0}$ when r_1 is positive and small.

On the second piece $\mathbf{r}_K = 1$,

$$\frac{\partial J}{\partial r_{K-1}} = (n-1)\omega_{n-1}\tau r_{K-1}^{n-2} + \omega_{n-1}\gamma(b-a)(-1)^{K-1}v(r_{K-1};\mathbf{r})r_{K-1}^{n-1}, \tag{2.27}$$

$$\frac{\partial J}{\partial r_K} = (n-1)\omega_{n-1}\tau + \omega_{n-1}\gamma(b-a)(-1)^{K}v(1;\mathbf{r}). \tag{2.28}$$

For $r \in (r_{K-1}, 1)$, $u(r;\mathbf{r}) = a < 0$ if K is odd and $u(r;\mathbf{r}) = b > 0$ if K is even. If $\frac{\partial J}{\partial r_{K-1}} = 0$, then $v(r_{K-1};\mathbf{r})$ is negative if K is odd and positive if K is even. Then for $r \in [r_{K-1}, 1]$, $v(r;\mathbf{r})$ is negative if K is odd and positive if K is even. In particular, $v(1;\mathbf{r})$ is negative if K is odd and positive if K is even. Then $\frac{\partial J}{\partial r_K}$ is always positive. Hence grad J is not $\vec{0}$ on the second piece of the boundary.

On the third piece of the boundary $r_j = r_{j+1}$,

$$\frac{\partial J}{\partial r_j} = (n-1)\omega_{n-1}\tau r_j^{n-2} + \omega_{n-1}\gamma(b-a)(-1)^{j}v(r_j;\mathbf{r})r_j^{n-1}, \tag{2.29}$$

$$\frac{\partial J}{\partial r_{j+1}} = (n-1)\omega_{n-1}\tau r_j^{n-2} + \omega_{n-1}\gamma(b-a)(-1)^{j+1}v(r_j;\mathbf{r})r_j^{n-1}. \tag{2.30}$$

These two partial derivatives cannot simultaneously be 0. Hence grad J is not $\vec{0}$ on the third piece of the boundary.

Now we can find a compact subset $\mathcal{K} \supset \mathcal{K}'$ of $\mathcal{A}_K^a$ so that for all $\gamma \in [0, \gamma']$, grad J is not $\vec{0}$ on the boundary of $\mathcal{K}$. Consequently we can define the topological degree of grad J in $\mathcal{K}$ about $\vec{0}$:

$$\mathrm{Deg}\,(\mathrm{grad}\, J, \mathcal{K}, \vec{0}). \tag{2.31}$$

Note that in part 2 of the lemma, γ is allowed to be 0. This is important, because when $\gamma = 0$, grad $J \neq \vec{0}$ in $\mathcal{A}_K^a$. Hence $\mathrm{Deg}\,(\mathrm{grad}\, J, \mathcal{K}, \vec{0}) = 0$ when $\gamma = 0$. By the invariance of the degree under continuous deformation, $\mathrm{Deg}\,(\mathrm{grad}\, J, \mathcal{K}, \vec{0}) = 0$ for all $\gamma \in [0, \gamma']$. This proves part 2 of the lemma.

The third part of the lemma follows from parts 1 and 2. For large γ, there is a minimum, say, $\mathbf{r}_*$, in $\mathcal{A}_K^a$. This gives one critical point of J. If this is the only critical point of J in $\mathcal{A}_K^a$, we can find an open ball $\mathcal{B}_\eta(\mathbf{r}_*)$ of radius η centered at $\mathbf{r}_*$ whose

closure is a subset of $\mathcal{A}_K^a$. Let $\mathcal{K}$ be a compact subset of $\mathcal{A}_K^a$ given in part 2 of the lemma, and it is large enough to contain $\mathcal{B}_\eta(\mathbf{r}_*)$ as a subset. Then

$$\text{Deg}\,(\text{grad}\,J, \mathcal{K}, \vec{0}) = \text{Deg}\,(\text{grad}\,J, \overline{\mathcal{B}}_\eta(\mathbf{r}_*), \vec{0}) + \text{Deg}\,(\text{grad}\,J, \mathcal{K}\backslash\mathcal{B}_\eta(\mathbf{r}_*), \vec{0}). \tag{2.32}$$

We know that $\text{Deg}\,(\text{grad}\,J, \mathcal{K}, \vec{0}) = 0$ from part 2. Also $\text{Deg}\,(\text{grad}\,J, \mathcal{B}_\eta(\mathbf{r}_*), \vec{0}) = 1$ because the minimum $\mathbf{r}_*$ is the only critical point in $\mathcal{B}_\eta(\mathbf{r}_*)$. Therefore $\text{Deg}\,(\text{grad}\,J, \mathcal{K}\backslash\mathcal{B}_\eta(\mathbf{r}_*), \vec{0}) = -1 \neq 0$. There is another critical point in $\mathcal{K}\backslash\mathcal{B}_\eta(\mathbf{r}_*)$. □

The reader is probably tempted to combine Lemmas 2.2 and 2.4. Since there is a minimum of J in $\mathcal{A}_K^a$ when γ is large, one would like to show that this minimum is isolated and then following Lemma 2.2 conclude that I_ϵ has a local minimizer near the minimum of J. When $K = 1$ or $K = 2$, it is indeed easy to show that the local minimum of J is isolated. However, for general K, we do not have a proof.

Moreover, in part 3 of Lemma 2.4 we have also found another critical point of J for large γ. This critical point is in general not a local minimum of J. Lemma 2.2 is hence not applicable.

Similarly, the local maxima found in Lemma 2.3 are not of much use in the Γ-convergence theory.

To make use of all the critical points of J found in Lemmas 2.3 and 2.4, we now abandon the Γ-convergence theory and proceed differently. Our new reduction approach may roughly be regarded as a convergence theory at the C^2 level, while the Γ-convergence theory is at the C^0 level. Using this argument we will be able to prove that in case III there are at least two critical points of I_ϵ with K interfaces when γ is sufficiently large (see Theorem 1.1). Similarly, in cases I and II there is a critical point of I_ϵ with one interface if γ is large (see Theorem 1.2).

3. Lyapunov–Schmidt reduction procedure. The Lyapunov–Schmidt reduction procedure involves the first and second derivatives of I_ϵ. For this reason we vaguely regard it as a reduction theory at the C^2 level.

We construct a manifold $\mathcal{M}$ of approximate solutions parameterized by $\mathbf{r} = (r_1, r_2, \dots, r_K)$. First define

$$s(r; \mathbf{r}) = a \text{ in } (0, r_1),\ b \text{ in } (r_1, r_2),\ a \text{ in } (r_2, r_3),\ \dots, \tag{3.1}$$

which gives a profile away from the interfaces. Clearly $s(\cdot; \mathbf{r}) \in \mathcal{A}_K^a$. We also identify the domain of $\mathbf{r}$ with $\mathcal{A}_K^a$. From now to the end of the paper we construct the two solutions in Theorem 1.1 that satisfy $\lim_{\epsilon\to 0} u_\epsilon^a(0) = a$ and the solution in part 2 of Theorem 1.2. Similar arguments can give the other solutions, starting with an $s(\cdot; \mathbf{r}) \in \mathcal{A}_K^b$. We leave the details to the reader.

The interface profile is the solution $H(t)$ of the differential equation

$$-H_{tt} + f(H) = 0,\ H(-\infty) = a,\ H(\infty) = b,\ H(0) = \frac{a+b}{2}. \tag{3.2}$$

$H(t)$ approaches a (or b, respectively) exponentially fast as t tends to $-\infty$ (or ∞, respectively) in the sense that there exist positive C_1, C_2 so that

$$0 < H(t) - a < C_1 e^{C_2 t} \text{ if } t < 0, \text{ and } 0 < b - H(t) < C_1 e^{-C_2 t} \text{ if } t > 0. \tag{3.3}$$

Near r_j we use $H((r - r_j)/\epsilon)$ if j is odd, or $H(-(r - r_j)/\epsilon)$ if j is even.

The outer approximation $s(\cdot;\mathbf{r})$ and the inner approximation H must be connected by a smooth cut-off function χ to make

$$w(r;\mathbf{r}) = \sum_{j=1}^{K} \chi(r-r_j) H\left((-1)^{j+1}\frac{r-r_j}{\epsilon}\right) + \left(1-\sum_{j=1}^{K}\chi(r-r_j)\right) s(r;\mathbf{r}), \tag{3.4}$$

where χ is defined to be

$$\chi(r) = \begin{cases} 1 & \text{in} \quad (-\epsilon^\alpha, \epsilon^\alpha), \\ 0 & \text{in} \quad \mathbf{R}\backslash(-2\epsilon^\alpha, 2\epsilon^\alpha). \end{cases} \tag{3.5}$$

The exponent α in (3.5) satisfies

$$0 < \alpha < 1. \tag{3.6}$$

χ satisfies

$$\chi = O(1),\ \chi' = O(\epsilon^{-\alpha}),\ \chi'' = O(\epsilon^{-2\alpha}). \tag{3.7}$$

The manifold $\mathcal{M}$ is

$$\mathcal{M} = \{w(\cdot;\mathbf{r}) :\ \mathbf{r} \in \mathcal{A}_K^a\}, \tag{3.8}$$

which is parameterized by $\mathbf{r}$ in $\mathcal{A}_K^a$.

We define two function spaces $\mathcal{X}$ and $\mathcal{Y}$,

$$\mathcal{X} = \{u \in W^{2,2}(\Omega) : u = u(|x|),\ u_r(1) = 0\};\quad \mathcal{Y} = \{q \in L^2(\Omega) : q = q(|x|)\}, \tag{3.9}$$

and a nonlinear operator $S_\epsilon : \mathcal{X} \to \mathcal{Y}$ by

$$S_\epsilon(u) = -\epsilon^2\Delta u + f(u) + \epsilon\gamma(1-\Delta)^{-1}u. \tag{3.10}$$

Equation (1.5) is $S_\epsilon(u) = 0$.

LEMMA 3.1. *$S_\epsilon(w) = O(\epsilon)$ locally uniformly in $\mathbf{r}$ and γ. More precisely, for each compact subset $\mathcal{K}$ of $\mathcal{A}_K^a$ and $[\gamma_1,\gamma_2]$, $0 < \gamma_1 < \gamma_2$, there exist $C > 0$ and $\epsilon_0 > 0$ such that for all $\mathbf{r} \in \mathcal{K}$, $\gamma \in [\gamma_1,\gamma_2]$ and $\epsilon < \epsilon_0$, $\|S_\epsilon(w(\cdot;\mathbf{r}))\|_\infty \le C\epsilon$.*

Proof. Given $\mathcal{K}$ and $[\gamma_1,\gamma_2]$, we let $\mathbf{r} \in \mathcal{K}$ and $\gamma \in [\gamma_1,\gamma_2]$. Then

$$\begin{aligned} S_\epsilon(w) &= -\epsilon^2\left(w_{rr} + \frac{n-1}{r}w_r\right) + f(w) + \epsilon\gamma(1-\Delta)^{-1}w \\ &= (-\epsilon^2 w_{rr} + f(w)) - \epsilon^2\frac{n-1}{r}w_r + \epsilon\gamma(1-\Delta)^{-1}w \\ &= O(e^{-C/\epsilon}) + O(\epsilon) + O(\epsilon) \ =\ O(\epsilon). \end{aligned} \tag{3.11}$$

The lemma follows. ∎

For each $j = 1, 2, \ldots, K$, let us define

$$h_j(r) = H'\left(\frac{r-r_j}{\epsilon}\right)\kappa\left(\frac{r-r_j}{\sqrt{\epsilon}}\right) = H'\left(\frac{r-r_j}{\epsilon}\right) + O(e^{-C/\sqrt{\epsilon}}), \tag{3.12}$$

where κ is a smooth, even, cut-off function

$$\kappa(s) = \begin{cases} 1 & \text{if } |s| \le 1, \\ 0 & \text{if } |s| \ge 2. \end{cases} \tag{3.13}$$

Here $O(e^{-C/\sqrt{\epsilon}})$ is an exponentially small quantity with respect to ϵ because of the exponentially fast decay rate of H': $H'(t) \le C_1 e^{-C_2|t|}$. Therefore $h_j'(0) = h_j'(1) = 0$, $\|h_j' - \epsilon^{-1} H''(\frac{\cdot - r_j}{\epsilon})\|_\infty = O(\epsilon^{-C/\sqrt{\epsilon}})$, and $\|h_j'' - \epsilon^{-2} H'''(\frac{\cdot - r_j}{\epsilon})\|_\infty = O(\epsilon^{-C/\sqrt{\epsilon}})$. Note that h_j depends on $\mathbf{r}$ so we sometimes write it as $h_j(r; \mathbf{r})$.

At each $w(\cdot; \mathbf{r})$ of the manifold we define the space

$$\mathcal{F}_{\mathbf{r}} = \{\phi \in \mathcal{X} : \ \phi \perp h_j, \ j = 1, 2, \dots, K\}, \tag{3.14}$$

where $\perp$ is defined from the inner product

$$\langle A, B \rangle = \int_0^1 A(r) B(r)\, r^{n-1} dr. \tag{3.15}$$

Then $w_{\mathbf{r}} + \mathcal{F}_{\mathbf{r}}$ is a subset of $\mathcal{X}$, which we call the $\mathbf{r}$-fiber of $\mathcal{M}$ in $\mathcal{X}$. Define $\mathcal{E}_{\mathbf{r}}$ to be the subspace

$$\mathcal{E}_{\mathbf{r}} = \{q \in \mathcal{Y} : \ q \perp h_j, \ j = 1, 2, \dots, K\} \tag{3.16}$$

of $\mathcal{Y}$. Let the projection from $\mathcal{Y}$ to $\mathcal{E}_{\mathbf{r}}$ be $\pi_{\mathbf{r}} : \mathcal{Y} \to \mathcal{E}_{\mathbf{r}}$, defined by

$$\pi_{\mathbf{r}}(q) = q - \sum_{j=1}^{K} \frac{\langle q, h_j \rangle}{\|h_j\|_2^2} h_j. \tag{3.17}$$

At each $w(\cdot; \mathbf{r})$ we look for a $\phi(\cdot; \mathbf{r}) \in \mathcal{F}_{\mathbf{r}}$ so that

$$\pi_{\mathbf{r}} \circ S_\epsilon(w(\cdot; \mathbf{r}) + \phi(\cdot; \mathbf{r})) = 0. \tag{3.18}$$

This means that we solve $S_\epsilon(u) = 0$ in the fiber direction. For each $\phi \in \mathcal{F}_{\mathbf{r}}$ we expand

$$S_\epsilon(w + \phi) = S_\epsilon(w) + L_{\mathbf{r}}(\phi) + R_{\mathbf{r}}(\phi), \tag{3.19}$$

where the linearized operator of S_ϵ at $w(\cdot; \mathbf{r})$ is denoted by $L_{\mathbf{r}}$: $\mathcal{X} \to \mathcal{Y}$, defined by

$$L_{\mathbf{r}}\phi := -\epsilon^2 \left(\phi_{rr} + \frac{n-1}{r} \phi_r \right) + f'(w(r; \mathbf{r}))\phi + \epsilon\gamma(1 - \Delta)^{-1}\phi, \tag{3.20}$$

and the remainder is

$$R_{\mathbf{r}}(\phi) = f(w + \phi) - f(w) - f'(w)\phi. \tag{3.21}$$

Then (3.18) is written as

$$\pi_{\mathbf{r}} \circ S_\epsilon(w) + \pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) + \pi_{\mathbf{r}} \circ R_{\mathbf{r}}(\phi) = 0. \tag{3.22}$$

Regarding the linear operator $\pi_{\mathbf{r}} \circ L_{\mathbf{r}}$,

$$\pi_{\mathbf{r}} \circ L_{\mathbf{r}} : \mathcal{F}_{\mathbf{r}} \to \mathcal{E}_{\mathbf{r}} \tag{3.23}$$

(note that it is defined on $\mathcal{F}_{\mathbf{r}}$—not on $\mathcal{X}$), we have the following lemma.

LEMMA 3.2.

1. *There exists $C_1 > 0$ independent of ϵ such that $\|\phi\|_\infty \le C_1 \|\pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi)\|_\infty$ for all $\phi \in \mathcal{F}_{\mathbf{r}}$. In particular, $\pi_{\mathbf{r}} \circ L_{\mathbf{r}}$ is one-to-one from $\mathcal{F}_{\mathbf{r}}$ to $\mathcal{E}_{\mathbf{r}}$.*
2. *$\pi_{\mathbf{r}} \circ L_{\mathbf{r}}$ is onto from $\mathcal{F}_{\mathbf{r}}$ to $\mathcal{E}_{\mathbf{r}}$.*

Proof. To prove part 1 we argue by contradiction. Suppose the conclusion is false. Then there exists $\psi_\epsilon \in \mathcal{F}_{\mathbf{r}}$ for each ϵ such that $\|\psi_\epsilon\|_\infty = 1$ and along a subsequence of $\epsilon \to 0$,

$$\|\pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\psi_\epsilon)\|_\infty \to 0. \tag{3.24}$$

To simplify notation, we write ψ instead of ψ_ϵ. We rewrite (3.24) as

$$-\epsilon^2 \left(\psi_{rr} + \frac{n-1}{r}\psi_r\right) + f'(w)\psi + \epsilon\gamma(1-\Delta)^{-1}\psi - \sum_{j=1}^{K} \beta_j h_j = o(1) \tag{3.25}$$

for some $\beta_j \in \mathbf{R}$. More specifically, β_j are given by

$$\beta_j = \frac{\langle L_{\mathbf{r}}(\psi), h_j \rangle}{\|h_j\|_2^2}. \tag{3.26}$$

We must estimate the size of β_j. To this end we multiply (3.25) by h_k and integrate. Then

$$\begin{aligned} &\int_0^1 \left[\left(-\epsilon^2 \left(\psi_{rr} + \frac{n-1}{r}\psi_r\right) + f'(w)\psi + \epsilon\gamma(1-\Delta)^{-1}\psi\right) h_k\right] r^{n-1} dr \\ &\quad + \sum_{j=1}^{K} \beta_j \langle h_j, h_k \rangle = o(\epsilon), \end{aligned} \tag{3.27}$$

Simple calculations simplify the second part on the left side, so

$$\begin{aligned} &\int_0^1 \left[\left(-\epsilon^2 \left(\psi_{rr} + \frac{n-1}{r}\psi_r\right) + f'(w)\psi + \epsilon\gamma(1-\Delta)^{-1}\psi\right) h_k\right] r^{n-1} dr \\ &\quad + \sum_{j=1}^{K} \beta_j (\epsilon\tau r_j^{n-1}\delta_{jk} + O(\epsilon^2)) = o(\epsilon), \end{aligned} \tag{3.28}$$

where $\delta_{jk} = 1$ if $j \ne k$ and 0 if $j = k$. Also we have used the fact that

$$\tau = \int_{\mathbf{R}} (H')^2 \, dt. \tag{3.29}$$

This τ is the same as the one given in (2.5). These two expressions give the same value because of (3.2), which H satisfies, and its first integral

$$-\frac{1}{2}(H'(t))^2 + W(H(t)) = 0. \tag{3.30}$$

The first part of the left side of (3.28) is estimated as follows:

(3.31)

$$\int_0^1 \left[\left(-\epsilon^2\left(\psi_{rr}+\frac{n-1}{r}\psi_r\right)+f'(w)\psi+\epsilon\gamma(1-\Delta)^{-1}\psi\right)h_k\right] r^{n-1}dr$$
$$=\int_0^1\left(-\epsilon^2\left(h_k''+\frac{n-1}{r}h_k'\right)\psi+f'(w)h_k\psi\right)r^{n-1}dr+\epsilon\gamma\int_0^1((1-\Delta)^{-1}\psi)h_k\,r^{n-1}dr$$
$$=\int_0^1 -\epsilon^2\frac{n-1}{r}h_k'\psi\,r^{n-1}dr+O(\epsilon^2)\;=\;O(\epsilon^2).$$

This simplifies (3.28) to

$$\sum_{j=1}^{2N}\beta_j(\epsilon\tau\delta_{jk}+O(\epsilon^2))=o(\epsilon). \tag{3.32}$$

Hence

$$\beta_j=o(1). \tag{3.33}$$

Let $y\in[0,1]$ such that, without loss of generality, $\psi(y)=\|\psi\|_\infty=1$. We claim that $y-r_j=O(\epsilon)$ for some j. Otherwise, at y,

$$\begin{aligned}L_{\mathbf{r}}(\psi)(y)&=-\epsilon^2\Delta\psi(y)+f'(w(y))\psi(y)+\epsilon\gamma((1-\Delta)^{-1}\psi)(y)\\&\geq 0+f'(w(y))+\epsilon\gamma((1-\Delta)^{-1}\psi)(y)\\&=f'(w(y))+O(\epsilon)\;=\;f'(a)+o(1).\end{aligned} \tag{3.34}$$

Combining (3.33) and (3.34), we obtain

$$\pi_{\mathbf{r}}\circ L_{\mathbf{r}}(\psi)(y)\geq f'(a)-\sum_{j=1}^{K}\beta_j h_j(y)+o(1)\geq f'(0)+o(1), \tag{3.35}$$

which contradicts (3.24).

We have thus proved that $y-\xi_j=O(\epsilon)$ for some j, along a subsequence of $\epsilon\to 0$. Define $\Psi(t)=\psi(r_j+\epsilon t)$. Then (3.25) and (3.33) imply

$$-\Psi''+f'(w_{\mathbf{r}}(r_j+\epsilon t))\Psi=o(1) \tag{3.36}$$

uniformly on any compact subset of $\mathbf{R}$. From here we may pass to the limit and find Ψ_∞ so that $\Psi\to\Psi_\infty$ in $C^2_{loc}(\mathbf{R})$. Moreover, $\Psi_\infty\neq 0$ since $\Psi((y-\xi_j)/\epsilon)=1$, and

$$-\Psi_\infty''+f'(H)\Psi_\infty=0. \tag{3.37}$$

The bounded solutions of this equation are scalar multiples of H'. Hence $\Psi_\infty=cH'$ for some $c\neq 0$.

On the other hand, since $\psi\in\mathcal{F}_{\mathbf{r}}$ means that $\psi\perp h_j$, we deduce that

$$\begin{aligned}0=\langle\psi,h_j\rangle&=\epsilon\int_{-r_j/\epsilon}^{(1-r_j)/\epsilon}\Psi(t)(H'(t)-O(e^{-C/\sqrt{\epsilon}}))(r_j+\epsilon t)^{n-1}\,dt\\&=\epsilon\left(cr_j^{n-1}\int_{\mathbf{R}}(H'(t))^2\,dt+o(1)\right),\end{aligned} \tag{3.38}$$

which is impossible, for $c \neq 0$. We have thus proved part 1 of the lemma.

To prove part 2 of the lemma we need to solve

$$\pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = p \tag{3.39}$$

in $\mathcal{F}_{\mathbf{r}}$ for any given $p \in \mathcal{E}_{\mathbf{r}}$. By applying $\pi_{\mathbf{r}} \circ (1-\Delta)^{-1}$ to both sides of (3.39) we consider the equation

$$\pi_{\mathbf{r}} \circ (1-\Delta)^{-1} \circ \pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = \pi_{\mathbf{r}} \circ (1-\Delta)^{-1} p. \tag{3.40}$$

The linear operator $\pi_{\mathbf{r}} \circ (1-\Delta)^{-1} \circ \pi_{\mathbf{r}} \circ L_{\mathbf{r}}$ on the left side maps from $\mathcal{F}_{\mathbf{r}}$ to itself. For this operator $\mathcal{F}_{\mathbf{r}}$ is viewed as a Banach space whose norm is inherited from the $W^{2,2}(\Omega)$ norm. The operator has the form

$$\epsilon^2(\text{identity operator}) + \text{compact operator}. \tag{3.41}$$

According to the Fredholm alternative, (3.40) is solvable if

$$\pi_{\mathbf{r}} \circ (1-\Delta)^{-1} \circ \pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = 0 \tag{3.42}$$

has only the trivial solution. To see this we write (3.42) as

$$(1-\Delta)^{-1} \circ \pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = \sum_{j=1}^{K} \alpha_j h_j \tag{3.43}$$

for some $\alpha_j \in \mathbf{R}$. Apply $1-\Delta$ to the last equation to find

$$\pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = \sum_{j=1}^{K} \alpha_j (-\Delta h_j + h_j). \tag{3.44}$$

We multiply it by h_k and integrate to deduce

$$\begin{aligned} 0 &= \sum_{j=1}^{K} \alpha_j \int_{\Omega} (\nabla h_j \cdot \nabla h_k + h_j h_k)\, dx \qquad (3.45) \\ &= \alpha_k \int_{\Omega} (|\nabla h_k|^2 + h_k^2)\, dx, \ \ k = 1, 2, \dots, K, \end{aligned}$$

which implies that $\alpha_j = 0$, $j = 1, 2, \dots, K$. Then (3.44) becomes

$$\pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = 0. \tag{3.46}$$

The first part of the lemma implies that $\phi = 0$.

Hence (3.40) is solvable; i.e., for any $p \in \mathcal{E}_{\mathbf{r}}$ there exist $\phi \in \mathcal{F}_{\mathbf{r}}$ and $\beta_j \in \mathbf{R}$ such that

$$(1-\Delta)^{-1} \circ \pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = (1-\Delta)^{-1} p + \sum_{j=1}^{K} \beta_j h_j. \tag{3.47}$$

Apply $1-\Delta$ to the last equation to deduce

$$\pi_{\mathbf{r}} \circ L_{\mathbf{r}}(\phi) = p + \sum_{j=1}^{K} \beta_j (-\Delta h_j + h_j). \tag{3.48}$$

We again multiply by h_k and integrate to obtain

$$0 = \sum_{j=1}^{K} \beta_j \int_\Omega \nabla h_j \cdot \nabla h_k \, dx = \beta_k \epsilon \int_\Omega (|\nabla h_k|^2 + h_k^2) \, dx, \ k = 1, 2, \dots, K, \tag{3.49}$$

which implies that $\beta_j = 0$ for all $j = 1, 2, \dots, K$. Then (3.48) becomes (3.39). ∎

We are now ready to solve (3.18).

LEMMA 3.3. *There exists $\phi(\cdot; \mathbf{r}) \in \mathcal{F}_\mathbf{r}$ with $\|\phi(\cdot; \mathbf{r})\|_\infty = O(\epsilon)$ so that $\pi_\mathbf{r} \circ S_\epsilon(w(\cdot; \mathbf{r}) + \phi(\cdot; \mathbf{r})) = 0$.*

Proof. We write (3.22) in a fixed point form:

$$\phi = (\pi_\mathbf{r} \circ L_\mathbf{r})^{-1}(-\pi_\mathbf{r} \circ S_\epsilon(w) - \pi_\mathbf{r} \circ R_\mathbf{r}(\phi)). \tag{3.50}$$

We define the operator $T_\mathbf{r}$ from $\mathcal{D}(T_\mathbf{r})$ to itself,

$$T_\mathbf{r}(\phi) = (\pi_\mathbf{r} \circ L_\mathbf{r})^{-1}(-\pi_\mathbf{r} \circ S_\epsilon(w) - \pi_\mathbf{r} \circ R_\mathbf{r}(\phi)), \tag{3.51}$$

where the domain $\mathcal{D}(T_\mathbf{r})$ of $T_\mathbf{r}$ is

$$\mathcal{D}(T_\mathbf{r}) = \{\phi \in L^\infty(0,1) : \ \phi \perp h_j, \ j = 1, 2, \dots, K\}. \tag{3.52}$$

Let $\mathcal{B}_\mathbf{r}$ be a closed ball in $\mathcal{D}(T_\mathbf{r})$ defined by

$$\mathcal{B}_\mathbf{r} = \{\phi \in \mathcal{D}(T_\mathbf{r}) : \|\phi\|_\infty \le C_2 \epsilon\}, \tag{3.53}$$

where C_2 is a constant independent of ϵ to be determined soon. For every $\phi \in \mathcal{B}_\mathbf{r}$, by Lemma 3.1

$$\begin{aligned} \|T_\mathbf{r}(\phi)\|_\infty &\le C_1 \|\pi_\mathbf{r} \circ S_\epsilon(w)\|_\infty + C_1 \|\pi_\mathbf{r} \circ R_\mathbf{r}(\phi)\|_\infty \\ &\le C_3 \epsilon + C_5 (1 + O(\|\phi\|_\infty)) \|\phi\|_\infty^2 \\ &\le C_3 \epsilon + C_6 C_2^2 (1 + C_2 \epsilon) \epsilon^2, \end{aligned} \tag{3.54}$$

where we have estimated $R_\mathbf{r}(\phi)$ as

$$\|R_\mathbf{r}(\phi)\|_\infty \le 2\|f(w_\mathbf{r} + \phi) - f(w_\mathbf{r}) - f'(w_\mathbf{r})\phi\|_\infty \le C_4 (1 + O(\|\phi\|_\infty)) \|\phi\|_\infty^2 \tag{3.55}$$

for some C_4 depending only on f. In (3.54) the constants C_3 and C_6 are again independent of ϵ. If we choose C_2 to be sufficiently large, then when ϵ is small enough (3.54) is bounded by $C_2 \epsilon$. Therefore by choosing such C_2 we see that $\mathcal{D}(T_\mathbf{r})$ maps $\mathcal{B}_\mathbf{r}$ to itself.

Next we prove that $T_\mathbf{r}$ is a contraction mapping in $\mathcal{D}(T_\mathbf{r})$. Take ϕ_1 and ϕ_2 in $\mathcal{D}(T_\mathbf{r})$. Then

$$\begin{aligned} \|T_\mathbf{r}(\phi_1) - T_\mathbf{r}(\phi_2)\|_\infty &\le C_1 \|\pi_\mathbf{r} \circ (R_\mathbf{r}(\phi_1) - R_\mathbf{r}(\phi_2))\|_\infty \ \le \ C_7 \|R_\mathbf{r}(\phi_1) - R_\mathbf{r}(\phi_2)\|_\infty \\ &\le C_8 \|f(w_\mathbf{r} + \phi_1) - f(w_\mathbf{r} + \phi_2) - f'(w_\mathbf{r})(\phi_1 - \phi_2)\|_\infty \\ &\le C_8 \|f'(w_\mathbf{r} + \phi_2 + \theta(\phi_1 - \phi_2))(\phi_1 - \phi_2) - f'(w_\mathbf{r})(\phi_1 - \phi_2)\|_\infty \\ &\le C_8 \|f'(w_\mathbf{r} + \phi_2 + \theta(\phi_1 - \phi_2)) - f'(w_\mathbf{r})\|_\infty \|\phi_1 - \phi_2\|_\infty \\ &\le O(\|\phi_1\|_\infty + \|\phi_2\|_\infty) \ \|\phi_1 - \phi_2\|_\infty \\ &\le C_9 \epsilon \|\phi_1 - \phi_2\|_\infty, \end{aligned} \tag{3.56}$$

which implies that $T_\mathbf{r}$ is a contraction mapping if ϵ is sufficiently small. In these estimates $\theta = \theta(x) \in (0,1)$ comes from the mean value theorem. ∎

4. C^1-convergence of the reduced problem. We now define

$$Q_\epsilon(\mathbf{r}) = I_\epsilon(w(\cdot;\mathbf{r}) + \phi(\cdot;\mathbf{r})), \tag{4.1}$$

where $w(\cdot;\mathbf{r})$ is the approximate solution constructed in (3.4) and $\phi(\cdot;\mathbf{r})$ is given in Lemma 3.3. We may view Q_ϵ as a function defined on $\mathcal{A}_K^a$.

LEMMA 4.1. *If $\mathbf{r} \in A_K^a$ is a critical point of Q_ϵ, then $S_\epsilon(w(\cdot;\mathbf{r}) + \phi(\cdot;\mathbf{r})) = 0$.*

Proof. Let $\mathbf{r}_*$ be a critical point of Q_ϵ. Set $g(r;\mathbf{r}) = (w(\cdot;\mathbf{r}) + \phi(\cdot;\mathbf{r}))$. At $\mathbf{r} = \mathbf{r}_*$ we have, for each l,

$$\begin{aligned} 0 = \frac{\partial Q_\epsilon(\mathbf{r}_*)}{\partial r_l} &= \int_0^1 (-\epsilon^2 \Delta g + f(g) + \epsilon\gamma(1-\Delta)^{-1} g) \frac{\partial g}{\partial r_l}\, r^{n-1} dr \\ &= \sum_{m=1}^K c_m \int_0^1 h_m \frac{\partial g}{\partial r_l}\, r^{n-1} dr. \end{aligned}$$

Here we have assumed that at $\mathbf{r}_*$, $S_\epsilon(g) = \sum_{m=1}^K c_m h_m$, because $\pi_{\mathbf{r}}(S_\epsilon(g)) = 0$. The last equation asserts that the coefficients c_m satisfy a linear homogeneous system whose ml matrix entry is $\int_0^1 h_m \frac{\partial g}{\partial r_l}\, dx$ at $\mathbf{r} = \mathbf{r}_*$.

Recall that $g = w + \phi$ and $h_m \perp \phi$ for all $\mathbf{r}$. We differentiate $0 = \int_0^1 h_m \phi\, r^{n-1} dr$ with respect to r_l to obtain

$$\int_0^1 h_m \frac{\partial \phi(r;\mathbf{r})}{\partial r_l}\, r^{n-1} dr = -\int_0^1 \frac{\partial h_m(r;\mathbf{r})}{\partial r_l} \phi(r;\mathbf{r})\, r^{n-1} dr.$$

Therefore, since $\phi = O(\epsilon)$,

$$\int_0^1 h_m \frac{\partial g}{\partial r_l}\, r^{n-1} dr = \int_0^1 \left(h_m \frac{\partial w}{\partial r_l} - \frac{\partial h_m}{\partial r_l} \phi \right) r^{n-1} dr = \delta_{ml} r_l^{n-1} \int_R (H'(t))^2\, dt + O(\epsilon).$$

Therefore the coefficient matrix is nonsingular. This implies $c_m = 0$, i.e., $S_\epsilon(g(\cdot;\mathbf{r}_*)) = 0$. ∎

The reduced problem Q_ϵ, scaled by ϵ^{-1}, converges to J given in (2.4) in C^1 locally in $\mathbf{r}$ and locally in γ.

LEMMA 4.2. *Given a compact subset $\mathcal{K}$ of $\mathcal{A}_K^a$ and an interval $[\gamma_1, \gamma_2]$, with $0 < \gamma_1 < \gamma_2 < \infty$, we have that for every $\delta > 0$ there exists $\epsilon_0 > 0$ such that when $\epsilon < \epsilon_0$, $|\epsilon^{-1} Q_\epsilon(\mathbf{r}) - J(\mathbf{r})| < \delta$ and $|\text{grad } \epsilon^{-1} Q_\epsilon(\mathbf{r}) - \text{grad } J(\mathbf{r})| < \delta$ for all $\mathbf{r} \in \mathcal{K}$ and $\gamma \in [\gamma_1, \gamma_2]$.*

Proof. Given $\mathcal{K}$ and $[\gamma_1, \gamma_2]$, we let $\mathbf{r} \in \mathcal{K}$ and $\gamma \in [\gamma_1, \gamma_2]$. We first show the C^0-convergence. Expand $I_\epsilon(w(\cdot;\mathbf{r}) + \phi(\cdot;\mathbf{r}))$ to find

$$Q_\epsilon(\mathbf{r}) = I_\epsilon(w) + \int_0^1 S_\epsilon(w)\phi\, r^{n-1} dr + \frac{1}{2} \int_0^1 \phi L_{\mathbf{r}} \phi\, dx + O(\epsilon^3). \tag{4.2}$$

The equation $\pi_{\mathbf{r}} \circ S_\epsilon(w+\phi) = 0$ implies that $S_\epsilon(w+\phi) = \sum_{j=1}^K \beta_j h_j$ for some $\beta_j \in \mathbf{R}$, which can be written as

$$S_\epsilon(w) + L_{\mathbf{r}}\phi + O(\epsilon^2) = \sum_{j=1}^K \beta_j h_j. \tag{4.3}$$

Multiply (4.3) by ϕ and integrate to find

$$\int_0^1 S_\epsilon(w)\phi\, r^{n-1}dr + \int_0^1 \phi L_{\mathbf{r}}\phi\, r^{n-1}dr + O(\epsilon^3) = 0 \tag{4.4}$$

since $\phi \perp h_j$. Substituting (4.4) into (4.2), we deduce

$$Q_\epsilon(\mathbf{r}) = I_\epsilon(w) + \frac{1}{2}\int_0^1 S_\epsilon(w)\phi\, r^{n-1}dr + O(\epsilon^3). \tag{4.5}$$

By Lemma 3.1 we obtain

$$\int_0^1 S_\epsilon(w)\phi\, r^{n-1}dr = O(\epsilon^2). \tag{4.6}$$

Now (4.5) becomes

$$Q_\epsilon(\mathbf{r}) = I_\epsilon(w(\cdot;\mathbf{r})) + O(\epsilon^2). \tag{4.7}$$

So we turn our attention to $I_\epsilon(w(\cdot;\mathbf{r}))$. Note that

$$\begin{aligned} I_\epsilon(w(\cdot;\mathbf{r})) &= \omega_{n-1}\int_0^1 \left[\frac{\epsilon^2}{2}|w_r|^2 + W(w)\right] r^{n-1}dr + \frac{\omega_{n-1}\epsilon\sigma}{2}\int_0^1 |(1-\Delta)^{-1/2}w|^2\, r^{n-1}dr \\ &= \omega_{n-1}\epsilon\tau\sum_{j=1}^K r_j^{n-1} + \frac{\omega_{n-1}\epsilon\gamma}{2}\int_0^1 |(1-\Delta)^{-1/2}w|^2\, r^{n-1}dr + O(\epsilon^2) \\ &= \epsilon J(\mathbf{r}) + O(\epsilon^2). \end{aligned} \tag{4.8}$$

Before arriving at (4.8) we have used the fact that

$$\int_{\mathbf{R}} \left[\frac{1}{2}(H')^2 + W(H)\right] dt = \tau,$$

which follows from the first integral (3.30) of H and the definition of τ. Therefore,

$$Q_\epsilon(\mathbf{r}) = \epsilon J(\mathbf{r}) + O(\epsilon^2), \tag{4.9}$$

proving the convergence at the C^0 level.

Next we show the convergence of grad $(\epsilon^{-1}Q_\epsilon)$. We calculate

$$\begin{aligned} \frac{\partial Q_\epsilon(\mathbf{r})}{\partial r_j} &= \frac{\partial}{\partial r_j} I_\epsilon(w(\cdot;\mathbf{r}) + \phi(\cdot;\mathbf{r})) \\ &= \int_\Omega S_\epsilon(w+\phi)\frac{\partial(w(\cdot;\mathbf{r}) + \phi(\cdot;\mathbf{r}))}{\partial r_j}\, dx \\ &= \int_\Omega S_\epsilon(w+\phi)\frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\, dx + \int_\Omega S_\epsilon(w+\phi)\frac{\partial \phi(\cdot;\mathbf{r})}{\partial r_j}\, dx. \end{aligned} \tag{4.10}$$

We estimate the second integral in (4.10) first. Note that since $\pi_{\mathbf{r}}(S_\epsilon(w+\phi)) = 0$,

$$S_\epsilon(w+\phi) = \sum_{l=1}^K \beta_l h_l$$

for some $\beta_l \in \mathbf{R}$. Since when $l \neq m$, h_l, and h_m are supported in disjoint sets, the h_l's are perpendicular to each other. We find

$$\beta_l = \frac{\langle S_\epsilon(w+\phi), h_l\rangle}{\|h_l\|_2^2}.$$

To estimate the numerator on the right side we write

$$S_\epsilon(w+\phi) = S_\epsilon(w) + L_{\mathbf{r}}\phi + R_{\mathbf{r}}(\phi).$$

From Lemma 3.1 we find

$$|\langle S_\epsilon(w), h_l\rangle| \leq \|S_\epsilon(w)\|_\infty \|h_l\|_1 = O(\epsilon)O(\epsilon) = O(\epsilon^2).$$

From Lemma 3.3 we have

$$|\langle L_{\mathbf{r}}\phi, h_l\rangle| = |\langle L_{\mathbf{r}}h_l, \phi\rangle| \leq \|L_{\mathbf{r}}h_l\|_1 \|\phi\|_\infty = O(\epsilon^2)O(\epsilon) = O(\epsilon^3),$$

and

$$|\langle R_{\mathbf{r}}(\phi), h_l\rangle| \leq \|R_{\mathbf{r}}(\phi)\|_\infty \|h_l\|_1 = O(\epsilon^2)O(\epsilon) = O(\epsilon^3).$$

Combing the last three estimates we obtain

$$\langle S_\epsilon(w+\phi), h_l\rangle = O(\epsilon^2).$$

Since $\|h_l\|_2^2$ is of order ϵ, we deduce

$$\beta_l = O(\epsilon). \tag{4.11}$$

The fact $\phi \perp h_l$ implies, after differentiating $\langle \phi, h_l\rangle = 0$ with respect to r_j, that

$$\int_0^1 \frac{\partial \phi(r;\mathbf{r})}{\partial r_j} h_j \, r^{n-1} dr + \int_0^1 \frac{\partial h_l(r;\mathbf{r})}{\partial r_j} \phi \, r^{n-1} dr = 0. \tag{4.12}$$

Hence the second integral in (4.10) becomes

$$\begin{aligned} \int_\Omega S_\epsilon(w+\phi) \frac{\partial \phi(\cdot;\mathbf{r})}{\partial r_j}\, dx &= \sum_{l=1}^K \omega_{n-1}\beta_l \int_0^1 h_j \frac{\partial \phi}{\partial r_j}\, r^{n-1} dr \\ &= -\sum_{l=1}^K \omega_{n-1}\beta_l \int_0^1 \phi \frac{\partial h_l}{\partial r_j}\, r^{n-1} dr. \end{aligned} \tag{4.13}$$

Our estimate of β_l, (4.11), and Lemma 3.3 imply that the last quantity of (4.13) is of order ϵ^2:

$$\int_\Omega S_\epsilon(w+\phi) \frac{\partial \phi(\cdot;\mathbf{r})}{\partial r_j}\, dx = O(\epsilon^2). \tag{4.14}$$

It remains to calculate the first integral in (4.10). We again write $S_\epsilon(w+\phi) = S_\epsilon(w) + L_{\mathbf{r}}\phi + R_{\mathbf{r}}(\phi)$ so that

$$\int_\Omega S_\epsilon(w+\phi) \frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\, dx = \int_\Omega \left[S_\epsilon(w) \frac{\partial w(\cdot;\mathbf{r})}{\partial r_j} + L_{\mathbf{r}}\phi \frac{\partial w(\cdot;\mathbf{r})}{\partial r_j} + R_{\mathbf{r}}(\phi) \frac{\partial w(\cdot;\mathbf{r})}{\partial r_j} \right] dx.$$

We must separate two cases: j is odd and j is even. When j is odd, $w(r)$ is $H(\frac{r-r_j}{\epsilon})$ for r near r_j. Moreover,

$$\frac{\partial w(r;\mathbf{r})}{\partial r_j} = -\epsilon^{-1}H'\left(\frac{r-r_j}{\epsilon}\right) + O(e^{-C/\epsilon}). \tag{4.15}$$

In this case we argue as in the estimations leading to (4.11) to conclude that

$$\int_\Omega L_{\mathbf{r}}\phi \frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\,dx = O(\epsilon^2), \quad \int_\Omega R_{\mathbf{r}}(\phi)\frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\,dx = O(\epsilon^2).$$

Therefore,

(4.16)

$$\begin{aligned}
&\int_\Omega S_\epsilon(w+\phi)\frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\,dx = \int_\Omega S_\epsilon(w)\frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\,dx + O(\epsilon^2)\\
&= \omega_{n-1}\int_0^1 \left[-\epsilon^2\left(w_{rr}+\frac{n-1}{r}w_r\right) + f(w) + \epsilon\gamma(1-\Delta)^{-1}w\right]\frac{\partial w(r;\mathbf{r})}{\partial r_j}\,r^{n-1}dr\\
&\quad + O(\epsilon^2)\\
&= \omega_{n-1}\int_0^1 \left[-\epsilon^2\frac{n-1}{r}w_r + \epsilon\gamma(1-\Delta)^{-1}w\right]\left(-\epsilon^{-1}H'\left(\frac{r-r_j}{\epsilon}\right)\right)r^{n-1}dr + O(\epsilon^2)\\
&= \omega_{n-1}\int_0^1 \left[-\epsilon\frac{n-1}{r}H'\left(\frac{r-r_j}{\epsilon}\right) + \epsilon\gamma(1-\Delta)^{-1}w\right]\left(-\epsilon^{-1}H'\left(\frac{r-r_j}{\epsilon}\right)\right)r^{n-1}dr\\
&\quad + O(\epsilon^2)\\
&= \omega_{n-1}\epsilon[(n-1)\tau r_j^{n-2} - (b-a)\gamma r_j^{n-1}((1-\Delta)^{-1}w(\cdot;\mathbf{r}))(r_j)] + O(\epsilon^2).
\end{aligned}$$

Similarly when j is even, $w(r)$ is $H(-\frac{r-r_j}{\epsilon})$ for r near r_j and

$$\frac{\partial w(r;\mathbf{r})}{\partial r_j} = \epsilon^{-1}H'\left(-\frac{r-r_j}{\epsilon}\right) + O(e^{-C/\epsilon}). \tag{4.17}$$

Then

$$\begin{aligned}
&\int_\Omega S_\epsilon(w+\phi)\frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\,dx\\
&= \omega_{n-1}\epsilon[(n-1)\tau r_j^{n-2} + (b-a)\gamma r_j^{n-1}((1-\Delta)^{-1}w(\cdot;\mathbf{r}))(r_j)] + O(\epsilon^2)\\
&= \omega_{n-1}\epsilon[(n-1)\tau r_j^{n-2} + (b-a)\gamma r_j^{n-1}((1-\Delta)^{-1}s(\cdot;\mathbf{r}))(r_j)] + O(\epsilon^2). \qquad (4.18)
\end{aligned}$$

Recall that $s \in \mathcal{A}_K^a$ is the outer part of w defined in (3.1). In conclusion, by (4.10), (4.14), (4.16), (4.18), and the calculations of $\partial J/\partial r_j$ in section 2, we have that

$$\frac{\partial Q_\epsilon(\mathbf{r})}{\partial r_j} = \int_\Omega S_\epsilon(w+\phi)\frac{\partial w(\cdot;\mathbf{r})}{\partial r_j}\,dx + O(\epsilon^2) = \epsilon\frac{\partial J(\mathbf{r})}{\partial r_j} + O(\epsilon^2). \tag{4.19}$$

This proves the lemma. □

We are now ready to prove our main results.

Proof of Theorem 1.1. Lemma 2.4, part 1, asserts that J is minimized at a point in the interior of $\mathcal{A}_K^a$ if γ is large. By Lemma 4.2 we conclude that Q_ϵ has a local

minimum $\mathbf{r}_\epsilon$ in $\mathcal{A}_K^a$. As $\epsilon \to 0$, $\mathbf{r}_\epsilon$ converges, possibly along a subsequence, to $\mathbf{r}_* \in \mathcal{A}_K^a$, which is a minimizer of J. Choose a small neighborhood $\mathcal{K}_1$ of r_* so that $\mathbf{r}_\epsilon \in \mathcal{K}_1$. If there are several critical points of Q_ϵ in $\mathcal{K}_1$, we are finished. If there is only one, it is an isolated strict local minimum and hence has index 1. On the other hand, there exists a neighborhood $\mathcal{K}$ of $\mathcal{K}_1$ on which the degree of $\operatorname{grad} J$ is zero by Lemma 2.4, part 2. Hence, by continuity, the same is true for $\operatorname{grad} Q_\epsilon$ for small ϵ. Hence there must be another critical point of Q_ϵ in $\mathcal{K}$, as required. □

Proof of Theorem 1.2. We combine Lemmas 2.3 and 4.2. □

5. Solutions with layers near the boundary. In this section, we construct solutions with multiple layers near the boundary of Ω and prove Theorem 1.3. Let $m^2 = f'(a) = f'(b) = (b-a)^2/2 > 0$. First, we construct an approximate solution.

Let $\xi(t)$ be a smooth function, such that $0 \le \xi \le 1$, $\xi(t) = 0$ for $t \le \frac{1}{2}$, and $\xi = 1$ for $t \ge \frac{2}{3}$. Let $\mathbf{r} = (r_0, \bar{r}_1, r_1, \dots, \bar{r}_k, r_k) \in D_{\epsilon,k}$, where $D_{\epsilon,k}$ is the set containing all $\mathbf{r}$ satisfying

$$1 - M\epsilon \ln \frac{1}{\epsilon} < r_0 < \bar{r}_1 < r_1 < \cdots < \bar{r}_k < r_k < 1 - \alpha\epsilon \ln \frac{1}{\epsilon},$$

and

$$\bar{r}_j - r_{j-1} \ge \alpha\epsilon \ln \frac{1}{\epsilon}, \quad r_j - \bar{r}_j \ge \alpha\epsilon \ln \frac{1}{\epsilon}, \quad j = 1, \dots, k,$$

where $\alpha > 0$ is a small constant and $M > 0$ is a large constant.

Define

$$v_{\epsilon,j}(r) = (1-\xi)b + \xi(r) H\left(\frac{r_j - r}{\epsilon}\right), \quad j = 0, 1, \dots, k,$$

and

$$\bar{v}_{\epsilon,j}(r) = (1-\xi)a + \xi(r) H\left(\frac{r - \bar{r}_j}{\epsilon}\right), \quad j = 1, \dots, k.$$

It is easy to check that $v_{\epsilon,j}$ and $\bar{v}_{\epsilon,j}$ satisfy

$$-\epsilon^2 \Delta v = -f(v) + O(\epsilon^2 |v'| + \epsilon^2) = -f(v) + O(\epsilon). \tag{5.1}$$

Let

$$w_{\epsilon,k}(r) = v_{\epsilon,0} + \sum_{j=1}^{k} \big(v_{\epsilon,j} + \bar{v}_{\epsilon,j} - a - b\big).$$

Then, using (5.1), we obtain

$$\begin{aligned}
&-\epsilon^2 \Delta w_{\epsilon,k} + m^2 w_{\epsilon,k} \\
&= -\epsilon^2 \Delta v_{\epsilon,0} - \sum_{j=1}^{k} \big(\epsilon^2 \Delta v_{\epsilon,j} + \epsilon^2 \Delta \bar{v}_{\epsilon,j}\big) + m^2 w_{\epsilon,k} \\
&= -f(v_{\epsilon,0}) - \sum_{j=1}^{k} \big(f(v_{\epsilon,j}) + f(\bar{v}_{\epsilon,j})\big) + m^2 w_{\epsilon,k} \\
&\quad + \epsilon O\left(\epsilon |v'_{\epsilon,0}| + \epsilon \sum_{j=1}^{k} \big(|v'_{\epsilon,j}| + |\bar{v}'_{\epsilon,j}|\big) + \epsilon\right).
\end{aligned} \tag{5.2}$$

Since $w_{\epsilon,k}$ does not satisfy $w'_{\epsilon,k}(1)=0$, we need to make a projection as follows. Let $Pw_{\epsilon,k}$ be the solution of

$$\begin{cases} -\epsilon^2\Delta Pw_{\epsilon,k}+m^2Pw_{\epsilon,k}=-\epsilon^2\Delta w_{\epsilon,k}+m^2w_{\epsilon,k} & \text{in } \Omega,\\ (Pw_{\epsilon,k})'(1)=0. \end{cases} \tag{5.3}$$

We denote $\varphi_\epsilon = w_{\epsilon,k}-Pw_{\epsilon,k}$. Then φ_ϵ satisfies

$$\begin{cases} -\epsilon^2\Delta\varphi_\epsilon+m^2\varphi_\epsilon=0 \ \text{ in } \Omega,\\ \varphi'_\epsilon(1)=w'_{\epsilon,k}(1). \end{cases} \tag{5.4}$$

By the maximum principle, we see $\varphi_\epsilon<0$.

We have the following estimates for φ_ϵ.

LEMMA 5.1. *For any small $\theta>0$, there is a constant $C>0$, such that*

$$|\varphi_\epsilon(r)|\le Ce^{-m(1-r_k)/\epsilon}e^{-m(1-\theta)(1-r)/\epsilon}.$$

In particular,

$$|\varphi_\epsilon(r)|\le Ce^{-m(1-\theta)|r-r_k|/\epsilon}.$$

Proof. Let $G_\epsilon(Y,y)$ and $G(Y,y)$ be the Green function of $-\epsilon^2\Delta+m^2I$ in Ω and $-\Delta+m^2I$ in $\Omega_{\epsilon,y}=\{Y:\epsilon Y+y\in\Omega\}$ subject to the Neumann boundary condition, respectively. Then

$$G_\epsilon(Y,y)=\frac{1}{\epsilon^n}G\left(\frac{Y-y}{\epsilon},0\right).$$

We have

$$\varphi_\epsilon(y)=\epsilon^2\int_{\partial\Omega}G_\epsilon(Y,y)w'_{\epsilon,k}(1)\,dY.$$

This, together with $|w'_{\epsilon,k}(1)|=\epsilon^{-1}\big|\sum_{j=1}^k H'\big(\frac{1-\bar r_j}{\epsilon}\big)-\sum_{j=0}^k H'\big(\frac{1-r_j}{\epsilon}\big)\big|\le C\epsilon^{-1}e^{-m(1-r_k)/\epsilon}$, gives

$$\begin{aligned} |\varphi_\epsilon(y)| &\le C\epsilon e^{-m(1-r_k)/\epsilon}\int_{\partial\Omega}|G_\epsilon(Y,y)|dY\\ &= C\epsilon e^{-m(1-r_k)/\epsilon}\frac{1}{\epsilon^n}\int_{\partial\Omega}\left|G\left(\frac{Y-y}{\epsilon},0\right)\right|dY\\ &= Ce^{-m(1-r_k)/\epsilon}\int_{\partial\Omega_{\epsilon,y}}|G(Y,0)|\,dY\le Ce^{-m(1-r_k)/\epsilon}e^{-m(1-\theta)(1-r)/\epsilon}, \end{aligned} \tag{5.5}$$

since $G(Y,0)\sim\frac{1}{|Y|^{N-2}}$ as $|Y|\to0$, and $|G(Y,0)|\le Ce^{-m|Y|}$ as $|Y|\to+\infty$.

Since for $r\in[0,1]$ we have

$$|r-r_k|\le|1-r_k|+|r-1|=1-r_k+1-r,$$

as a result

$$|\varphi_\epsilon(y)|\le Ce^{-m(1-r_k)/\epsilon}e^{-m(1-\theta)(1-r)/\epsilon}\le Ce^{-m(1-\theta)|r-r_k|/\epsilon}.$$

So, we have proved the lemma. □

Define

$$I^*(u) = \frac{\epsilon^2}{2}\int_\Omega |Du|^2 + \int_\Omega W(u),$$

where $W(t) = \int_a^t f(s)\,ds$.

Next, we estimate $I^*\big(Pw_{\epsilon,k}\big)$. We have

PROPOSITION 5.2.

$$I^*(Pw_{\epsilon,k}) = \epsilon A r_0^{n-1} + \epsilon A \sum_{j=1}^{k} (\bar r_j^{n-1} + r_j^{n-1})$$

$$-B\epsilon \sum_{j=1}^{k} \big(e^{-m(\bar r_j - r_{j-1})/\epsilon} + e^{-m(r_j - \bar r_j)/\epsilon}\big) - B_\epsilon \omega_{n-1}\epsilon e^{-2m(1-r_k)/\epsilon}$$

$$+ \epsilon O\left(\sum_{j=1}^{k} \big(e^{-(1+\sigma)m(\bar r_j - r_{j-1})/\epsilon} + e^{-(1+\sigma)m(r_j - \bar r_j)/\epsilon}\big) + e^{-2(1+\sigma)m(1-r_k)/e} + \epsilon\right),$$

where $A > 0$ and $B > 0$ are some constants independent of ϵ, $B_\epsilon > 0$ is a constant depending on ϵ, satisfying $b_2 \ge B_\epsilon \ge b_1 > 0$ for some constants b_2 and b_1, and $\sigma > 0$ is a constant.

From Proposition 5.2, we see that there are three factors that affect the energy of $Pw_{\epsilon,k}$:

(i) The contribution from the layers is

$$\epsilon A r_0^{n-1} + \epsilon A \sum_{j=1}^{k} (\bar r_j^{n-1} + r_j^{n-1}).$$

(ii) The contribution from the interaction between the layers is

$$-B\epsilon \sum_{j=1}^{k} \big(e^{-m(\bar r_j - r_{j-1})/\epsilon} + e^{-m(r_j - \bar r_j)/\epsilon}\big).$$

(iii) The contribution from the Neumann boundary condition is

$$B_\epsilon \omega_{n-1}\epsilon e^{-2m(1-r_k)/\epsilon}.$$

So, we conclude that the energy will decrease if the layer moves away from the boundary, or the layers move toward each other, or the layer moves toward the boundary. As a result, if $I^*(Pw_{\epsilon,k})$ attains its maximum, the layers must be suitably separated and stay suitably close to the boundary.

We will prove Proposition 5.2 by proving three lemmas.

Let

$$\hat f(t) = f(t + a).$$

Then $\hat F(t) = \int_0^t \hat f(s)\,ds = W(t + a)$.

Denote $\psi_{\epsilon,j} = v_{\epsilon,j} - a$, $j = 0, \dots, k$, $\bar{\psi}_{\epsilon,j} = \bar{v}_{\epsilon,j} - a$, $j = 1, \dots, k$, and $\tilde{v}_{\epsilon,j} = \psi_{\epsilon,j} + \bar{\psi}_{\epsilon,j} - (b-a) = v_{\epsilon,j} + \bar{v}_{\epsilon,j} - b - a$, $j = 1, \dots, k$, $\tilde{v}_{\epsilon,0} = \psi_{\epsilon,0}$. So, with this notation, we see that

$$f(v_{\epsilon,j}) = \hat{f}(\psi_{\epsilon,j}), \quad f(\bar{v}_{\epsilon,j}) = \hat{f}(\bar{\psi}_{\epsilon,j}),$$

and

$$W(w_{\epsilon,k}) = \hat{F}(\bar{w}_{\epsilon,k}),$$

where $\bar{w}_{\epsilon,k} = w_{\epsilon,k} - a = \sum_{j=0}^{k} \tilde{v}_j$.

It is easy to see that

$$P\bar{w}_{\epsilon,k} = Pw_{\epsilon,k} - a.$$

Let

$$\hat{I}^*(u) = \frac{\epsilon^2}{2} \int_\Omega |Du|^2 + \int_\Omega \hat{F}(u).$$

By (5.2) and (5.3), we have

(5.6)

$$\begin{aligned}
&I^*\big(Pw_{\epsilon,k}\big) = \hat{I}^*\big(P\bar{w}_{\epsilon,k}\big) \\
&= \frac{1}{2}\int_\Omega \big(\epsilon^2 |DP\bar{w}_{\epsilon,k}|^2 + m^2 (P\bar{w}_{\epsilon,k})^2\big) + \int_\Omega \hat{F}(P\bar{w}_{\epsilon,k}) - \frac{1}{2} m^2 \int_\Omega (P\bar{w}_{\epsilon,k})^2 \\
&= \frac{1}{2}\int_\Omega \left(-\hat{f}(\psi_{\epsilon,0}) - \sum_{j=1}^{k} \big(\hat{f}(\psi_{\epsilon,j}) + \bar{f}(\hat{\psi}_{\epsilon,j})\big) + m^2 \hat{w}_{\epsilon,k} \right) P\bar{w}_{\epsilon,k} \\
&\quad + \epsilon O\left(\int_\Omega \left(\epsilon |\psi'_{\epsilon,0}| + \epsilon \sum_{j=1}^{k} (|\psi'_{\epsilon,j}| + |\bar{\psi}'_{\epsilon,j}|) \right) |Pw_{\epsilon,k}| \right) \\
&\quad + \int_\Omega \left(\hat{F}(P\bar{w}_{\epsilon,k}) - \frac{m^2}{2} (P\bar{w}_{\epsilon,k})^2 \right) \\
&= \frac{1}{2}\int_\Omega \left(-\hat{f}(\psi_{\epsilon,0}) - \sum_{j=1}^{k} \big(\hat{f}(\psi_{\epsilon,j}) + \hat{f}(\bar{\psi}_{\epsilon,j})\big) \right) \bar{w}_{\epsilon,k} + \int_\Omega \hat{F}(\bar{w}_{\epsilon,k}) \\
&\quad + \int_\Omega \left(-\hat{f}(\bar{w}_{\epsilon,k}) + \frac{1}{2} \left(\hat{f}(\psi_{\epsilon,0}) + \sum_{j=1}^{k} \big(\hat{f}(\psi_{\epsilon,j}) + \hat{f}(\bar{\psi}_{\epsilon,j})\big) \right) + \frac{1}{2} m^2 \bar{w}_{\epsilon,k} \right) \varphi_\epsilon \\
&\quad + \int_\Omega \left(\hat{F}(P\bar{w}_{\epsilon,k}) - \hat{F}(\bar{w}_{\epsilon,k}) + \hat{f}(\bar{w}_{\epsilon,k})\varphi_\epsilon - \frac{1}{2} m^2 \varphi_\epsilon^2 \right) + O(\epsilon^2) \\
&=: \hat{I}_1 + \hat{I}_2 + \hat{I}_3 + O(\epsilon^2).
\end{aligned}$$

LEMMA 5.3. *We have*

$$|\hat{I}_3| \le C\epsilon e^{-(3-\theta)m(1-r_k)/\epsilon},$$

where $\theta > 0$ is any small constant.

Proof. By definition, we have

$$\hat{I}_3 = \frac{1}{2}\int_\Omega \big(\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2\big) + O\left(\int_\Omega |\varphi_\epsilon|^3\right). \tag{5.7}$$

Write

$$\int_\Omega \big(\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2\big) = \int_{r_k<r\le 1} \big(\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2\big) + \int_{r\le r_k} \big(\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2\big)$$

$$= \int_{r_k<r\le 1} \big(\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2\big) + O\left(\int_{r\le r_k} \varphi_\epsilon^2\right). \tag{5.8}$$

Using Lemma 5.1, we obtain for $r \le r_k$

$$|\varphi_\epsilon(r)| \le Ce^{-m(1-r_k)/\epsilon} e^{-m(1-\theta)(1-r)/\epsilon} \le Ce^{-(2-\theta)m(1-r_k)/\epsilon}.$$

Thus,

$$\begin{aligned}\int_{r\le r_k} |\varphi_\epsilon|^2 &\le Ce^{-(2-\theta)2m(1-r_k)/\epsilon} \int_{r\le r_k} |\varphi_\epsilon|^\theta \\ &\le Ce^{-(2-\theta)2m(1-r_k)/\epsilon} \int_{r\le r_k} e^{-\theta(1-\theta)m|r-r_k|/\epsilon} \\ &\le C\epsilon e^{-(2-\theta)2m(1-r_k)/\epsilon}.\end{aligned} \tag{5.9}$$

On the other hand, for $r \ge r_k$, we have

$$\begin{aligned}\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2 &= \hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - \hat{f}'(0)\varphi_\epsilon^2 \\ &= O\big(|\bar{w}_{\epsilon,k}|\varphi_\epsilon^2\big) = O\Big(e^{-m(r-r_k)/\epsilon} e^{-(2-\theta)m(1-r_k)/\epsilon} e^{-m(2-\theta)(1-\theta)(1-r)/\epsilon} |\varphi_\epsilon^\theta|\Big) \\ &= O\Big(e^{-m(3-\theta)(1-r_k)/\epsilon} |\varphi_\epsilon^\theta|\Big).\end{aligned} \tag{5.10}$$

Using (5.10), we obtain

$$\begin{aligned}&\int_{r_k<r\le 1} \big(\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2\big) \\ &\qquad \le Ce^{-m(3-\theta)(1-r_k)/\epsilon} \int_\Omega |\varphi_\epsilon^\theta| \le C\epsilon e^{-m(3-\theta)(1-r_k)/\epsilon}.\end{aligned} \tag{5.11}$$

Combining (5.9) and (5.10), we are led to

$$\int_\Omega \big(\hat{f}'(\bar{w}_{\epsilon,k})\varphi_\epsilon^2 - m^2\varphi_\epsilon^2\big) = O\Big(\epsilon e^{-m(3-\theta)(1-r_k)/\epsilon}\Big). \tag{5.12}$$

Finally,

$$\begin{aligned}\int_\Omega |\varphi_\epsilon|^3 &= \int_{r_k<r\le 1} |\varphi_\epsilon|^3 + \int_{r\le r_k} |\varphi_\epsilon|^3 \le C\epsilon e^{-m(3-\theta)(1-r_k)/\epsilon} + C\epsilon e^{-(2-\theta)2m(1-r_k)/\epsilon} \\ &\le C\epsilon e^{-(3-\theta)m(1-r_k)/\epsilon}.\end{aligned} \tag{5.13}$$

Combining (5.7), (5.12), and (5.13), we obtain the result. □

LEMMA 5.4. *We have*

$$\hat{I}_1 = \epsilon A r_0^{n-1} + \epsilon A \sum_{j=1}^{k} (\bar{r}_j^{n-1} + r_j^{n-1}) - B\epsilon \sum_{j=1}^{k} \big(e^{-m(\bar{r}_j - r_{j-1})/\epsilon} + e^{-m(r_j - \bar{r}_j)/\epsilon}\big)$$

$$+ \epsilon O\left(\sum_{j=1}^{k} \big(e^{-(1+\sigma)m(\bar{r}_j - r_{j-1})/\epsilon} + e^{-(1+\sigma)m(r_j - \bar{r}_j)/\epsilon}\big) + e^{-(2+\sigma)m(1-r_k)/\epsilon} + \epsilon \right),$$

where $A = \omega_{n-1} \int_{-\infty}^{+\infty} \big(W(H(t)) - \frac{1}{2} f(H(t))H(t)\big)\, dt > 0$, $B > 0$ *is a constant, and* $\sigma > 0$ *is a small constant.*

Proof. It is easy to check that for any bounded t_1 and t_2,

$$\hat{F}(t_1 + t_2) - \hat{F}(t_1) - \hat{F}(t_2) = \hat{f}(t_1)t_2 + \big(\hat{f}(t_2) - \hat{f}'(0)t_2\big)t_1 + O\big(|t_1 t_2|^2\big).$$

Thus,

$$\hat{F}(\bar{w}_{\epsilon,k}) = \sum_{j=0}^{k} \hat{F}(\tilde{v}_{\epsilon,j}) + \sum_{i<j} \hat{f}(\tilde{v}_{\epsilon,i})\tilde{v}_{\epsilon,j} + \sum_{i=1}^{k-1} \left(\hat{f}\left(\sum_{j=i+1}^{k} \tilde{v}_{\epsilon,j} \right) - \hat{f}'(0) \sum_{j=i+k}^{k} \tilde{v}_{\epsilon,j} \right) \tilde{v}_{\epsilon,i}$$

$$+ O\left(\sum_{i \neq j} |\tilde{v}_{\epsilon,i}\tilde{v}_{\epsilon,j}|^2 \right). \tag{5.14}$$

Using (5.14), we can write

$$\hat{I}_1 = \sum_{j=0}^{k} \int_\Omega \left(\hat{F}(\tilde{v}_{\epsilon,j}) - \frac{1}{2}\hat{f}(\tilde{v}_{\epsilon,j})\tilde{v}_{\epsilon,j} \right) + \frac{1}{2} \int_\Omega \sum_{j=1}^{k} \Big(\hat{f}(\tilde{v}_{\epsilon,j}) - \hat{f}(\psi_{\epsilon,j}) - \hat{f}(\bar{\psi}_{\epsilon,j}) \Big) \bar{w}_{\epsilon,k}$$

$$+ \sum_{i=1}^{k-1} \int_\Omega \left(\hat{f}\left(\sum_{j=i+1}^{k} \tilde{v}_{\epsilon,j} \right) - \hat{f}'(0) \sum_{j=i+1}^{k} \tilde{v}_{\epsilon,j} \right) \tilde{v}_{\epsilon,i} + O\left(\sum_{i \neq j} \int_\Omega |\tilde{v}_{\epsilon,i}\tilde{v}_{\epsilon,j}|^2 \right). \tag{5.15}$$

It is easy to prove that

$$\int_\Omega |\tilde{v}_{\epsilon,i}\tilde{v}_{\epsilon,j}|^2 = \epsilon O\Big(e^{-2m|\bar{r}_j - r_i|/\epsilon}\Big), \quad j > i. \tag{5.16}$$

On the other hand, since for any $t \in (0, b-a)$,

$$\hat{f}(t) - \hat{f}'(0)t < 0,$$

we see

$$\int_\Omega \left(\hat{f}\left(\sum_{j=i+1}^{k} \tilde{v}_{\epsilon,j} \right) - \hat{f}'(0) \sum_{j=i+1}^{k} \tilde{v}_{\epsilon,j} \right) \tilde{v}_{\epsilon,i} = -(B + o(1))\epsilon e^{-m(\bar{r}_{i+1} - r_i)/\epsilon}, \tag{5.17}$$

where $o(1) \to 0$ as $\epsilon \to 0$, and

$$B = -\int_{-\infty}^{+\infty} \big(\hat{f}(H(t) - a) - \hat{f}'(0)(H(t) - a)\big) e^{-mt}\, dt > 0.$$

Since $\hat{f}(t) = -\hat{f}(b-a-t)$, we obtain

$$\begin{aligned}
&\hat{f}(\tilde{v}_{\epsilon,j}) - \hat{f}(\psi_{\epsilon,j}) - \hat{f}(\bar{\psi}_{\epsilon,j}) \\
&= -\hat{f}(b-a-\psi_{\epsilon,j}+b-a-\bar{\psi}_{\epsilon,j}) + \hat{f}(b-a-\psi_{\epsilon,j}) + \hat{f}(b-a-\bar{\psi}_{\epsilon,j}) \\
&= O\big(|(b-a-\psi_{\epsilon,j}||(b-a-\bar{\psi}_{\epsilon,j}|\big).
\end{aligned}$$

Thus, for $i \neq j$, we see

$$\int_\Omega \Big(\hat{f}(\tilde{v}_{\epsilon,j}) - \hat{f}(\psi_{\epsilon,j}) - \hat{f}(\bar{\psi}_{\epsilon,j})\Big)\tilde{v}_i = \epsilon O\left(\sum_{j=1}^{k}\big(e^{-2m|\bar{r}_j - r_{j-1}|/\epsilon} + e^{-2m|r_j-\bar{r}_j|/\epsilon}\big)\right).$$

So, we have

$$\begin{aligned}
&\sum_{j=0}^{k}\int_\Omega \left(\hat{F}(\tilde{v}_{\epsilon,j}) - \frac{1}{2}\hat{f}(\tilde{v}_{\epsilon,j})\tilde{v}_{\epsilon,j}\right) + \frac{1}{2}\int_\Omega \sum_{j=1}^{k}\Big(\hat{f}(\tilde{v}_{\epsilon,j}) - \hat{f}(\psi_{\epsilon,j}) - \hat{f}(\bar{\psi}_{\epsilon,j})\Big)\bar{w}_{\epsilon,k} \\
&= \sum_{j=0}^{k}\int_\Omega \left(\hat{F}(\tilde{v}_{\epsilon,j}) - \frac{1}{2}\hat{f}(\tilde{v}_{\epsilon,j})\tilde{v}_{\epsilon,j}\right) + \frac{1}{2}\int_\Omega \sum_{j=1}^{k}\Big(\hat{f}(\tilde{v}_{\epsilon,j}) - \hat{f}(\psi_{\epsilon,j}) - \hat{f}(\bar{\psi}_{\epsilon,j})\Big)\tilde{v}_j \\
&\quad + \epsilon O\left(\sum_{j=1}^{k}\big(e^{-2m|\bar{r}_j - r_{j-1}|/\epsilon} + e^{-2m|r_j-\bar{r}_j|/\epsilon}\big)\right) \\
&= \sum_{j=0}^{k}\int_\Omega \hat{F}(\tilde{v}_{\epsilon,j}) - \frac{1}{2}\int_\Omega \sum_{j=1}^{k}\Big(\hat{f}(\psi_{\epsilon,j}) + \hat{f}(\bar{\psi}_{\epsilon,j})\Big)\tilde{v}_j \\
&\quad + \epsilon O\left(\sum_{j=1}^{k}\big(e^{-2m|\bar{r}_j - r_{j-1}|/\epsilon} + e^{-2m|r_j-\bar{r}_j|/\epsilon}\big)\right). \qquad (5.18)
\end{aligned}$$

But from $\hat{f}(t) = -\hat{f}(b-a-t)$, we see

$$\hat{f}(\psi_{\epsilon,j})(\bar{\psi}_{\epsilon,j} - (b-a)) = \hat{f}(b-a-\psi_{\epsilon,j})(b-a-\bar{\psi}_{\epsilon,j}),$$

and

$$\hat{f}(\bar{\psi}_{\epsilon,j})(\psi_{\epsilon,j} - (b-a)) = \hat{f}(b-a-\bar{\psi}_{\epsilon,j})(b-a-\psi_{\epsilon,j}).$$

We also have

$$\begin{aligned}\hat{F}(\tilde{v}_{\epsilon,j}) &= \hat{F}(b-a-\tilde{v}_{\epsilon,j}) = \hat{F}(b-a-\psi_{\epsilon,j}+b-a-\bar{\psi}_{\epsilon,j})\\ &= \hat{F}(b-a-\psi_{\epsilon,j}) + \hat{F}(b-a-\bar{\psi}_{\epsilon,j})\\ &\quad + \frac{1}{2}\hat{f}(b-a-\psi_{\epsilon,j})(b-a-\bar{\psi}_{\epsilon,j})\\ &\quad + \frac{1}{2}\Big(\hat{f}(b-a-\bar{\psi}_{\epsilon,j}) - \hat{f}'(0)(b-a-\bar{\psi}_{\epsilon,j})\Big)(b-a-\psi_{\epsilon,j})\\ &\quad + \frac{1}{2}\hat{f}(b-a-\bar{\psi}_{\epsilon,j})(b-a-\psi_{\epsilon,j})\\ &\quad + \frac{1}{2}\Big(\hat{f}(b-a-\psi_{\epsilon,j}) - \hat{f}'(0)(b-a-\psi_{\epsilon,j})\Big)(b-a-\bar{\psi}_{\epsilon,j})\\ &\quad + O\Big(|b-a-\psi_{\epsilon,j}|^2|b-a-\bar{\psi}_{\epsilon,j}|^2\Big). \qquad (5.19)\end{aligned}$$

Using (5.19), we obtain

$$\begin{aligned}&\int_\Omega \hat{F}(\tilde{v}_{\epsilon,j}) - \frac{1}{2}\int_\Omega \Big(\hat{f}(\psi_{\epsilon,j}) + \hat{f}(\bar{\psi}_{\epsilon,j})\Big)\tilde{v}_j\\ &= \int_\Omega \left(\hat{F}(\psi_{\epsilon,j}) - \frac{1}{2}\hat{f}(\psi_{\epsilon,j})\psi_{\epsilon,j}\right) + \int_\Omega \left(\hat{F}(\bar{\psi}_{\epsilon,j}) - \frac{1}{2}\hat{f}(\bar{\psi}_{\epsilon,j})\bar{\psi}_{\epsilon,j}\right)\\ &\quad -(B+o(1))\epsilon e^{-m(r_j-\bar{r}_j)/\epsilon} + \epsilon O\left(\sum_{j=1}^k e^{-2m(r_j-\bar{r}_j)/\epsilon}\right). \qquad (5.20)\end{aligned}$$

Combining (5.15), (5.16), (5.17), (5.18), and (5.20), we obtain

$$\begin{aligned}\hat{I}_1 &= \sum_{j=0}^k \int_\Omega \left(\hat{F}(\psi_{\epsilon,j}) - \frac{1}{2}\hat{f}(\psi_{\epsilon,j})\psi_{\epsilon,j}\right) + \sum_{j=1}^k \int_\Omega \left(\hat{F}(\bar{\psi}_{\epsilon,j}) - \frac{1}{2}\hat{f}(\bar{\psi}_{\epsilon,j})\bar{\psi}_{\epsilon,j}\right)\\ &\quad - (B+o(1))\epsilon \sum_{j=1}^k \big(e^{-m(r_j-\bar{r}_j)/\epsilon} + e^{-m(\bar{r}_j-r_{j-1})/\epsilon}\big)\\ &\quad + \epsilon O\left(\sum_{j=1}^k \big(e^{-2m(r_j-\bar{r}_j)/\epsilon} + e^{-2m(\bar{r}_j-r_{j-1})/\epsilon}\big)\right). \qquad (5.21)\end{aligned}$$

Finally, let $\hat{H}(t) = H(t) - a$. Then we have

(5.22)

$$\begin{aligned}\int_\Omega \hat{F}(\psi_{\epsilon,j}) &= \omega_{n-1}\epsilon r_j^{n-1}\int_{-(1-r_j)/\epsilon}^{r_j/\epsilon} \hat{F}(\hat{H}(t))\,dt + O(\epsilon^2)\\ &= \omega_{n-1}\epsilon r_j^{n-1}\int_{-\infty}^{+\infty} \hat{F}(\hat{H}(t))\,dt - \omega_{n-1}\epsilon r_j^{n-1}\int_{-\infty}^{-(1-r_j)/\epsilon} \hat{F}(\hat{H}(t))\,dt + O(\epsilon^2),\end{aligned}$$

and

$$\int_\Omega \hat f(\psi_{\epsilon,j})\psi_{\epsilon,j} = \omega_{n-1}\epsilon r_j^{n-1}\int_{-\infty}^{+\infty} \hat f(\hat H(t))\hat H(t)\,dt - \omega_{n-1}\epsilon r_j^{n-1} \times \int_{-\infty}^{-(1-r_j)/\epsilon} \hat f(\hat H(t))\hat H(t)\,dt + O(\epsilon^2). \tag{5.23}$$

But from $\hat F(0) = \hat f(0) = 0$, we see

$$\left|\int_{-\infty}^{-(1-r_j)/\epsilon} \hat F(\hat H(t))\,dt - \frac{1}{2}\int_{-\infty}^{-(1-r_j)/\epsilon} \hat f(\hat H(t))\hat H(t)\,dt\right| \le \int_{-\infty}^{-(1-r_j)/\epsilon} \hat H^3(t)\,dt = O\left(\int_{-\infty}^{-(1-r_j)/\epsilon} e^{3mt}\,dt\right)$$

$$\le Ce^{-3m(1-r_j)/\epsilon}. \tag{5.24}$$

Combining (5.22), (5.23), and (5.24), we are led to

$$\int_\Omega \left(\hat F(\psi_{\epsilon,j}) - \frac{1}{2}\hat f(\psi_{\epsilon,j})\psi_{\epsilon,j}\right) = \omega_{n-1}\epsilon r_j^{n-1}\int_{-\infty}^{+\infty} \hat F(\hat H(t))\,dt - \frac{1}{2}\omega_{n-1}\epsilon r_j^{n-1}\int_{-\infty}^{+\infty} \hat f(\hat H(t))\hat H(t)\,dt + \epsilon O\big(\epsilon + e^{-3m(1-r_j)/\epsilon}\big)$$

$$= \omega_{n-1}\epsilon r_j^{n-1}A + \epsilon O\big(\epsilon + e^{-3m(1-r_k)/\epsilon}\big). \tag{5.25}$$

Similarly, we can obtain

$$\int_\Omega \left(\hat F(\bar\psi_{\epsilon,j}) - \frac{1}{2}\hat f(\bar\psi_{\epsilon,j})\bar\psi_{\epsilon,j}\right) = \omega_{n-1}\epsilon \bar r_j^{n-1}A + \epsilon O\big(\epsilon + e^{-3m(1-r_k)/\epsilon}\big). \tag{5.26}$$

Combining (5.21), (5.25), and (5.26), we prove this lemma. □

LEMMA 5.5. *We have*

$$\hat I_2 = -\epsilon B_\epsilon \omega_{n-1}e^{-2m(1-r_k)/\epsilon} + \epsilon O\Big(e^{-2m(1-r_k)/\epsilon} + \epsilon\Big) + \epsilon O\left(\sum_{j=1}^{k}\big(e^{-2(1+\sigma)(r_j-\bar r_j)/\epsilon} + e^{-2(1+\sigma)(\bar r_j - r_{j-1})/\epsilon}\big)\right), \tag{5.27}$$

where B_ϵ is a constant with $b_2 \ge B_\epsilon \ge b_1 > 0$, and $\sigma > 0$ is a small constant.

Proof. We have

$$\hat I_2 = \frac{1}{2}\int_\Omega \left(-\hat f(\psi_{\epsilon,0}) - \sum_{j=1}^{k}\big(\hat f(\psi_{\epsilon,j}) + \hat f(\bar\psi_{\epsilon,j})\big) + m^2\bar w_{\epsilon,k}\right)\varphi_\epsilon - \int_\Omega \left(\hat f(\bar w_{\epsilon,k}) - \hat f(\psi_{\epsilon,0}) - \sum_{j=1}^{k}\big(\hat f(\psi_{\epsilon,j}) + \hat f(\bar\psi_{\epsilon,j})\big)\right)\varphi_\epsilon$$

$$=: \hat I_{2,1} + \hat I_{2,2}. \tag{5.28}$$

It is easy to see that

$$|\hat{I}_{2,2}| \le C\epsilon \left(\sum_{j=1}^{k} \big(e^{-m(r_j-\bar{r}_j)/\epsilon} + e^{-m(\bar{r}_j-r_{j-1})/\epsilon} \big) \right) \|\varphi_\epsilon\|_\infty^{1-\theta}$$

$$\le C\epsilon \left(\sum_{j=1}^{k} \big(e^{-m(\bar{r}_j-r_{j-1})/\epsilon} + e^{-m(r_j-\bar{r}_j)/\epsilon} \big) \right) e^{-m(1-\theta)(1-r_k)/\epsilon}$$

$$\le C\epsilon \left(\sum_{j=1}^{k} \big(e^{-(1+\sigma)m(\bar{r}_j-r_{j-1})/\epsilon} + e^{-(1+\sigma)m(r_j-\bar{r}_j)/\epsilon} \big) + e^{-(2+\sigma)m(1-r_k)/\epsilon} \right). \tag{5.29}$$

On the other hand, using (5.2) and (5.4), we see

(5.30)

$$\hat{I}_{2,1} = \frac{1}{2} \int_\Omega \left(-\epsilon^2 \Delta \bar{w}_{\epsilon,k} + m^2 \bar{w}_{\epsilon,k} + \epsilon O\left(\epsilon |\psi'_{\epsilon,0}| + \epsilon \sum_{j=1}^{k} \big(|\psi'_{\epsilon,j}| + |\bar{\psi}'_{\epsilon,j}| \big) + \epsilon \right) \right) \varphi_\epsilon$$

$$= \frac{1}{2} \omega_{n-1} \epsilon^2 \big(-\bar{w}'_{\epsilon,k}(1)\varphi_\epsilon(1) + \bar{w}_{\epsilon,k}(1)\bar{w}'_{\epsilon,k}(1) \big)$$

$$+ \epsilon O\left(\int_\Omega \left(\epsilon |\psi'_{\epsilon,0}| + \epsilon \sum_{j=1}^{k} \big(|\psi'_{\epsilon,j}| + |\bar{\psi}'_{\epsilon,j}| \big) + \epsilon \right) |\varphi_\epsilon| \right)$$

$$= -B_\epsilon \omega_{n-1} \epsilon e^{-2m(1-r_k)/\epsilon} + \epsilon O\left(\int_\Omega \left(\epsilon |\psi'_{\epsilon,0}| + \epsilon \sum_{j=1}^{k} \big(|\psi'_{\epsilon,j}| + |\bar{\psi}'_{\epsilon,j}| \big) + \epsilon \right) |\varphi_\epsilon| \right),$$

where $B_\epsilon = \frac{1}{2} e^{2m(1-r_j)/\epsilon} (\epsilon \bar{w}'_{\epsilon,k}(1)\varphi_\epsilon(1) - \epsilon \bar{w}_{\epsilon,k}(1)\bar{w}'_{\epsilon,k}(1))$. By Lemma 5.1, noting that $\varphi_\epsilon < 0$, $w_{\epsilon,k}(1) \sim e^{-m(1-r_k)/\epsilon}$, and $\epsilon w'_{\epsilon,k}(1) \sim -e^{-m(1-r_k)/\epsilon}$, we see easily that there are $b_2 > b_1 > 0$, independent of ϵ, such that $b_2 \ge B_\epsilon \ge b_1$.

We have

$$\epsilon \int_\Omega |\psi'_{\epsilon,j}||\varphi_\epsilon| = \epsilon \int_{r \le r_k} |\psi'_{\epsilon,j}||\varphi_\epsilon| + \epsilon \int_{r \ge r_k} |\psi'_{\epsilon,j}||\varphi_\epsilon|$$

$$\le C\epsilon e^{-(2-\theta)m(1-r_k)/\epsilon} + C e^{-m(1-r_k)/\epsilon} \int_{r \ge r_k} e^{-m(r-r_j)/\epsilon} e^{-(1-\theta)m(1-r)/\epsilon}$$

$$\le C\epsilon e^{-(2-\theta)m(1-r_k)/\epsilon} + C e^{-(2-\theta)m(1-r_k)/\epsilon} \epsilon \ln \frac{1}{\epsilon}$$

$$\le C\big(e^{-(2+\sigma)m(1-r_k)/\epsilon} + \epsilon \big), \tag{5.31}$$

$$\epsilon \int_\Omega |\bar{\psi}'_{\epsilon,j}||\varphi_\epsilon| \le C\big(e^{-(2+\sigma)m(1-r_k)/\epsilon} + \epsilon \big), \tag{5.32}$$

and

$$\epsilon \int_\Omega |\varphi_\epsilon| \le C\big(e^{-(2+\sigma)m(1-r_k)/\epsilon} + \epsilon\big). \tag{5.33}$$

Combining (5.30), (5.31), (5.32), and (5.33), we obtain

$$\hat{I}_{2,1} = -B_\epsilon \omega_{n-1} \epsilon e^{-2m(1-r_k)/\epsilon} + \epsilon O\big(e^{-(2+\sigma)m(1-r_k)/\epsilon} + \epsilon\big). \tag{5.34}$$

So, the result follows from (5.28), (5.29), and (5.34). ∎

Proof of Proposition 5.2. The proof follows from Lemmas 5.3, 5.4, and 5.5. ∎

Now, we look at the reduction. We have

$$S_\epsilon(Pw_{\epsilon,k}) = -\epsilon^2 \Delta P w_{\epsilon,k} + f(Pw_{\epsilon,k}) + \epsilon\gamma(1-\Delta)^{-1} P w_{\epsilon,k}$$

$$= f(w_{\epsilon,k}) - f(v_{\epsilon,0}) - \sum_{j=1}^{k} \big(f(v_{\epsilon,j}) + f(\bar{v}_{\epsilon,j})\big) \tag{5.35}$$

$$+ f(Pw_{\epsilon,k}) - f(w_{\epsilon,k}) + m^2(w_{\epsilon,k} - Pw_{\epsilon,k}) + O(\epsilon) \tag{5.36}$$

$$= O\left(e^{-m(1-r_k)/\epsilon} + \sum_{j=1}^{k} \big(e^{-m(\bar{r}_j - r_{j-1})/\epsilon} + e^{-m(r_j - \bar{r}_j)/\epsilon}\big) + \epsilon\right). \tag{5.37}$$

Similar to the discussion in section 3, we can prove the following result.

LEMMA 5.6. *There is a* $\phi_\epsilon \in \mathcal{F}_{\mathbf{r}}$, *such that* $\pi_{\mathbf{r}} \circ S_\epsilon(w_{\epsilon,k} + \phi_\epsilon) = 0$. *Moreover,*

$$\|\phi_\epsilon\|_\infty \le C\left(e^{-m(1-r_k)/\epsilon} + \sum_{j=1}^{k} \big(e^{-m(\bar{r}_j - r_{j-1})/\epsilon} + e^{-m(r_j - \bar{r}_j)/\epsilon}\big) + \epsilon\right).$$

LEMMA 5.7. *We have*

$$I_\epsilon(Pw_{\epsilon,k} + \phi_\epsilon) = I_\epsilon(Pw_{\epsilon,k})$$

$$+ \epsilon O\left(e^{-(2+\sigma)m(1-r_k)/\epsilon} + \sum_{j=1}^{k} \big(e^{-(1+\sigma)m(\bar{r}_j - r_{j-1})/\epsilon} + e^{-(1+\sigma)m(r_j - \bar{r}_j)/\epsilon}\big) + \epsilon\right),$$

where $\sigma > 0$ *is a small constant.*

Proof. First, we estimate $\|\phi_\epsilon\|_2$.

We have

$$\begin{aligned}
\|\phi_\epsilon\|_2 &\le C\|S_\epsilon(Pw_{\epsilon,k})\|_2 + C\|R_{\mathbf{r}}(\phi_\epsilon)\|_2 \\
&\le C\|S_\epsilon(Pw_{\epsilon,k})\|_2 + C\|\phi_\epsilon^2\|_2 \\
&\le C\|S_\epsilon(Pw_{\epsilon,k})\|_2 + C\|\phi_\epsilon\|_\infty \|\phi_\epsilon\|_2 \\
&\le C\|S_\epsilon(Pw_{\epsilon,k})\|_2 + o(1)\|\phi_\epsilon\|_2,
\end{aligned}$$

where $o(1) = \|\phi_\epsilon\|_\infty \to 0$ as $\epsilon \to 0$. Thus,

$$\|\phi_\epsilon\|_2 \le C\|S_\epsilon(Pw_{\epsilon,k})\|_2. \tag{5.38}$$

We obtain, from (5.37),

$$\begin{aligned}\|S_\epsilon(Pw_{\epsilon,k})\|_2 &\le \|\hat f(\bar w_{\epsilon,k}) - \sum_{j=0}^{k}\hat f(\tilde v_{\epsilon,j})\|_2\\ &+ \left\|\sum_{j=1}^{k}\big(\hat f(\tilde v_{\epsilon,j}) - \hat f(\psi_{\epsilon,j}) - \hat f(\bar\psi_{\epsilon,j})\big)\right\|_2\\ &+ \|f(Pw_{\epsilon,k}) - f(w_{\epsilon,k}) + m^2(w_{\epsilon,k} - Pw_{\epsilon,k})\|_2 + C\epsilon. \qquad (5.39)\end{aligned}$$

It is easy to prove that

$$\begin{aligned}&\left\|\hat f(\bar w_{\epsilon,k}) - \sum_{j=0}^{k}\hat f(\tilde v_{\epsilon,j})\right\|_2^2\\ &\le C\epsilon\sum_{j=1}^{k}\big(e^{-(1+\sigma)m(r_j-\bar r_j)/\epsilon} + e^{-(1+\sigma)m(\bar r_j - r_{j-1})/\epsilon}\big), \qquad (5.40)\end{aligned}$$

and

$$\begin{aligned}&\left\|\sum_{j=1}^{k}\big(\hat f(\tilde v_{\epsilon,j}) - \hat f(\psi_{\epsilon,j}) - \hat f(\bar\psi_{\epsilon,j})\big)\right\|_2^2\\ &\le C\epsilon\sum_{j=1}^{k}\big(e^{-(1+\sigma)m(r_j-\bar r_j)/\epsilon} + e^{-(1+\sigma)m(\bar r_j - r_{j-1})/\epsilon}\big). \qquad (5.41)\end{aligned}$$

Moreover, from

$$f(Pw_{\epsilon,k}) - f(w_{\epsilon,k}) + m^2(w_{\epsilon,k} - Pw_{\epsilon,k}) = \big(m^2 - f'(w_{\epsilon,k})\big)\varphi_\epsilon + O(\varphi_\epsilon^2),$$

we obtain

$$\begin{aligned}&\|f(Pw_{\epsilon,k}) - f(w_{\epsilon,k}) + m^2(w_{\epsilon,k} - Pw_{\epsilon,k})\|_2^2\\ &\le C\int_\Omega \varphi_\epsilon^4 + C\int_\Omega \big(m^2 - f'(w_{\epsilon,k})\big)^2\varphi_\epsilon^2\\ &\le C\epsilon e^{-3m(1-r_k)/\epsilon} + C\int_\Omega \big(m^2 - f'(w_{\epsilon,k})\big)^2\varphi_\epsilon^2. \qquad (5.42)\end{aligned}$$

But

$$\begin{aligned}&\int_\Omega \big(m^2 - f'(w_{\epsilon,k})\big)^2\varphi_\epsilon^2\\ &\le C\int_{r\le r_k}\varphi_\epsilon^2 + C\int_{r\ge r_k}|w_{\epsilon,k} - a|\varphi_\epsilon^2\\ &\le C\epsilon e^{-(2+\sigma)m(1-r_k)/\epsilon}. \qquad (5.43)\end{aligned}$$

Combining (5.39)–(5.43), we obtain

$$\|S_\epsilon(Pw_{\epsilon,k})\|_2^2 \le C\epsilon e^{-(2+\sigma)m(1-r_k)/\epsilon} + C\epsilon\sum_{j=1}^{k}\left(e^{-(1+\sigma)m(r_j-\bar r_j)/\epsilon}+e^{-(1+\sigma)m(\bar r_j-r_{j-1})/\epsilon}\right). \tag{5.44}$$

Using (5.38), we obtain

$$\|\phi_\epsilon\|_2^2 \le C\epsilon e^{-(2+\sigma)m(1-r_k)/\epsilon} + C\epsilon\sum_{j=1}^{k}\left(e^{-(1+\sigma)m(r_j-\bar r_j)/\epsilon}+e^{-(1+\sigma)m(\bar r_j-r_{j-1})/\epsilon}\right). \tag{5.45}$$

Similar to section 4, using (5.44) and (5.45), we have

$$\begin{aligned} &I_\epsilon(Pw_{\epsilon,k}+\phi_\epsilon)\\ &= I_\epsilon(Pw_{\epsilon,k}) + O\Big(\|S_\epsilon(Pw_{\epsilon,k})\|_2\|\phi_\epsilon\|_2 + \|\phi_\epsilon\|_2^2\Big)\\ &= I_\epsilon(Pw_{\epsilon,k})\\ &\quad + \epsilon O\left(\epsilon + e^{-(2+\sigma)m(1-r_k)/\epsilon} + \sum_{j=1}^{k}\left(e^{-(1+\sigma)m(r_j-\bar r_j)/\epsilon}+e^{-(1+\sigma)m(\bar r_j-r_{j-1})/\epsilon}\right)\right). \end{aligned}$$

So we have proved this lemma. □

Proof of Theorem 1.3. We just need to consider the case $b > 0$. For $a < b \le 0$, we let $u_1 = -u$ and $v_1 = -v$. Then u_1 and v_1 will satisfy a similar system with $f_1(t) = (t+a)(t+\frac{a+b}{2})(t+b)$, and $-a > -b \ge 0$. From now on, we always assume that $b > 0$.

Consider

$$\max_{\mathbf{r}_\epsilon \in D_{\epsilon,k}} Q_\epsilon(\mathbf{r}), \tag{5.46}$$

where $Q_\epsilon(\mathbf{r}) = I_\epsilon(Pw_{\epsilon,k}+\phi_\epsilon)$.

Let $\mathbf{r}_\epsilon \in D_{\epsilon,k}$ be a maximum point of (5.46). We will prove that $\mathbf{r}_\epsilon$ is an interior point of $D_{\epsilon,k}$. So it is a critical point of $I_\epsilon(Pw_{\epsilon,k}+\phi_\epsilon)$.

Let $L > 0$ be a large number, such that $mL > 4$. Choose $\mathbf{r}_\epsilon^* \in D_{\epsilon,k}$, such that $r_k^* = 1 - L\epsilon\ln\frac{1}{\epsilon}$, $\bar r_j^* = r_j^* - L\epsilon\ln\frac{1}{\epsilon}$, and $r_{j-1}^* = \bar r_j^* - L\epsilon\ln\frac{1}{\epsilon}$. For this r_ϵ^*, using Proposition 5.2, we have

$$I^*(Pw_{\epsilon,k}) = \epsilon(2k+1)A + \epsilon O\left(\epsilon\ln\frac{1}{\epsilon}\right). \tag{5.47}$$

On the other hand, we have

$$\int_\Omega Pw_{\epsilon,k}(1-\Delta)^{-1}Pw_{\epsilon,k} = \int_\Omega b(1-\Delta)^{-1}b + O\left(\epsilon\ln\frac{1}{\epsilon}\right). \tag{5.48}$$

Combining (5.47) and (5.48), we obtain

$$Q_\epsilon(\mathbf{r}_\epsilon^*) = \epsilon(2k+1)A + \frac{1}{2}\gamma\epsilon\int_\Omega b(1-\Delta)^{-1}b + \epsilon O\left(\epsilon\ln\frac{1}{\epsilon}\right). \tag{5.49}$$

We have, from $Q_\epsilon(\mathbf{r}_\epsilon) \geq Q_\epsilon(\mathbf{r}_\epsilon^*)$,

$$\begin{aligned}
&\epsilon A r_{\epsilon,0}^{n-1} + \epsilon A\sum_{j=1}^{k}(\bar{r}_{\epsilon,j}^{n-1} + r_{\epsilon,j}^{n-1}) + \frac{1}{2}\gamma\epsilon\int_\Omega Pw_{\epsilon,k}(1-\Delta)^{-1}Pw_{\epsilon,k}\\
&-(B+o(1))\epsilon\sum_{j=1}^{k}\left(e^{-m(\bar{r}_{\epsilon,j}-r_{\epsilon,j-1})/\epsilon} + e^{-m(r_{\epsilon,j}-\bar{r}_{\epsilon,j})/\epsilon}\right)\\
&-(B_\epsilon+o(1))\omega_{n-1}\epsilon e^{-2m(1-r_{\epsilon,k})/\epsilon} + O(\epsilon^2)\\
&\geq \epsilon(2k+1)A + \frac{1}{2}\gamma\epsilon\int_\Omega b(1-\Delta)^{-1}b + \epsilon O\left(\epsilon\ln\frac{1}{\epsilon}\right).
\end{aligned} \tag{5.50}$$

Since $|b - Pw_{\epsilon,k}| \leq C\epsilon^2$ if $r \leq 1 - 2M\epsilon\ln\frac{1}{\epsilon}$, it is easy to check that

$$\int_\Omega |(1-\Delta)^{-1}(b-Pw_{\epsilon,k})|^2 \leq C\int_\Omega |b-Pw_{\epsilon,k}|^2 \leq C\epsilon\ln\frac{1}{\epsilon}.$$

As a result,

$$\begin{aligned}
&\int_\Omega b(1-\Delta)^{-1}b - \int_\Omega Pw_{\epsilon,k}(1-\Delta)^{-1}Pw_{\epsilon,k}\\
&= -2\int_\Omega (Pw_{\epsilon,k}-b)(1-\Delta)^{-1}b + \int_\Omega (Pw_{\epsilon,k}-b)(1-\Delta)^{-1}(Pw_{\epsilon,k}-b)\\
&= 2\int_\Omega (b-w_{\epsilon,k})(1-\Delta)^{-1}b + 2\int_\Omega \varphi_\epsilon(1-\Delta)^{-1}b + O\left(\epsilon^2\ln^2\frac{1}{\epsilon}\right).
\end{aligned} \tag{5.51}$$

But

$$\left|\int_\Omega \varphi_\epsilon(1-\Delta)^{-1}b\right| \leq C\int_\Omega |\varphi_\epsilon| \leq C\epsilon. \tag{5.52}$$

Combining (5.51) and (5.52), we obtain

$$\begin{aligned}
&\int_\Omega b(1-\Delta)^{-1}b - \int_\Omega Pw_{\epsilon,k}(1-\Delta)^{-1}Pw_{\epsilon,k}\\
&= 2\int_\Omega (b-w_{\epsilon,k})(1-\Delta)^{-1}b + O(\epsilon).
\end{aligned} \tag{5.53}$$

So, it follows from (5.50) and (5.53) that

$$\begin{aligned}
&\epsilon A(1-r_{\epsilon,0}^{n-1})+\epsilon A\sum_{j=1}^{k}(1-\bar{r}_{\epsilon,j}^{n-1}+1-r_{\epsilon,j}^{n-1})\\
&+\frac{1}{2}\gamma\epsilon\int_{\Omega}(b-w_{\epsilon,k})(1-\Delta)^{-1}b\\
&+(B+o(1))\epsilon\sum_{j=1}^{k}\big(e^{-m(\bar{r}_{\epsilon,j}-r_{\epsilon,j-1})/\epsilon}+e^{-m(r_{\epsilon,j}-\bar{r}_{\epsilon,j})/\epsilon}\big)\\
&+(B_\epsilon+o(1))\omega_{n-1}\epsilon e^{-2m(1-r_{\epsilon,k})/\epsilon}\\
&\le C\epsilon^2\ln\frac{1}{\epsilon}. \qquad (5.54)
\end{aligned}$$

Since all the terms in the left-hand side of (5.54) are positive, we obtain

$$1-r_{\epsilon,0}^{n-1}\le C\epsilon\ln\frac{1}{\epsilon},$$

and

$$e^{-m(\bar{r}_{\epsilon,j}-r_{\epsilon,j-1})/\epsilon},\quad e^{-m(r_{\epsilon,j}-\bar{r}_{\epsilon,j})/\epsilon},\quad e^{-2m(1-r_{\epsilon,k})/\epsilon}\le C\epsilon\ln\frac{1}{\epsilon}.$$

So $|\bar{r}_{\epsilon,j}-r_{\epsilon,j-1}|\ge c'\epsilon\ln\frac{1}{\epsilon}$, $|r_{\epsilon,j}-\bar{r}_{\epsilon,j}|\ge c'\epsilon\ln\frac{1}{\epsilon}$, $1-r_{\epsilon,0}\le C\epsilon\ln\frac{1}{\epsilon}$, and $1-r_{\epsilon,k}\ge c'\epsilon\ln\frac{1}{\epsilon}$. This shows that $\mathbf{r}_\epsilon$ is an interior point of $D_{\epsilon,k}$ if $M>0$ is large and $\alpha>0$ is small. □

Remark 5.8. In the case $b>0$, if $\gamma=0$, or $\gamma>0$ and $a\le 0$, then we can use the above techniques to show that (1.1) has a solution, which is close to $Pw_{\epsilon,k}^*$, where

$$w_{\epsilon,k}^*=\sum_{j=1}^{k-1}(v_{\epsilon,j}+\bar{v}_{\epsilon,j}-a-b)+\bar{v}_{\epsilon,k}.$$

In this case, similar to (5.54), we have

$$\begin{aligned}
&\epsilon A(1-\bar{r}_{\epsilon,k}^{n-1})+\epsilon A\sum_{j=1}^{k-1}(1-\bar{r}_{\epsilon,j}^{n-1}+1-r_{\epsilon,j}^{n-1})\\
&+\frac{1}{2}\gamma\epsilon\int_{\Omega}(a-w_{\epsilon,k}^*)(1-\Delta)^{-1}a\\
&+(B+o(1))\epsilon\sum_{j=1}^{k-1}\big(e^{-m(\bar{r}_{\epsilon,j}-r_{\epsilon,j-1})/\epsilon}+e^{-m(r_{\epsilon,j}-\bar{r}_{\epsilon,j})/\epsilon}\big)\\
&+(B_\epsilon+o(1))\omega_{n-1}\epsilon e^{-2m(1-\bar{r}_{\epsilon,k})/\epsilon}\\
&\le C\epsilon^2\ln\frac{1}{\epsilon}. \qquad (5.55)
\end{aligned}$$

We know that $a - w^*_{\epsilon,k}$ is always nonpositive. So, if $a \le 0$, the term $\int_\Omega (a - w^*_{\epsilon,k})(1 - \Delta)^{-1} a$ is nonnegative. Thus, if $\gamma = 0$, or $\gamma > 0$ and $a \le 0$, the left-hand side of (5.55) is always nonnegative. As a result, we can use (5.55) to deduce that $\mathbf{r}_\epsilon$ is an interior point of $D_{\epsilon,k}$.

REFERENCES

[1] X. Chen and Y. Oshita, *Periodicity and uniqueness of global minimizers of an energy functional containing a long-range interaction*, SIAM J. Math. Anal., 37 (2006), pp. 1299–1332.

[2] E. N. Dancer and S. Yan, *Multi-layer solutions for an elliptic problem*, J. Differential Equations, 194 (2003), pp. 382–405.

[3] E. N. Dancer and S. Yan, *Peak solutions for an elliptic system of FitzHugh-Nagumo type*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 2 (2003), pp. 679–709.

[4] E. N. Dancer and S. Yan, *A minimization problem associated with elliptic systems of FitzHugh-Nagumo type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 21 (2004), pp. 237–257.

[5] E. N. Dancer and S. Yan, *Peak solutions for the Neumann problem of an elliptic system of FitzHugh-Nagumo type*, Proc. London Math. Soc., 90 (2005), pp. 209–244.

[6] E. N. Dancer and S. Yan, *Multipeak solutions for an elliptic system of FitzHugh-Nagumo type*, Math. Ann., 335 (2006), pp. 527–569.

[7] E. N. Dancer and S. Yan, *Solutions with interior and boundary peaks for the Neumann problem of an elliptic system of FitzHugh-Nagumo type*, Indiana Univ. Math. J., 55 (2006), pp. 217–258.

[8] E. De Giorgi, *Sulla convergenza di alcune successioni d'integrali del tipo dell'area*, Rend. Mat. (6), 8 (1975), pp. 277–294.

[9] P. C. Fife, *Boundary and interior transition layer phenomena for pairs of second order differential equations*, J. Math. Anal., 54 (1976), pp. 497–521.

[10] R. A. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.

[11] M. Itô, *A remark on singular perturbation methods*, Hiroshima Math. J., 14 (1985), pp. 619–629.

[12] G. A. Klaasen and E. Mitidieri, *Standing wave solutions for a system derived from the FitzHugh–Nagumo equations for nerve conduction*, SIAM J. Math. Anal., 17 (1986), pp. 74–83.

[13] R. Kohn and P. Sternberg, *Local minimisers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.

[14] M. Mimura, M. Tabata, and Y. Hosono, *Multiple solutions of two-point boundary value problems of Neumann type with a small parameter*, SIAM J. Math. Anal., 11 (1980), pp. 613–631.

[15] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Ration. Mech. Anal., 98 (1987), pp. 123–142.

[16] L. Modica and S. Mortola, *Un esempio di Γ^--convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[17] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.

[18] J. Nagumo, S. Arimoto, and Y. Yoshikawa, *An active pulse transmission line simulating nerve axton*, Proc. IRE, 50 (1962), p. 2061.

[19] Y. Nishiura, *Coexistence of infinitely many stable solutions to reaction-diffusion system in the singular limit*, in Dynamics Reported: Expositions in Dynamical Systems, vol. 3, C. R. K. T. Jones, U. Kirchgraber, and H. O. Walther, eds., Springer-Verlag, New York, 1994, pp. 25–103.

[20] Y. Nishiura and H. Fujii, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 1726–1770.

[21] C. Reinecke and G. Sweers, *Solutions with internal jump for an autonomous elliptic system of FitzHugh-Nagumo type*, Math. Nachr., 251 (2003), pp. 64–87.

[22] X. Ren and J. Wei, *Concentrically layered energy equilibria of the di-block copolymer problem*, European J. Appl. Math., 13 (2002), pp. 479–496.

[23] X. Ren and J. Wei, *On energy minimizers of the di-block copolymer problem*, Interfaces Free Bound., 5 (2003), pp. 193–238.

[24] X. Ren and J. Wei, *Nucleation in the FitzHugh-Nagumo system: Interface-spike solutions*, J. Differential Equations, 209 (2005), pp. 266–301.

SIAM J. MATH. ANAL.
Vol. 38, No. 6, pp. 2042–2056

WHITE NOISE HYPOTHESIS FOR UNIFORM QUANTIZATION ERRORS*

DAVID JIMENEZ†, LONG WANG‡, AND YANG WANG†

Abstract. The white noise hypothesis (WNH) assumes that in the uniform pulse code modulation (PCM) quantization scheme the errors in individual channels behave like white noise; i.e., they are independent and identically distributed random variables. The WNH is key to estimating the mean square quantization error (MSE). But is the WNH valid? In this paper we take a close look at the WNH. We show that in a redundant system the errors from individual channels can never be independent. Thus to an extent the WNH is invalid. Our numerical experiments also indicate that with coarse quantization the WNH is far from valid. However, as the main result of this paper we show that with fine quantizations the WNH is essentially valid in that the errors from individual channels become asymptotically *pairwise* independent, each uniformly distributed in $[-\Delta/2, \Delta/2)$, where Δ denotes the stepsize of the quantization.

Key words. frames, tight frame, pulse code modulation, quantization, white noise hypothesis

AMS subject classification. 42C15

DOI. 10.1137/050636929

1. Introduction.

In processing, transmitting, and storing analogue signals it is often necessary to make atomic decompositions of the signal using a given set of *atoms*, or *dictionary* $\{\mathbf{v}_j\}$. In this approach, a signal $\mathbf{x}$ is represented as a linear combination of $\{\mathbf{v}_j\}$,

$$\mathbf{x} = \sum_j c_j \mathbf{v}_j.$$

In practice $\{\mathbf{v}_j\}$ is a finite set. Furthermore, for the purpose of error correction, recovery from data erasures or robustness, redundancy is built into $\{\mathbf{v}_j\}$; i.e., more elements than needed are in $\{\mathbf{v}_j\}$. Instead of a true basis, $\{\mathbf{v}_j\}$ is chosen to be a frame. Since $\{\mathbf{v}_j\}$ is a finite set, we may without loss of generality assume $\{\mathbf{v}_j\}_{j=1}^N$ are vectors in $\mathbb{R}^d$ with $N \geq d$.

Let $F = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$ be the $d \times N$ matrix whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_N$. We say $\{\mathbf{v}_j\}_{j=1}^N$ is a *frame* if F has rank d. Let $\lambda_{\max} \geq \lambda_{\min} > 0$ be the maximal and minimal eigenvalues of FF^T, respectively. It is easily checked that

$$\lambda_{\min}\|\mathbf{x}\|^2 \leq \sum_{j=1}^N |\mathbf{x} \cdot \mathbf{v}_j|^2 \leq \lambda_{\max}\|\mathbf{x}\|^2. \tag{1.1}$$

$\lambda_{\max}$ and $\lambda_{\min}$ are called the *upper and lower frame bounds* for the frame, respectively. If $\lambda_{\max} = \lambda_{\min} = \lambda$, in which case $FF^T = \lambda I_d$, we call $\{\mathbf{v}_j\}_{j=1}^N$ a *tight frame* with frame constant λ. Note that any signal $\mathbf{x} \in \mathbb{R}^d$ can be easily reconstructed using the

*Received by the editors July 26, 2005; accepted for publication (in revised form) October 31, 2006; published electronically April 6, 2007.
http://www.siam.org/journals/sima/38-6/63692.html

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (djimenez@math.gatech.edu, wang@math.gatech.edu). The third author is supported in part by the National Science Foundation, grant DMS-0456538.

‡Department of Mathematics, Southern Polytechnic State University, Marietta, GA 30060 (lwang@spsu.edu).

data $\{\mathbf{x}\cdot\mathbf{v}_j\}_{j=1}^N$. Set $\mathbf{y}=[\mathbf{x}\cdot\mathbf{v}_1,\mathbf{x}\cdot\mathbf{v}_2,\ldots,\mathbf{x}\cdot\mathbf{v}_N]^T$. Then $\mathbf{y}=F^T\mathbf{x}$ and

$$(FF^T)^{-1}F\mathbf{y}=(FF^T)^{-1}FF^T\mathbf{x}=\mathbf{x}.$$

Let $G=(FF^T)^{-1}F=[\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_N]$. The set of columns $\{\mathbf{u}_j\}_{j=1}^N$ of G is called the *canonical dual frame* of the frame $\{\mathbf{v}_j\}_{j=1}^N$. We have the reconstruction

$$\mathbf{x}=\sum_{j=1}^N(\mathbf{x}\cdot\mathbf{v}_j)\,\mathbf{u}_j. \tag{1.2}$$

If $\{\mathbf{v}_j\}_{j=1}^N$ is a tight frame with frame constant λ, then $G=\lambda^{-1}F$, and we have the reconstruction

$$\mathbf{x}=\frac{1}{\lambda}\sum_{j=1}^N(\mathbf{x}\cdot\mathbf{v}_j)\,\mathbf{v}_j. \tag{1.3}$$

In digital applications, quantizations will have to be performed. The simplest scheme is the pulse code modulation (PCM) quantization scheme, in which the coefficients $\{\mathbf{x}\cdot\mathbf{v}_j\}_{j=1}^N$ are quantized. In this paper we consider exclusively *uniform quantizations*. Let $\mathcal{A}=\Delta\mathbb{Z}$, where $\Delta>0$ is the quantization step. With uniform quantization a real value t is replaced with the value in $\mathcal{A}$ that is the closest to t. So, in our setting, t is replaced with $Q_\Delta(t)$ given by

$$Q_\Delta(t):=\left\lfloor\frac{t}{\Delta}+\frac{1}{2}\right\rfloor\Delta.$$

Thus, given a frame $\{\mathbf{v}_j\}_{j=1}^N$ and its canonical dual frame $\{\mathbf{u}_j\}_{j=1}^N$, instead of using the data $\{\mathbf{x}\cdot\mathbf{v}_j\}_{j=1}^N$ and (1.2) to obtain a perfect reconstruction, we use the data $\{Q_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\}_{j=1}^N$ and obtain an imperfect reconstruction

$$\tilde{\mathbf{x}}=\sum_{j=1}^N Q_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\mathbf{u}_j. \tag{1.4}$$

This raises the following question: How good is the reconstruction? This question has been studied in terms of both the worst case error and the mean square error (MSE); see, e.g., [13]. Note that the error from the reconstruction is

$$\mathbf{x}-\tilde{\mathbf{x}}=\sum_{j=1}^N\tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\mathbf{u}_j, \tag{1.5}$$

where $\tau_\Delta(t):=t-Q_\Delta(t)=\left(\left\{\frac{t}{\Delta}+\frac{1}{2}\right\}-\frac{1}{2}\right)\Delta$, with $\{\cdot\}$ denoting the fractional part. While an a priori error bound is relatively straightforward to obtain, the MSE, $\mathrm{MSE}:=\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)$, assuming certain probability distribution for $\mathbf{x}$, is much harder. To simplify the problem, the so-called *white noise hypothesis* (WNH), is employed by engineers and mathematicians in this area (see, e.g., [2, 3, 13]). The WNH asserts the following:

- Each $\tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)$ is uniformly distributed in $[-\Delta/2,\Delta/2)$; hence it has mean 0 and variance $\Delta^2/12$.
- $\{\tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\}_{j=1}^N$ are independent random variables.

With the WNH it is an easy derivation, which we furnish in the next section, that the MSE is given by

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)=\frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1}=\frac{\Delta^2}{12}\sum_{j=1}^{N}\|\mathbf{u}_j\|^2, \tag{1.6}$$

where $\{\lambda_j\}$ are the eigenvalues of FF^T.

Note that using (1.6) the MSE for quantization decreases by a factor of 4 if we decrease Δ by a factor of 2. It amounts to an increase in signal to noise ratio of approximately 6dB ($10\log_{10}4\approx 6$). This is often referred to as the *6dB-per-bit-rule*.

The WNH is often called *Bennett's white noise assumption* [2, 3]. Bennett studied quantization error (distortion) in his fundamental paper [4] in the scalar setting. He demonstrated that under the assumption that the scalar random variable has a smooth density, the quantization error behaves like uniformly distributed "random noise" when Δ is small, resulting in the MSE of approximately $\Delta^2/12$. Bennett also studied quantization errors in the nonuniform quantization setting, which can often be reduced to the uniform setting by the use of companders. The current interest in the WNH stems from the study of vector quantization, in which several correlated signals are quantized simultaneously such as in our setting. A vast literature on vector quantization and on vector quantization errors exists, and for an excellent and comprehensive survey on vector quantization see Gray and Neuhoff [14]. A weaker form of the WNH, which states that the error components are approximately uncorrelated in the high resolution setting, i.e., when Δ is small, is often found in engineering literature without rigorous proof (see [11] and the discussion in [22]). A rigorous proof of this weaker form of the WNH was first given in Viswanathan and Zamir [22]. More precisely, they proved that if two random variables X, Y have a joint density function, then $\frac{1}{\Delta^2}\mathcal{E}\left(\tau_\Delta(X)\tau_\Delta(Y)\right)\longrightarrow 0$ as $\Delta\to 0$. Viswanathan and Zamir also proved similar results in the nonuniform quantization setting, under much stronger assumptions.

It should be pointed out that much of the advantage of vector quantization comes from the fact that the quantizations are *not* necessarily performed independently on each channel. As a result there are many interesting and challenging mathematical problems in nonuniform vector quantization. While the focus of this paper is on uniform quantization, we hope it will be a very first step in resolving the problem in the more general setting.

The objective of this paper is a rather modest one. Given the vast literature on quantization errors and some of the general confusion regarding the WNH, this paper aims to provide a complete analysis and rigorous mathematical theorems on the behavior of quantization errors. These results are by no means difficult, and they are also rather intuitive. Nevertheless we feel there is a need to have them written. If nothing else we hope this paper will serve to clarify the WNH in the uniform quantization setting. As a very simple result we show that, under the assumption that the distribution of $\mathbf{x}$ has a density (absolutely continuous), the components of the quantization errors $\left\{\tau_\Delta\big(\mathbf{x}\cdot\mathbf{v}_j\big)\right\}_{j=1}^N$ can *never* be independent if $N>d$. However, we show that asymptotically the WNH is almost valid by proving stronger and more general results than those in [22]. More precisely, we prove that if a set of vectors $\{\mathbf{u}_1,\mathbf{u}_2,\dots,\mathbf{u}_k\}$ is linearly independent, then the normalized quantization errors $\left\{\frac{1}{\Delta}\tau_\Delta\left(\mathbf{x}\cdot\mathbf{u}_j\right)\right\}_{j=1}^k$ converge in distribution to independent and uniformly distributed random variables as $\Delta\to 0^+$. Applying this to the frame setting, we show that if the vectors $\{\mathbf{v}_j\}_{j=1}^N$ are pairwise linearly independent, then $\{\tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\}_{j=1}^N$ becomes asymptotically *pairwise*

independent and thus *pairwise uncorrelated*, and each $\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)$ becomes asymptotically uniformly distributed on $[-\Delta/2,\Delta/2]$. These slightly weaker assumptions are sufficient to lead to the MSE given by (1.6) asymptotically. Furthermore, we also characterize completely the asymptotic behavior of the MSE if some $\mathbf{v}_j$'s are parallel. These and other results are stated and proved in subsequent sections.

2. A priori error bound and MSE under the WNH. In this section we derive an a priori error bound and a formula for the MSE under the WNH. These results are not new. We include them for self-containment. We use the following settings throughout this section: Let $\{\mathbf{v}_j\}_{j=1}^N$ be a frame in $\mathbb{R}^d$ with corresponding frame matrix $F=[\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_N]$. The eigenvalues of FF^T are $\lambda_{\max}=\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_d=\lambda_{\min}>0$. Let $\{\mathbf{u}_j\}_{j=1}^N$ be the canonical dual frame with corresponding matrix $G=(FF^T)^{-1}F$. For any $\mathbf{x}=\sum_{j=1}^N(\mathbf{x}\cdot\mathbf{v}_j)\,\mathbf{u}_j$, using the quantization alphabet $\mathcal{A}=\Delta\mathbb{Z}$, we have the PCM quantized reconstruction

$$\tilde{\mathbf{x}}=\sum_{j=1}^N Q_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\,\mathbf{u}_j.$$

PROPOSITION 2.1. *For any $\mathbf{x}\in\mathbb{R}^d$ we have*

$$\|\mathbf{x}-\tilde{\mathbf{x}}\|\le\frac{1}{2}\sqrt{\frac{N}{\lambda_{\min}}}\Delta. \tag{2.1}$$

If in addition $\{\mathbf{v}_j\}_{j=1}^N$ is a tight frame with frame constant λ, then

$$\|\mathbf{x}-\tilde{\mathbf{x}}\|\le\frac{1}{2}\sqrt{\frac{N}{\lambda}}\Delta. \tag{2.2}$$

Proof. We have $\mathbf{x}-\tilde{\mathbf{x}}=\sum_{j=1}^N\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\,\mathbf{u}_j=G\mathbf{y}$, where $\mathbf{y}=[\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_1),\dots,\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_N)]^T$. Thus $\|\mathbf{x}-\tilde{\mathbf{x}}\|^2=\mathbf{y}^TG^TG\mathbf{y}\le\rho\left(G^TG\right)\|\mathbf{y}\|^2$, where $\rho(\cdot)$ denotes the spectral radius. Now $\rho(G^TG)=\rho(GG^T)=\rho((FF^T)^{-1})=\lambda_{\min}^{-1}$. Observe that $|\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)|\le\Delta/2$. Thus $\|\mathbf{y}\|^2\le N(\Delta/2)^2$. This yields an a priori error bound (2.1). The bound (2.2) is an immediate corollary. ∎

PROPOSITION 2.2. *Under the WNH, the MSE is*

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)=\frac{\Delta^2}{12}\sum_{j=1}^d\lambda_j^{-1}=\frac{\Delta^2}{12}\sum_{j=1}^N\|\mathbf{u}_j\|^2. \tag{2.3}$$

In particular, if $\{\mathbf{v}_j\}_{j=1}^N$ is a tight frame with frame constant λ, then

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)=\frac{d}{12\lambda}\Delta^2. \tag{2.4}$$

Proof. Denote $G^TG=[b_{ij}]_{i,j=1}^N$ and again let $\mathbf{y}=[\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_1),\dots,\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_N)]^T$. Note that with the WNH, $\mathcal{E}(y_iy_j)=\mathcal{E}(\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_i)\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j))=(\Delta^2/12)\delta_{ij}$. Now $\mathbf{x}-\tilde{\mathbf{x}}=G\mathbf{y}$, and hence

$$\mathcal{E}\left(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\right)=\mathcal{E}\left(\mathbf{y}^TG^TG\mathbf{y}\right)=\sum_{i,j=1}^N b_{ij}\mathcal{E}(y_iy_j)=\sum_{i=1}^N b_{ii}\frac{\Delta^2}{12}=\frac{\Delta^2}{12}\mathrm{tr}(G^TG).$$

Finally, $\operatorname{tr}(G^TG) = \sum_{j=1}^N \|\mathbf{u}_j\|^2$, and $\operatorname{tr}(G^TG) = \operatorname{tr}(GG^T) = \operatorname{tr}((FF^T)^{-1}) = \sum_{j=1}^d \lambda_j^{-1}$. ∎

Remark. The MSE formulae (2.3)–(2.4) still hold if the independence of $\{\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\}_{j=1}^N$ in the WNH is replaced with the weaker condition that $\{\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\}_{j=1}^N$ are uncorrelated.

3. A closer look at the WNH. The WNH asserts that the error components $\{\tau_\Delta(\mathbf{x}\cdot\mathbf{v}_j)\}_{j=1}^N$ are independent and identically distributed random variables. Intuitively this cannot be true if $N > d$. This is indeed the case in general. The following is a simple result.

THEOREM 3.1. *Let $\mathbf{X} \in \mathbb{R}^d$ be an absolutely continuous random vector. Let $\{\mathbf{v}_j\}_{j=1}^N$ be nonzero vectors in $\mathbb{R}^d$ with $N > d$. Then the random variables $\{\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_j)\}_{j=1}^N$ are not independent.*

Proof. Let F be the frame matrix for the frame $\{\mathbf{v}_j\}$. Then $\dim(\operatorname{range}(F^T)) \le d$, and therefore $\mathcal{L}(\operatorname{range}(F^T)) = 0$, where $\mathcal{L}$ is the Lebesgue measure on $\mathbb{R}^N$. Let $\mathbf{Y} = [Y_1, \dots, Y_N]^T := F^T\mathbf{X}$, and let $\tilde{\mathbf{Y}} = [Q_\Delta(Y_1), \dots, Q_\Delta(Y_N)]^T$ be the quantized $\mathbf{Y}$. Denote $\mathbf{Z} = \mathbf{Y} - \tilde{\mathbf{Y}} = [Z_1, \dots, Z_N]^T$. Note that $Y_j = \mathbf{v}_j \cdot \mathbf{X}$, so each Y_j is absolutely continuous, and therefore so is each Z_j. If $\{Z_j\}$ are independent, then $\mathbf{Z}$ must be absolutely continuous.

Now, set $\Omega := \operatorname{range}(F^T) + \Delta\mathbb{Z}^N$. Then $\mathcal{L}(\Omega) = 0$ because $\Delta\mathbb{Z}^N$ is a countable set. However, $\mathbf{Z}$ takes values in Ω, so $P(\mathbf{Z} \in \Omega) = 1$. This contradicts the absolute continuity of $\mathbf{Z}$. ∎

Remark. Actually for Theorem 3.1 to hold we need only to assume that $\mathbf{X}$ has an absolutely continuous component, i.e., $\mathbf{X} = \mathbf{X}_c + \mathbf{X}_s$, where $\mathbf{X}_c \neq 0$ is absolutely continuous and $\mathbf{X}_s$ is singular. However, the theorem can fail without the absolute continuity condition, even if each component of $\mathbf{X}$ may be absolutely continuous. The simplest example is to take $\mathbf{X} = [X, -X]^T$, where X is any random variable and $\mathbf{v}_1 = [1, 1]^T$ and $\mathbf{v}_2 = [1, -1]^T$.

Even when $N = d$ the WNH holds only under rather strict conditions. The following is another simple result.

PROPOSITION 3.2. *Let $\mathbf{X} = [X_1, \dots, X_m]^T$ be a random vector in $\mathbb{R}^m$ whose distribution has density function $g(x_1, \dots, x_m)$.*

(1) *The error components $\{\tau_\Delta(X_j)\}_{j=1}^m$ are independent if and only if there exist complex numbers $\{\beta_j(n) : 1 \le j \le m, n \in \mathbb{Z}\}$ such that*

$$\widehat{g}\left(\frac{a_1}{\Delta}, \dots, \frac{a_m}{\Delta}\right) = \beta_1(a_1)\cdots\beta_m(a_m) \tag{3.1}$$

for all $[a_1, \dots, a_m]^T \in \mathbb{Z}^m$.

(2) *Let $h_j(t)$ be the marginal density of X_j. Then $\{\tau_\Delta(X_j)\}_{j=1}^m$ are identically distributed if and only if*

$$\sum_{n\in\mathbb{Z}} h_j(t - n\Delta) = H(t) \quad \text{a.e.}$$

for some $H(t)$ independent of j. They are uniformly distributed on $[-\Delta/2, \Delta/2]$ if and only if $H(t) = 1/\Delta$ a.e.

Proof. To prove (1) denote $\mathcal{I}_\Delta = [-\Delta/2, \Delta/2]$ and $\mathbf{Y} = [\tau_\Delta(X_1), \dots, \tau_\Delta(X_m)]^T$. Observe that $\mathbf{Y}$ has a density

$$G(\mathbf{y}) := \sum_{\mathbf{a}\in\mathbb{Z}^m} g(\mathbf{y} - \Delta\mathbf{a}) \tag{3.2}$$

for $\mathbf{y} \in \mathcal{I}_\Delta^m$. The density $G(\mathbf{y})$ is periodic with period Δ, and it is well known that its Fourier series is given by $G(\mathbf{y}) = \sum_{\mathbf{a}\in\mathbb{Z}^m} c_{\mathbf{a}} e^{2i\pi \frac{\mathbf{a}}{\Delta}\cdot\mathbf{y}}$, where $c_{\mathbf{a}} = \widehat{g}\left(\frac{\mathbf{a}}{\Delta}\right)$. But $\{Y_j\}_{j=1}^m$ are independent if and only if on $\mathcal{I}_\Delta^m$ we have $g(y_1,\dots,y_m) = g_1(y_1)\cdots g_m(y_m)$. This happens if and only if

$$\widehat{g}\left(\frac{a_1}{\Delta},\frac{a_2}{\Delta},\dots,\frac{a_m}{\Delta}\right) = h_1\left(\frac{a_1}{\Delta}\right) h_2\left(\frac{a_2}{\Delta}\right)\cdots h_m\left(\frac{a_m}{\Delta}\right)$$

for all $\mathbf{a} = [a_1,\dots,a_m]^T \in \mathbb{Z}^m$, with $h_j(\xi) = \widehat{g}_i(\xi)$. This part of the theorem is proved by setting $\beta_j(n) = h_j(n)$.

The proof of (2) follows directly from the fact that the density of $\tau_\Delta(X_j)$ is $\sum_{n\in\mathbb{Z}} h_j(t-\Delta n)$ for $t \in \mathcal{I}_\Delta$. □

Proposition 3.2 puts strong constraints on the distribution of $\mathbf{x}$ for the WNH to hold. Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with joint density $f(\mathbf{x})$. Let $\{\mathbf{v}_j\}_{j=1}^d$ be linearly independent, and let $\mathbf{Y} = [\mathbf{X}\cdot\mathbf{v}_1, \mathbf{X}\cdot\mathbf{v}_2,\dots,\mathbf{X}\cdot\mathbf{v}_d]^T$. Then the joint density of $\mathbf{Y}$ is $g(\mathbf{y}) = |\det(F)|^{-1} f\left((F^T)^{-1}\mathbf{y}\right)$, where $F = [\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_d]$. Thus, both the independence and the identical distribution assumptions in the WNH, even for $N = d$, will be false unless very exact conditions are met. For instance, if we take $\mathbf{X}$ to be Gaussian and F to be unitary, then the independence property is satisfied only when F diagonalizes the covariance matrix of $\mathbf{X}$.

COROLLARY 3.3. *Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with joint density $f(\mathbf{x})$ and $\{\mathbf{v}_j\}_{j=1}^d$ be linearly independent vectors in $\mathbb{R}^d$. Let $\mathbf{Y} = F^T\mathbf{X} = [\mathbf{X}\cdot\mathbf{v}_1,\dots,\mathbf{X}\cdot\mathbf{v}_N]^T$ and $g(\mathbf{y}) = |\det(F)|^{-1} f\left((F^T)^{-1}\mathbf{y}\right)$, where $F = [\mathbf{v}_1,\dots,\mathbf{v}_d]$.*

(1) *$\{\tau_\Delta(Y_j)\}_{j=1}^d$ are independent random variables if and only if there exist complex numbers $\{\beta_j(n) : 1 \le j \le d, n \in \mathbb{Z}\}$ such that*

$$\widehat{g}\left(\frac{a_1}{\Delta},\dots,\frac{a_d}{\Delta}\right) = \beta_1(a_1)\cdots\beta_d(a_d) \tag{3.3}$$

for all $[a_1,\dots,a_d]^T \in \mathbb{Z}^d$.

(2) *Let $h_j(t) = \int_{\mathbb{R}^{d-1}} g(x_1,\dots,x_{j-1},t,x_{j+1},\dots,x_d)\, dx_1\cdots dx_{j-1}\, dx_{j+1}\dots dx_d$. Then $\{\tau_\Delta(X_j)\}_{j=1}^d$ are identically distributed if and only if $\sum_{n\in\mathbb{Z}} h_j(t - n\Delta) = H(t)$ a.e. for some $H(t)$ independent of j. They are uniformly distributed on $[-\frac{\Delta}{2},\frac{\Delta}{2}]$ if and only if $H(t) = 1/\Delta$ a.e.*

Proof. We have only to observe that $g(\mathbf{y})$ is the density of $\mathbf{Y}$ and that h_j is the marginal density of Y_j. The corollary now follows directly from the theorem. □

From a practical point of view, with coarse quantization the MSE cannot be estimated simply by (1.6). Thus the "6-dB-per-bit" rule may not apply. We shall demonstrate this with numerical results. However, with high resolution quantization the formula (1.6) becomes increasingly accurate. We show this in the next section.

4. Asymptotic behavior of errors: Linear independence case. In many practical applications such as a music CD, fine quantizations with 16 bits or more have been adopted. Although the WNH is not valid in general, with fine quantizations we prove here that a weaker version of the WNH is close to being valid, which yields an asymptotic formula for the PCM quantized MSE. Our result here strengthens an asympototic result in [22].

We again consider the same setup as before. Let $\{\mathbf{v}_j\}_{j=1}^N$ be a frame in $\mathbb{R}^d$ with corresponding frame matrix $F = [\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_N]$. The eigenvalues of FF^T are $\lambda_{\max} = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d = \lambda_{\min} > 0$. Let $\{\mathbf{u}_j\}_{j=1}^N$ be the canonical dual frame with corresponding matrix $G = (FF^T)^{-1}F$. For any $\mathbf{x} \in \mathbb{R}^d$ we have $\mathbf{x} = \sum_{j=1}^N (\mathbf{x}\cdot\mathbf{v}_j)\,\mathbf{u}_j$.

Using the quantization alphabet $\mathcal{A} = \Delta\mathbb{Z}$ we have the PCM reconstruction (1.4). Note that $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}(\Delta)$ as it depends on Δ. With the WNH we obtain the MSE

$$\text{MSE} = \mathcal{E}\left(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2\right) = \frac{\Delta^2}{12}\sum_{j=1}^{N}\lambda_j^{-1}.$$

To study the asymptotic behavior of the error components, we study as $\Delta \to 0^+$ the normalized quantization error

$$\frac{1}{\Delta}(\mathbf{x} - \tilde{\mathbf{x}}) = \sum_{j=1}^{N}\frac{1}{\Delta}\tau_\Delta\left(\mathbf{x}\cdot\mathbf{v}_j\right)\mathbf{u}_j. \tag{4.1}$$

THEOREM 4.1. *Let $\mathbf{X} \in \mathbb{R}^d$ be an absolutely continuous random vector. Let $\mathbf{w}_1, \ldots, \mathbf{w}_m$ be linearly independent vectors in $\mathbb{R}^d$. Then*

$$\left[\frac{1}{\Delta}\tau_\Delta\left(\mathbf{X}\cdot\mathbf{w}_1\right), \ldots, \frac{1}{\Delta}\tau_\Delta\left(\mathbf{X}\cdot\mathbf{w}_m\right)\right]^T$$

converges in distribution as $\Delta \to 0^+$ to a random vector uniformly distributed in $[-1/2, 1/2]^m$.

Proof. Denote $Y_j = \mathbf{X}\cdot\mathbf{w}_j$. Since $\{\mathbf{w}_j\}$ are linearly independent, $\mathbf{Y} = [Y_1, \ldots, Y_m]^T$ is absolutely continuous with some joint density $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^m$. As a consequence of (3.2) one has that the distribution of $\mathbf{Z} = [Z_1, \ldots, Z_m]^T$, where $Z_j = \frac{1}{\Delta}\tau_\Delta(Y_j) = \left\{\frac{Y_j}{\Delta} + \frac{1}{2}\right\} - \frac{1}{2}$, is

$$f_\Delta(\mathbf{x}) := \Delta^m \sum_{\mathbf{a}\in\mathbb{Z}^m} f(\Delta\mathbf{x} - \Delta\mathbf{a}) \tag{4.2}$$

for $\mathbf{x} \in [-1/2, 1/2]^m$. Again denote $\mathcal{I}_1 := [-1/2, 1/2]$. It is easy to see that $\|f_\Delta\|_{L^1(\mathcal{I}_1^m)} \le \|f\|_{L^1(\mathbb{R}^m)}$ for

$$\begin{aligned}
\|f_\Delta\|_{L^1(\mathcal{I}_1^m)} &= \int_{\mathcal{I}_1^m} |f_\Delta(\mathbf{x})|\, d\mathbf{x} \\
&\le \sum_{\mathbf{a}\in\mathbb{Z}^m}\int_{\mathcal{I}_1^m} \Delta^m \left|f(\Delta\mathbf{x} - \Delta\mathbf{a})\right|\, d\mathbf{x} \\
&= \sum_{\mathbf{a}\in\mathbb{Z}^m}\int_{\Delta{\mathcal{I}_1}^m + \Delta\mathbf{a}} \left|f\left(\mathbf{y}\right)\right|\, d\mathbf{y} \\
&= \int_{\mathbb{R}^m} \left|f\left(\mathbf{y}\right)\right|\, d\mathbf{y} \\
&= \|f\|_{L^1(\mathbb{R}^m)}.
\end{aligned}$$

Now, if $\Omega = [a_1, b_1] \times \cdots \times [a_m, b_m]$ and $f(\mathbf{x}) = \mathbf{1}_\Omega(\mathbf{x})$, then for $\mathbf{x} \in \mathcal{I}_1^m$ observe that $f_\Delta(\mathbf{x}) = \Delta^m K_\Delta$, where $K_\Delta(\mathbf{x}) = \#\{\mathbf{a} \in \mathbb{Z}^m : \Delta\mathbf{x} + \Delta\mathbf{a} \in \Omega\}$. Obviously, $K_\Delta(\mathbf{x}) = s/\Delta^m + O(\Delta^{-m+1})$, where $s = \mathcal{L}(\Omega)$ is the Lebesgue measure of Ω. Then $f_\Delta \to s\mathbf{1}_{\mathcal{I}_1^m}$ in $L^1(\mathcal{I}_1^m)$ as $\Delta \to 0^+$.

We return to the case when $f(\mathbf{x})$ is the density of $\mathbf{Y}$. For any $\varepsilon > 0$ it is possible to choose a $g(\mathbf{x}) \in L^1(\mathbb{R}^m)$ such that $\|f - g\|_{L^1} < \frac{\varepsilon}{3}$, and furthermore, $g(\mathbf{x}) = \sum_{j=1}^{N} c_j \mathbf{1}_{E_j}(\mathbf{x})$ is a simple function where $c_j \in \mathbb{R}$ and each E_j is a product

of finite intervals. Observe that $\int_{\mathbb{R}^m} g = \sum_{j=1}^N c_j \mathcal{L}(E_j)$. Since $(\mathbf{1}_{E_j})_\Delta \to \mathcal{L}(E_j)\mathbf{1}_{\mathcal{I}_1^m}$ in L^1 we have $g_\Delta \to \left(\int_{\mathbb{R}^m} g\right)\mathbf{1}_{\mathcal{I}_1^m}$ as $\Delta \to 0$. Hence there exists a $\delta > 0$ such that $\|g_\Delta - (\int_{\mathbb{R}^m} g)\mathbf{1}_{\mathcal{I}_1^m}\|_{L^1} < \varepsilon/3$ whenever $\Delta < \delta$. Now, for $\Delta < \delta$,

$$\begin{aligned}\left\|f_\Delta - \mathbf{1}_{\mathcal{I}_1^m}\right\|_{L^1(\mathcal{I}_1^m)} &= \|f_\Delta - g_\Delta\|_{L^1(\mathcal{I}_1^m)} + \left\|g_\Delta - \left(\int_{\mathbb{R}^m} g\right)\mathbf{1}_{\mathcal{I}_1^m}\right\|_{L^1(\mathcal{I}_1^m)} \\ &\quad + \left|1 - \left(\int_{\mathbb{R}^m} g\right)\right| \|\mathbf{1}_{\mathcal{I}_1^m}\|_{L^1(\mathcal{I}_1^m)} \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \left|1 - \left(\int_{\mathbb{R}^m} g\right)\right| \\ &= \frac{2\varepsilon}{3} + \left|\left(\int_{\mathbb{R}^m} f\right) - \left(\int_{\mathbb{R}^m} g\right)\right| \\ &< \varepsilon. \qquad \square\end{aligned}$$

Remark. We in fact proved a stronger result, namely, that the densities converge in L^1. Applying the above theorem to the MSE, if $\{\mathbf{v}_j\}_{j=1}^N$ are pairwise linearly independent, then the error components $\{\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_j)\}_{j=1}^N$ become asymptotically pairwise independent and each is uniformly distributed in $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$.

COROLLARY 4.2. *Let $\mathbf{X} \in \mathbb{R}^d$ be an absolutely continuous random vector. If $\{\mathbf{v}_j\}_{j=1}^N$ are pairwise linearly independent, then as $\Delta \to 0^+$ we have*

$$\mathcal{E}\left(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2\right) = \frac{\Delta^2}{12}\sum_{j=1}^d \lambda_j^{-1} + o(\Delta^2) = \frac{\Delta^2}{12}\sum_{j=1}^N \|\mathbf{u}_j\|^2 + o(\Delta^2). \tag{4.3}$$

Proof. As usual, denote by F and G the frame matrices associated with the frame $\{\mathbf{v}_j\}_{j=1}^N$ and the dual frame $\{\mathbf{u}_j\}_{j=1}^N$, respectively. Let $H = G^T G$, $Y_j = \mathbf{X}\cdot\mathbf{v}_j$, $Z_j = \left\{\frac{Y_j}{\Delta} + \frac{1}{2}\right\} - \frac{1}{2}$, and $\mathbf{Z} = [Z_1, \ldots, Z_N]^T$. By Theorem 4.1, $\mathcal{E}(Z_i) \to 0$ and $\mathcal{E}(Z_i Z_j) \to \frac{1}{12}\delta_{ij}$ as $\Delta \to 0^+$. Now $\mathbf{X} - \tilde{\mathbf{X}} = G\mathbf{Z}$. It follows from the proof of Proposition 2.2 that

$$\begin{aligned}\frac{1}{\Delta^2}\mathcal{E}(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2) &= \mathcal{E}(\mathbf{Z}^T H \mathbf{Z}) \\ &= \mathcal{E}\left(\sum_{i,j=1}^N Z_i Z_j h_{ij}\right) \\ &= \sum_{i,j=1}^N h_{ij}\mathcal{E}(Z_i Z_j) \\ &= \frac{1}{12}\sum_{i=1}^N h_{ii} + o(1) \\ &= \frac{1}{12}\sum_{j=1}^d \lambda_j^{-1} + o(1),\end{aligned}$$

and hence

$$\mathcal{E}\left(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2\right) = \frac{\Delta^2}{12}\sum_{j=1}^d \lambda_j^{-1} + o(\Delta^2) = \frac{\Delta^2}{12}\sum_{j=1}^N \|\mathbf{u}_j\|^2 + o(\Delta^2). \qquad \square$$

5. Asymptotic behavior of errors: Linear dependence case. In this section we consider the case in which some vectors in the frame may be parallel. This can happen, for example, if the frame contains redundant elements. Mathematically it would be interesting to understand how the MSE behaves as $\Delta \to 0^+$. We return to previous calculations and note that

$$\mathcal{E}(\|\mathbf{X}-\tilde{\mathbf{X}}\|^2) = \sum_{i,j=1}^{N} h_{ij}\mathcal{E}\left(\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_i)\tau_\Delta(\mathbf{X}\cdot\mathbf{v}_j)\right).$$

Our main result in this section is the following theorem.

THEOREM 5.1. *Let X be an absolutely continuous real random variable. Let $\alpha \in \mathbb{R}\setminus\{0\}$. Then*

$$\lim_{\Delta\to 0^+} \frac{1}{\Delta^2}\mathcal{E}\left(\tau_\Delta(X)\tau_\Delta(\alpha X)\right) = \begin{cases} 0, & \alpha \notin \mathbb{Q}, \\ \dfrac{1}{12pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is even}, \\ -\dfrac{1}{24pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is odd}, \end{cases} \tag{5.1}$$

where p, q are coprime integers.

Proof. Denote $g(x) := \{x+\frac{1}{2}\} - \frac{1}{2}$. Let $\phi(x) \geq 0$ be an even C^∞ function such that $\operatorname{supp}(\phi) \subseteq [-1,1]$ and $\int_{\mathbb{R}} \phi = 1$. Let $g_n(x) = g * \phi_n$, where $\phi_n(x) = n\phi(nx)$. It is standard to check that

(a) $|g_n(x)| \leq 1/2$;
(b) $\operatorname{supp}(g(x) - g_n(x)) \subseteq [\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}] + \mathbb{Z}$;
(c) $g_n(x) \in C^\infty$ and is $\mathbb{Z}$-periodic; and
(d) $\int_{\mathbb{R}} g_n(x)\,dx = 0$.

$g_n(x)$ represents a small perturbation of $g(x)$ that "smoothes out" the discontinuities of $g(x)$. Now, set

$$\begin{aligned} E(\Delta) &:= \mathcal{E}\left(\frac{1}{\Delta^2}\tau_\Delta(X)\tau_\Delta(\alpha X)\right) \\ &= \mathcal{E}\left(g\left(\frac{X}{\Delta}\right) g\left(\frac{\alpha X}{\Delta}\right)\right) \\ &= \int_{\mathbb{R}} g\left(\frac{x}{\Delta}\right) g\left(\frac{\alpha x}{\Delta}\right) f(x)\,dx, \end{aligned}$$

and

$$E_n(\Delta) := \int_{\mathbb{R}} g_n\left(\frac{x}{\Delta}\right) g_n\left(\frac{\alpha x}{\Delta}\right) f(x)\,dx.$$

CLAIM. *$E_n(\Delta) \to E(\Delta)$ as $n \to \infty$ uniformly for all $\Delta > 0$.*

Proof of the claim. Let f be the density of $\mathbf{X}$. For any $\varepsilon > 0$,

$$\begin{aligned} |E_n(\Delta) - E(\Delta)| &= \left| \int_{\mathbb{R}} \left[g_n\left(\frac{x}{\Delta}\right) g_n\left(\frac{\alpha x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) g\left(\frac{\alpha x}{\Delta}\right) \right] f(x)\,dx \right| \\ &\leq \frac{1}{2}\int_{\mathbb{R}} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right) \right| f(x)\,dx \\ &\quad + \frac{1}{2}\int_{\mathbb{R}} \left| g_n\left(\frac{\alpha x}{\Delta}\right) - g\left(\frac{\alpha x}{\Delta}\right) \right| f(x)\,dx. \end{aligned}$$

Now there exists an $M > 0$ such that $\int_{[-M,M]^c} f(x)\,dx < \frac{\varepsilon}{2}$. So

$$\int_{\mathbb{R}} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right)\right| f(x)\,dx \le \int_{-M}^{M} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right)\right| f(x)\,dx + \frac{\varepsilon}{2}.$$

Furthermore, let $A_n(\Delta, M) := \operatorname{supp}(g_n(x/\Delta) - g(x/\Delta)) \cap [-M, M]$. Then we have

$$A_n(\Delta, M) \subseteq \Delta\left(\left[\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right] + \mathbb{Z}\right) \cap [-M, M].$$

Hence $\mathcal{L}(A_n(\Delta, M)) \le \frac{2M}{\Delta} \cdot \frac{2\Delta}{n} = \frac{4M}{n}$, and thus

$$\int_{-M}^{M} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right)\right| f(x)\,dx \le \int_{A_n(\Delta,M)} f(x)\,dx < \frac{\varepsilon}{2}$$

for $n > 4M/\varepsilon$ (which is independent of Δ). This yields

$$\int_{\mathbb{R}} \left| g_n\left(\frac{x}{\Delta}\right) - g\left(\frac{x}{\Delta}\right)\right| f(x)\,dx < \varepsilon.$$

Similarly we have

$$\int_{\mathbb{R}} \left| g_n\left(\frac{\alpha x}{\Delta}\right) - g\left(\frac{\alpha x}{\Delta}\right)\right| f(x)\,dx < \varepsilon$$

for sufficiently large n, proving the claim. □

Now consider the Fourier series of $g_n(t)$,

$$g_n(t) = \sum_{k\in\mathbb{Z}} c_k^{(n)} e^{2\pi i k t}.$$

It is well known that the Fourier series converges to $g_n(t)$ uniformly for all t; see, e.g., [24]. Furthermore, since $g_n(t)$ is C^∞ we have $|c_k^{(n)}| = o\left((|k|+1)^{-L}\right)$ for all $L > 0$, giving absolute convergence of the Fourier series. Thus

$$\begin{aligned} E_n(\Delta) &= \lim_{K\to\infty} \int_{\mathbb{R}} \Big(\sum_{|k|\le K} c_k^{(n)} e^{2\pi i k t \Delta^{-1}} \Big) \Big(\sum_{|k|\le K} c_k^{(n)} e^{2\pi i k \alpha t \Delta^{-1}} \Big) f(t)\,dt \\ &= \lim_{K\to\infty} \sum_{|k|,|\ell|\le K} c_k^{(n)} c_\ell^{(n)} \widehat{f}\Big(-\frac{k+\alpha\ell}{\Delta}\Big). \end{aligned}$$

Observe that $|\widehat{f}(\xi)| \le \|f\|_{L^1} = 1$, and $|c_k^{(n)}| = o\left((|k|+1)^{-L}\right)$ for any $L > 0$. So the series converges absolutely and uniformly in Δ. Thus

$$E_n(\Delta) = \sum_{k,\ell\in\mathbb{Z}} c_k^{(n)} c_\ell^{(n)} \widehat{f}\Big(-\frac{k+\alpha\ell}{\Delta}\Big). \tag{5.2}$$

For any $n > 0$ we have

$$\lim_{\Delta\to 0^+} E_n(\Delta) = \sum_{k,\ell\in\mathbb{Z}} c_k^{(n)} c_\ell^{(n)} \lim_{\Delta\to 0^+} \widehat{f}\Big(-\frac{k+\alpha\ell}{\Delta}\Big)$$

because the series converges absolutely and uniformly. Suppose $\alpha \notin \mathbb{Q}$. Then $k + \alpha\ell \neq 0$ if either $k \neq 0$ or $\ell \neq 0$. Thus $\left|-\frac{k+\alpha\ell}{\Delta}\right| \to \infty$ as $\Delta \to \infty$, and hence $\lim_{\Delta\to 0^+} \widehat{f}\big(-\frac{k+\alpha\ell}{\Delta}\big) = 0$ as $f \in L^1(\mathbb{R})$. Note also that $c_0^{(n)} = \int_{\mathbb{R}} g_n = 0$. It follows that

$$\lim_{\Delta\to 0^+} E_n(\Delta) = 0.$$

But $E_n(\Delta) \to E(\Delta)$ as $n \to \infty$ uniformly in Δ, which yields $E(\Delta) \to 0$ as $\Delta \to 0^+$.

Next, suppose $\alpha = \frac{p}{q}$, where $p, q \in \mathbb{Z}$, $(p, q) = 1$. We observe that $k + \alpha\ell = 0$ if and only if $k = pm$ and $\ell = -qm$ for some $m \in \mathbb{Z}$. In such a case

$$\widehat{f}\Big(-\frac{k+\alpha\ell}{\Delta}\Big) = \widehat{f}(0) = \int_{\mathbb{R}} f = 1.$$

It follows that

$$\lim_{\Delta\to 0^+} E_n(\Delta) = \sum_{m\in\mathbb{Z}} c_{pm}^{(n)} c_{-qm}^{(n)} \widehat{f}(0) = \sum_{m\in\mathbb{Z}} c_{pm}^{(n)} c_{-qm}^{(n)} = \sum_{m\in\mathbb{Z}} c_{pm}^{(n)} \overline{c_{qm}^{(n)}}.$$

For $r \in \mathbb{Z}, r \neq 0$ set

$$G_r^{(n)}(x) := \sum_{m\in\mathbb{Z}} c_{rm}^{(n)} e^{2\pi i m x}.$$

By Parseval we have

$$\lim_{\Delta\to 0} E_n(\Delta) = \Big\langle G_q^{(n)}, G_p^{(n)} \Big\rangle_{L^2([0,1])}.$$

It is easy to check that

$$G_r^{(n)} = \frac{1}{|r|} \sum_{j=0}^{|r|-1} g_n\Big(\frac{x+j}{r}\Big).$$

Hence $G_r^{(n)}$ converges in $L^2([0,1])$ to $G_r(x) := \frac{1}{|r|}\sum_{j=0}^{|r|-1} g\big(\frac{x+j}{r}\big)$, which has Fourier series $G_r(x) = \sum_{m\in\mathbb{Z}} c_{rm} e^{2\pi i m x}$ with $c_0 = 0$ and $c_k = \frac{(-1)^{k-1}}{2\pi i k}$ for $k \neq 0$. This yields

$$\lim_{n\to\infty} \lim_{\Delta\to 0^+} E_n(\Delta) = \lim_{n\to\infty} \Big\langle G_q^{(n)}, G_p^{(n)} \Big\rangle = \langle G_q, G_p \rangle = \sum_{m\in\mathbb{Z}} c_{qm} \overline{c_{pm}}.$$

Finally,

$$\begin{aligned} \sum_{m\in\mathbb{Z}} c_{qm}\overline{c_{pm}} &= \sum_{m\in\mathbb{Z}\setminus\{0\}} \Big(\frac{(-1)^{qm-1}}{2\pi i m q}\Big)\overline{\Big(\frac{(-1)^{pm-1}}{2\pi i m p}\Big)} \\ &= \frac{1}{2pq\pi^2} \sum_{m=1}^{\infty} \frac{(-1)^{(p+q)m}}{m^2}. \end{aligned}$$

Note that if $p + q$ is even then $\sum_{m=1}^{\infty} \frac{(-1)^{(p+q)m}}{m^2} = \sum_{m=1}^{\infty} \frac{1}{m^2} = \frac{\pi^2}{6}$. On the other hand, if $p + q$ is odd then $\sum_{m=1}^{\infty} \frac{(-1)^{(p+q)m}}{m^2} = \sum_{m=1}^{\infty} \frac{(-1)^m}{m^2} = -\frac{\pi^2}{12}$. The theorem follows. ∎

COROLLARY 5.2. *Let* $\mathbf{X}$ *be an absolutely continuous random vector in* $\mathbb{R}^d$, $\mathbf{w} \neq 0$, $\mathbf{w} \in \mathbb{R}^d$, *and* $\alpha \in \mathbb{R} \setminus \{0\}$. *Then*

$$\lim_{\Delta \to 0^+} \frac{1}{\Delta^2} \mathcal{E}\left(\tau_\Delta(\mathbf{w}\cdot\mathbf{X})\tau_\Delta(\alpha\mathbf{w}\cdot\mathbf{X})\right) = \begin{cases} 0, & \alpha \notin \mathbb{Q}, \\ \dfrac{1}{12pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is even}, \\ -\dfrac{1}{24pq}, & \alpha = \dfrac{p}{q} \text{ and } p+q \text{ is odd}, \end{cases} \tag{5.3}$$

where p, q *are coprime integers.*

Proof. We need only to note that $\mathbf{w}\cdot\mathbf{X}$ is an absolutely continuous random variable. The corollary follows immediately from Theorem 5.1. □

We can now characterize completely the asymptotic behavior of the MSE in all cases. For any two vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ define $r(\mathbf{w}_1, \mathbf{w}_2)$ by

$$r(\mathbf{w}_1, \mathbf{w}_2) = \begin{cases} \dfrac{1}{pq}\mathbf{w}_1\cdot\mathbf{w}_2, & \mathbf{w}_1 = \dfrac{p}{q}\mathbf{w}_2, \text{ and } p+q \text{ is even}, \\ -\dfrac{1}{2pq}\mathbf{w}_1\cdot\mathbf{w}_2, & \mathbf{w}_1 = \dfrac{p}{q}\mathbf{w}_2, \text{ and } p+q \text{ is odd}, \\ 0, & \text{otherwise}, \end{cases}$$

where p, q are coprime integers.

COROLLARY 5.3. *Let* $\mathbf{X} \in \mathbb{R}^d$ *be an absolutely continuous random vector. Then as* $\Delta \longrightarrow 0^+$ *the MSE satisfies*

$$\mathcal{E}\left(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2\right) = \frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1} + \frac{\Delta^2}{6}\sum_{1\le i<j\le N} r(\mathbf{u}_i, \mathbf{u}_j) + o(\Delta^2). \tag{5.4}$$

Proof. In the proof of (4.2) we showed that

$$\lim_{\Delta\to 0^+}\frac{1}{\Delta^2}\mathcal{E}\left(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2\right) = \sum_{i,j} h_{ij}\mathcal{E}\left(Z_i Z_j\right)$$

with the notation there. Observe that $h_{ij} = \mathbf{u}_i \cdot \mathbf{u}_j$. The result is immediate from Corollary 5.2. □

For fixed quantization step $\Delta > 0$ we shall denote

$$\mathrm{MSE}_{ideal} = \frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1} + \frac{\Delta^2}{6}\sum_{1\le i<j\le N} r(\mathbf{u}_i, \mathbf{u}_j) \tag{5.5}$$

and call it the *ideal* MSE. If $\{\mathbf{v}_j\}_{j=1}^N$ are pairwise linearly independent, then the MSE_{ideal} is simply $\frac{\Delta^2}{12}\sum_{j=1}^{d}\lambda_j^{-1}$, the MSE under the WNH.

We should point out that even though the WNH is not true aysmpototically if some vectors in a frame are parallel, the contribution from the second part of (5.5) is often small enough that the MSE under the WNH is close enough to the ideal MSE. In the next section we shall show some numerical data comparing the actual MSE with the ideal MSE.

Appendix A. Numerical results. Here we present data from our computer experiments comparing the ideal MSE to the actual MSE. We performed Monte Carlo

TABLE 1
The harmonic frame in $\mathbb{R}^2$.

N	Actual MSE	Ideal MSE	Ratio
9	0.00934342	0.00925926	1.009090
17	0.00479521	0.00252525	0.976808
33	0.00246669	0.00490196	0.978223
65	0.00122499	0.00128205	0.955496
129	0.00065858	0.000645995	1.019480
257	0.00057971	0.00032425	1.787810
513	0.00056039	0.00016244	3.449740
1025	0.00052914	0.00008130	6.508450
2049	0.00053895	0.00004067	13.25180
4097	0.00058846	0.00002034	28.93090

TABLE 2
The randomly generated frame in $\mathbb{R}^4$.

k/N	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
$k = 0$	1.581960	2.232260	3.697160	6.497800	12.20670
$k = 1$	1.076590	1.130510	1.397840	1.649530	2.480920
$k = 2$	1.003680	0.995214	1.008370	1.033280	1.196680
$k = 3$	0.967138	0.990876	0.999648	0.981633	1.010090
$k = 4$	0.989295	1.009840	1.032110	1.002630	1.002260
$k = 5$	1.011720	1.035590	1.020870	1.002350	1.022250
$k = 6$	0.978712	1.006760	0.992207	1.001490	0.979342
$k = 7$	0.997524	1.017840	0.995852	0.972120	0.976273
$k = 8$	0.998725	1.011380	1.040270	0.978204	0.973284
$k = 9$	0.982450	1.038580	0.994463	1.021580	1.037800
$k = 10$	0.993099	1.002340	1.009930	1.009870	0.974017
$k = 11$	0.981428	0.998280	0.975881	1.049010	1.009570

simulations for several different sets of frames. We also experimented with various distributions for $\mathbf{X} \in \mathbb{R}^d$. As it turned out, we got very similar results for the distributions we used for most of the frames we tried. In the examples shown, the random vectors $\mathbf{X}$ are all chosen to be uniformly distributed in $[-5, 5]^d$.

Example A.1. Let $\{\mathbf{v}_j\}_{j=1}^N$ be the harmonic frame in $\mathbb{R}^2$, with $\mathbf{v}_j = \left[\cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}\right]^T$. This is a tight frame with frame constant $\lambda = \frac{N}{2}$. The ideal MSE is $\frac{\Delta^2}{3N}$ for N odd. Taking $\Delta = \frac{1}{2}$, Table 1 displays the actual MSE, the ideal MSE, and the ratio between them. It shows that as N gets larger than 129, the actual MSE does not improve, which shows that the WNH is invalid for large Δ.

Example A.2. Let $\{\mathbf{v}_j\}_{j=1}^N$ be N independently and randomly generated vectors uniformly distributed on the unit sphere in $\mathbb{R}^4$. Table 2 shows the ratio between the actual MSE and the ideal MSE, where $\mathrm{MSE}_{ideal} = \frac{\Delta^2}{12}(\sum_{j=1}^d \lambda_j^{-1})$, with $\Delta = 2^{-k}$.

Example A.3. Let $\{\mathbf{v}_j\}_{j=0}^{N-1}$ be the harmonic frame in $\mathbb{R}^4$, with

$$\mathbf{v}_j = \sqrt{\frac{1}{2}}\left[\cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}, \cos\frac{4\pi j}{N}, \sin\frac{4\pi j}{N}\right]^T.$$

This is a tight frame with frame constant $\lambda = \frac{N}{4}$, and the ideal MSE is $\frac{4\Delta^2}{3N}$. Table 3 shows the ratio between the actual MSE and the ideal MSE, where $\Delta = 2^{-k}$.

TABLE 3
The harmonic frame in $\mathbb{R}^4$.

k/N	$N=64$	$N=128$	$N=256$	$N=512$	$N=1024$
$k=0$	0.997218	0.928318	1.287990	2.312710	4.497050
$k=1$	1.005460	1.004720	0.950783	1.339810	2.395180
$k=2$	0.990253	1.001070	0.977474	0.960994	1.354320
$k=3$	0.995848	0.993963	0.981683	0.992655	0.955345
$k=4$	0.987371	1.007310	1.028120	1.016760	1.002570
$k=5$	0.993840	1.015230	1.026680	1.003770	1.023820
$k=6$	1.012230	1.012280	0.996363	0.999742	1.004120
$k=7$	1.020450	1.025820	1.031120	1.003770	1.004770
$k=8$	1.004710	1.010820	0.999289	0.973596	0.970415
$k=9$	0.993542	1.003380	0.981550	0.984594	0.981001
$k=10$	1.015610	1.008740	0.997469	0.986705	1.004360
$k=11$	1.010690	1.009080	0.994975	1.010510	0.998485

TABLE 4
The frame of Example A.4 *in* $\mathbb{R}^3$.

k	Actual MSE	Ideal MSE	MSE under WNH
2	0.012234100000	0.011880200000	0.011363600000
3	0.002935150000	0.002970040000	0.002840910000
4	0.000732567000	0.000742510000	0.000710227000
5	0.000188331000	0.000185628000	0.000177557000
6	0.000046664900	0.000046406900	0.000044389200
7	0.000011626300	0.000001160170	0.000011097300
8	0.000002953720	0.000002900430	0.000002774330
9	0.000000724800	0.000000725108	0.000000693581
10	0.000000180127	0.000000181277	0.000000173395
11	0.000000045856	0.000000045319	0.000000043349

Example A.4. Let $\{\mathbf{v}_j\}_{j=1}^5$ be a frame in $\mathbb{R}^3$, with the corresponding matrix

$$F = \begin{pmatrix} 1 & 1 & 0 & 1 & 3 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Note that the set contains many parallel vectors. The dual frame matrix is

$$\begin{pmatrix} \frac{1}{11} & 0 & 0 & \frac{1}{11} & \frac{3}{11} \\ \frac{1}{11} & -1 & 0 & \frac{1}{11} & \frac{3}{11} \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

The MSE under the WNH is $0.181818\Delta^2$ and, by our result, the ideal MSE is $0.190083\Delta^2$, which is closer to the actual MSE. The difference between the two estimates comes from the second part in (5.5). Table 4 shows the actual MSE, the ideal MSE, and the MSE under the WNH, where $\Delta = 2^{-k}$.

Acknowledgments. Several people have given us helpful suggestions on this paper. But in particular we wish to express our gratitude to the referee, who not only read the manuscript very carefully but provided us with a number of valuable suggestions, particularly on the vast engineering literature regarding vector quantization.

REFERENCES

[1] J. Benedetto and M. Fickus, *Finite normalized tight frames*, Adv. Comput. Math., 18 (2003), pp. 357–385.

[2] J. Benedetto, A. M. Powell, and Ö. Yilmaz, *Sigma-delta ($\Sigma\Delta$) quantization and finite frames*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1990–2005.

[3] J. Benedetto, A. M. Powell, and Ö. Yilmaz, *Second order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions*, Appl. Comput. Harmon. Anal., 20 (2006), pp. 126–148.

[4] W. Bennett, *Spectra of quantized signals*, Bell Syst. Tech. J., 27 (1948), pp. 446–472.

[5] S. Bochner and K. Chandrasekharan, *Fourier Transforms*, Princeton University Press, Princeton, NJ, 1949.

[6] P. G. Casazza and J. Kovačević, *Uniform tight frames with erasures*, Adv. Comput. Math., 18 (2003), pp. 387–430.

[7] I. Daubechies and R. DeVore, *Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order*, Ann. Math. (2), 158 (2003), pp. 679–710.

[8] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341–366.

[9] Y. Elder and G. D. Forney, *Optimal tight frames and quantum measurement*, IEEE Trans. Inform. Theory, 48 (2002), pp. 599–610.

[10] D. J. Feng, L. Wang, and Y. Wang, *Generation of finite tight frames by Householder transformations*, Adv. Comput. Math., 24 (2006), pp. 297–309.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, Boston, 1992.

[12] V. K. Goyal, J. Kovačcević, and J. Kelner, *Quantized frame expansions with erasures*, Appl. Comput. Harmon. Anal., 10 (2001), pp. 203–233.

[13] V. K. Goyal, M. Vetterli, and N. T. Thao, *Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms*, IEEE Trans. Inform. Theory, 44 (1998), pp. 16–31.

[14] R. M. Gray and D. L. Neuhoff, *Quantization*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2325–2383.

[15] R. Gray, *Quantized noise spectra*, IEEE Trans. Inform. Theory, 36 (1990), pp. 1220–1244.

[16] S. Güntürk, *Approximating a bandlimited function using very coarsely quantized data*, J. Amer. Math. Soc., 17 (2004), pp. 229–242.

[17] Y. Katznelson, *Harmonic Analysis*, Wiley, New York, 1968.

[18] T. Linder, R. Zamir, and K. Zeger, *High resolution source coding for nondifference distortion measure: Multidimensional companding*, IEEE Trans. Inform. Theory, 45 (1999), pp. 548–561.

[19] S. Na and D. L. Neuhoff, *Bennett's integral for vector quantizers*, IEEE Trans. Inform. Theory, 41 (1995), pp. 886–900.

[20] G. Rath and C. Guillemot, *Recent advances in DFT codes based on quantized finite frame expansions for erasure channels*, Digital Signal Processing, 14 (2004), pp. 332–354.

[21] N. Thao and M. Vetterli, *Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimate*, IEEE Trans. Signal Process., 42 (1994), pp. 519–531.

[22] H. Viswanathan and R. Zamir, *On the whiteness of high-resolution quantization errors*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2029–2038.

[23] R. Zamir and M. Feder, *On lattice quantization noise*, IEEE Trans. Inform. Theory, 42 (1996), pp. 1152–1159.

[24] A. Zygmund, *Trigonometrical Series*, Chelsea, New York, 1950.